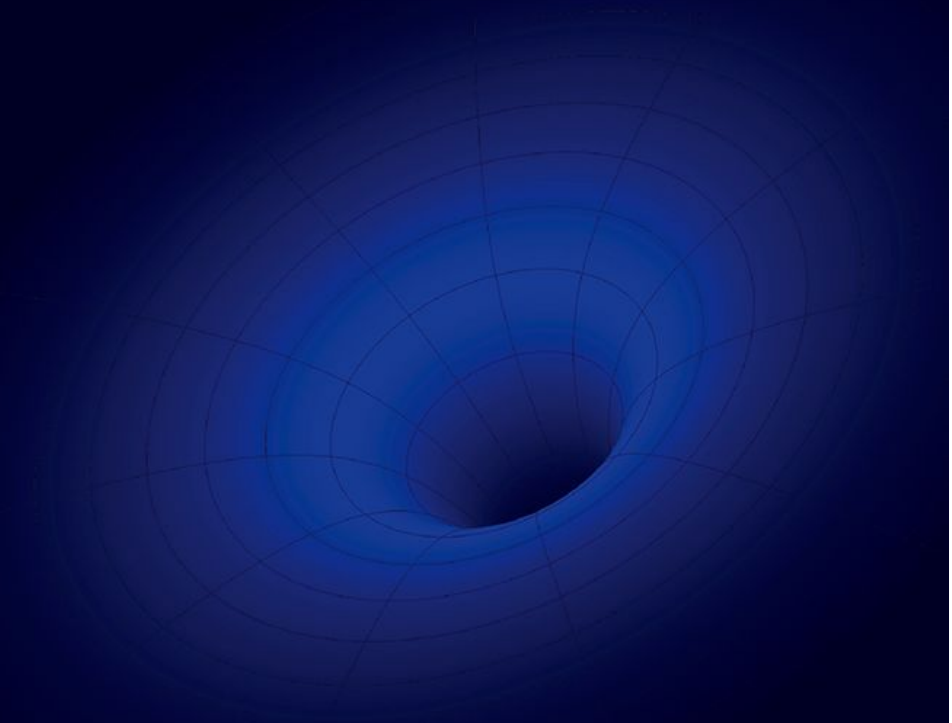
MOATAZ H. EMAM

COVARIANT PHYSICS

From Classical Mechanics to General Relativity & Beyond



OXFORD

Covariant Physics

Covariant Physics

From Classical Mechanics to General Relativity and Beyond

Moataz H. Emam
State University of New York College at Cortland

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First Edition published in 2021

Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data

Data available

Library of Congress Control Number: 2020951689

ISBN 978–0–19–886489–9 (hbk.)
ISBN 978–0–19–886500–1 (pbk.)

DOI: 10.1093/oso/9780198864899.001.0001

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

To Miriam, Maya, and Mourad

Preface

This is *not* "yet another book on the general theory of relativity." While the theory did inspire, and indeed plays a major role in, our discussion, it can be thought of as only one part of a greater whole. Initially developed as a theory of gravity, it was slowly realized that general relativity had a deeper meaning than just that. It teaches us, among other things, a completely new understanding of space and time, together renamed **spacetime**: a single fabric with its own particular set of symmetries and properties. While we learn that the familiar Euclidean geometry[1] that we have all learned (and loved?) in school no longer applies, we also discover that its most fundamental symmetries are not lost, but rather "upgraded" to new and deeper ones; leading to a "spacetime" that's a lot more dynamic and flexible than the one we learned about from Euclid, Galileo,[2] or Newton.[3] A byproduct of this dynamical flexibility, prominent in general relativity, is the phenomenon that we call "gravity;" but it is *only* a byproduct. It is actually possible, at least in principle, to go through an entire course on general relativity without mentioning the word "gravity" once, instead just focusing on its formal structure and the symmetries it inspires! Clearly such a choice would not be a very productive one, as it means that we would be missing out on what is undoubtedly the most important application of the theory; nevertheless, it *is* possible. The theory itself is the culmination of the efforts of many physicists and mathematicians, most notably Albert Einstein,[4] Hermann Minkowski,[5] Hendrik Lorentz,[6] Henri Poincaré,[7] David Hilbert,[8] and others. Einstein, of course, is responsible for the more revolutionary ideas in the theory. Today it is considered his greatest contribution to physics, as well as possibly the most amazing singular achievement by a human being in history.[9]

[1] Euclid of Alexandria was a Greek mathematician who lived around the third century BCE. His book *Elements* is considered the foundational text of modern geometry in two- and three- dimensional flat space. His axioms, theorems, and lemmas constitute the familiar geometry taught today to high school students. The term "Euclidean," coined in his honor, signifies flat space, as opposed to "non-Euclidean," or curved spaces.

[2] Galileo Galilei: Italian astronomer, physicist, engineer, philosopher, and mathematician (1564–1642).

[3] Sir Isaac Newton: English mathematician, physicist, astronomer, theologian, and author, requiring perhaps no introduction (1642–1726).

[4] German physicist (1879–1955).

[5] German mathematician (1864–1909).

[6] Dutch physicist (1853–1928).

[7] French mathematician and theoretical physicist (1854–1912).

[8] German mathematician (1862–1943).

[9] Charles Darwin's theory of natural selection might be a close contender.

One of the most fundamental characteristics of this new understanding of space and time that has arisen from, but, one can argue, has become even more important than, the theory of relativity, is what mathematicians sometimes call **diffeomorphism invariance**, which, as far as we are concerned, can be thought of as simply the "invariance of the laws of physics between coordinate systems," also known as **coordinate covariance**, or just **covariance**. This idea, which will be our main focus for most of this book, lies at the deepest foundations of physics and is certainly true in the special theory of relativity (**special covariance**), the general theory (**general covariance**), and even ordinary non-relativistic physics, i.e. the usual Euclidean/Galilean/Newtonian world view (**classical covariance**). What it means is this: The use of any type of coordinate system does *not* affect the physical outcome of any given problem. A specific physics formula applied to some problem should give *exactly* the same physical answers whether one has used Cartesian coordinates or any other type of coordinate system (and they are legion). This should be intuitively true even to the beginner, because coordinate systems are a human choice and nature should not, and indeed does not, care about our choices or conventions. On the other hand, the way the formula *looks* in Cartesian coordinates is significantly different from the way it looks in other coordinates. So an important question to ask is this: How can one write the laws of physics in a single form that is true in *any* coordinate system? In other words, in a form that is coordinate *covariant*, yet is easy enough to use once a choice of coordinates has been made? In this book we will approach our study of space and time from the perspective of covariance; as such, it is perhaps important to understand what we mean by this, even this early in the reading. Let's then clarify further: Consider Newton's second law as an example. It is usually first encountered in introductory courses in the following form:

$$\mathbf{F} = m\mathbf{a}, \tag{0.1}$$

where $\mathbf{F}$ is the vector representing the net sum of all external forces acting on a mass m and resulting in an acceleration $\mathbf{a}$.[10] Written in this form, eqn (0.1) is correct in any coordinate system, i.e. in a sense it is *already* covariant. But to actually calculate numerical results one still needs to make a specific choice of coordinates which will necessarily split the equation into its constituent components and have a form that is highly dependent on the coordinate system used. For example, in two-dimensional Cartesian coordinates (x, y), Newton's law reduces to the following set of equations:

$$\begin{aligned} F_x &= ma_x \\ F_y &= ma_y, \end{aligned} \tag{0.2}$$

where (F_x, F_y) and (a_x, a_y) are the Cartesian components of the force and acceleration respectively. With the knowledge that acceleration is the second derivative of the position we can write

[10] We will mostly denote vectors by bold face type font, except when there is a possibility of confusion. In such cases we will use an arrow over the symbol to denote a vector; for example, $\vec{F}$.

$$F_x = m\frac{d^2x}{dt^2}$$
$$F_y = m\frac{d^2y}{dt^2}, \tag{0.3}$$

or in a more compact form

$$F_x = m\ddot{x}$$
$$F_y = m\ddot{y}, \tag{0.4}$$

where an over dot is the standard notation for a derivative with respect to time ($\dot{x} = \frac{dx}{dt}$, $\ddot{x} = \frac{d^2x}{dt^2}$, and so on). This should be familiar even to the reader who has only had an introductory physics course, but now consider writing the equivalent equations in polar coordinates (ρ, φ). We find that they are *not*:

$$F_\rho = m\ddot{\rho}$$
$$F_\varphi = m\ddot{\varphi}. \qquad \text{Wrong!} \tag{0.5}$$

In other words, one does not simply replace $x \to \rho$ or $y \to \varphi$ and expect to get the right answer. (In fact, it is clear that the second equation in (0.5) does not even have the correct units.) The correct relationships are actually

$$F_\rho = m\ddot{\rho} - m\rho\dot{\varphi}^2$$
$$F_\varphi = m\rho\ddot{\varphi} + 2m\dot{\rho}\dot{\varphi}. \tag{0.6}$$

The two sets of eqns (0.4) and (0.6) are not identical in form, as (0.4) is with the (wrong) eqns (0.5). Specifically, there are "extra" terms and multiplicative factors in (0.6), and it is not immediately obvious where these came from or whether there is a pattern that is common to them. It is, however, true that applying either one of these choices to the same problem will describe the same physical behavior and give the same physical results. But where did the extra quantities come from? Mathematically, they arise from the relationships that relate polar coordinates to Cartesian coordinates, i.e. $x = \rho\cos\varphi$ and $y = \rho\sin\varphi$, as can be easily derived from Fig. [1.3b]. All one needs to do is to use calculus to find how $\dot{x}$ and $\dot{\rho}$ (for example) are related, and so on. In principle one needs to rederive the Newtonian equations of motion each and every time one makes a specific choice of a coordinate system. This is indeed the way it is usually done in physics classes. However, what if the coordinate system is *not* defined ahead of time? What if it depends on the curvature of space and time that is not *a priori* known? What if the coordinate system *itself* arises as the solution of the problem under study? Can one rewrite (0.1) or any other law of physics in such a way that it remains coordinate invariant and gives the correct set of equations for *any* coordinate system, whether it was known from the start or still to be derived? This is the problem of coordinate invariance in physics, or covariance as it is more commonly referred to.

Learning how to do all of this is a useful thing, whether one is going to study relativity or not, because it provides the student with a recipe for dealing with any choice of coordinate system in any physics problem, whether such a choice is one of the more common ones, like polar coordinates in two dimensions or cylindrical and spherical coordinates in three, or less common, such as parabolic coordinates or elliptic coordinates. In short, it provides the physicist (or engineer) with a powerful mathematical tool applicable to a variety of situations. It also acts as an easy (-*ish*) introduction to the subject of **tensor** mathematics, rarely discussed in an undergraduate setting, yet encountered in almost every branch of physics. While these "tensors" are particularly useful in a relativistic setting, they can also be found in non-relativistic mechanics and in electrodynamics, and knowledge of their properties and rules would be particularly useful for the student wishing to delve deeper into these topics, or at least hopes for a better understanding than is normally delivered to undergraduates.[11] We will touch upon all of this, and by the time we are done we will have a general understanding of (almost) all of fundamental (non-quantum) physics in a covariant setting and arbitrary spacetime backgrounds. We conclude the book with three chapters focusing on topics rarely, if ever, found in similar treatments. These are: classical fields, differential forms, and modified theories of gravity. These chapters are particularly geared towards the students interested in reading modern research papers. They should provide just a bit of extra preparation as well as possibly whet the reader's appetite for more.

A note to instructors:

This book is written from a minimalist's perspective. By no stretch of the imagination am I claiming that the discussions here are exhaustive or complete. In fact, I imagine that most mathematicians would be particularly appalled by the lack of rigor throughout, and most physicists will point out that many interesting applications and discussions commonly found in similar monographs are missing. This is all quite intentional. The intended audience of this book are those undergraduate students who may have only had introductory mechanics and electromagnetism at the level of Serway [1], for instance, as well as elementary calculus up to and including multivariable calculus. While the mathematical background gets a bit more intense in some places, it is my hope that students with only that much preparation can manage the majority of the book. As the students progress in their studies and reach more advanced courses in mechanics and electrodynamics, one hopes that this book, or the memory thereof, would provide a background tying all the various disciplines together in one continuous thread. The theme of covariance permeates all of physics, and yet is rarely discussed in an undergraduate setting. I have tried to make the book as self-contained as possible, such that a student reading it on their own would manage most of it with as little help as possible.

[11] It truly hurts when in teaching an undergraduate class I find myself forced to say something like "This is called the so-and-so tensor, but unfortunately we have no time in this class to explain what that *really* means."

One also hopes that this book can provide the minimum mathematical training needed by undergraduate students wishing to do research in certain areas of theoretical physics. In fact, the idea of the book, as well as the level of detail and sequence of topics, has arisen from my years of supervising undergraduate research. With graduation dates preset, one would like one's students to start their research experience as early as possible in their college years, and get to finish within the time they have available. In my undergraduate theoretical physics research program I have usually designed my students' learning experience in such a way that they would be able to get to the research part as fast as possible, amassing the needed mathematical training in the shortest possible period of time. In many cases, I have instructed them to focus on certain topics while skipping others, because of my preconceived knowledge of the nature of the research problem that I will eventually pose to them. For example, a student whose research will be calculating geodesic orbits of certain spacetime backgrounds need not waste their initial study sessions on understanding the details of how these spacetimes were derived in the first place; rather a cursory knowledge would be acceptable. If undergraduate research is the instructor's intention in assigning this book to their students, then it is advisable to plan their reading ahead of time. Of course, as they work on their projects they may go back and study in more detail those sections that they have initially only skimmed over. I have tried to make this book exhaustive enough for these purposes, although depending on what the research mentor has in mind, other books and/or notes can be supplemented.

On the exercises in the book:

I have tried to include as many exercises for practice and/or homework assignments as I could without overwhelming the reader. Many of the exercises in the book are chosen not only because they provide practice on the topics we discuss, but also because they include some additional material of their own. For example, some of the exercises require the reader to teach themselves computational techniques not discussed in the text or read on a topic that we didn't include in detail. Some may even be long enough to count as a final project or something similar, e.g. Exercise 6.49, or Exercises 7.26 and 7.27 combined. These increase in frequency, as well as intensity, as we get to the final chapters. A word of warning, however: Please be aware that some of the exercises list *many* different ways of practice; so many, in fact, that the student need not do all. For example, an exercise may say "do such and such for the following cases" and list *10* different cases; e.g. Exercise 2.22 which in turn is referred to multiple times in subsequent exercises. The student really need *not* do all of the listed cases; only enough to feel they have reached a certain degree of comfort with the techniques. Likewise, if this book is used in a formal course, the instructor should choose some of the cases for homework assignments, or even introduce their own. Many exercises depend on previous ones, however. For example, Exercise 2.23 needs the results of Exercise 2.22, so if the reader chooses to do, say three, of the cases in 2.22, then they would only be able to do the same three cases in 2.23, and so on.

Some of the exercises assigned already have solutions available in various places around the internet. This is unavoidable. However, it is understood that it is in your best interest to work out the exercises yourself, *then* check against the available solutions. It is also true that there are freely available software packages that would solve many of the exercises in this book for you. This is the case for most mathematical topics in this day and age. I actually *encourage* the reader to find and download these packages, or better yet learn to design and write your own. You can do so from scratch using any programming language or from within one of the commercially available symbolic manipulation software packages. In other words, you should treat the digital computer as a tool for learning, *not* as a way of avoiding work. The choice of which exercises to do by hand and which to do using the computer is mostly left to the reader or their instructors, even if not specifically stated. For example, once the instructor is satisfied that the students are able to calculate a specific quantity that usually takes hours to find by hand, say the components of a particularly large tensor, then they might just allow the students to use software packages for the remainder of the course.

Acknowledgements

It is very difficult, if not downright impossible, to give credit to every single person that I would like to. For I believe that any direct help I have had in writing this book is no more important than the indirect positive effect many people have had on my life in general; without which I would not have become the person who is able to write such a book in the first place. As such, I beg forgiveness to anyone that I won't mention and assure them that I appreciate them no less. To begin, I must acknowledge my parents' love, encouragement, and support. These are the foundations upon which we all grow and advance. Rest in peace mom and dad; I hope you are as proud of me as I am proud to be your son. Secondly, I acknowledge the support I have received from my family and close friends; you are too many to mention, but you were all supportive and understanding of my efforts in writing this book. I probably would have ended up nowhere without you. Thirdly I must acknowledge my students over at least the past fifteen years, in both the United States and Egypt. I have used the notes on which this book is based to teach them the material, and they have always honored me with valuable feedback and comments. I would also like to thank the colleagues, teachers, and friends who have had an immediate impact on this book, whether via discussions we have had, support and encouragement (sometimes such a long time ago), or direct feedback. These are, in no particular order:

- Dr. Mark Peterson of Mt. Holyoke College, MA: You allowed me to offer my very first class on the general theory of relativity. The notes I developed for that course were the earliest version of this book. Our discussions about teaching methods have had a lasting effect on me as a teacher and as an author. Your encouragement and faith in my abilities have stayed with me since then.
- Dr. Ahmed Alaa Eldin Abouelsaood of Cairo University, Egypt: Now where would I have ended without your advice and encouragement? You took a chance on a young man who came to you with nothing other than a passion for physics, and your gamble paid off (I hope). Not only did you allow me to absorb from you as much knowledge as I possibly could have at the time, you were also the one who provided the initial guidance which, quite literally, started the rest of my life.
- Dr. David Kastor of the University of Massachusetts at Amherst, MA: As my PhD advisor you did much more than supervise me. You allowed me to pick your brains in many topics in theoretical physics and patiently listened to me as I described how I understood things, even if my ideas sounded simple or mundane. These thoughts, which you encouraged, have slowly developed and matured, and became major contributions to this work. More recently, thanks for taking the time to read the manuscript and coming up with useful comments as well as suggestions for new sections. I am particularly grateful to you for suggesting the subtitle of the book.
- Dr. Eric Edlund of the State University of New York College at Cortland, NY: Thanks for the time you dedicated to reading the manuscript and providing excellent suggestions and comments.

- Mr. Tariq Saeed of the University of Mansoura, Egypt: Thanks for providing valuable input from a student perspective. Your suggestions on sections §3.3, §4.7, and others were particularly useful. I also truly appreciated your taking the time to check many of the calculations in this book.
- Eng Muhammad El-Nashash'ee, and Eng Akram Masoud: Thank you for taking the time to read the manuscript and providing valuable feedback and suggestions.
- My cousin and dear friend Brigadier General Eng Omar Farouk: Many thanks for the big push you gave me at a time I desperately needed one.
- Last, but certainly not least, my dear friend Mr. Ahmed Abuali: I will always be grateful for your valuable input, help, as well as many hours of discussion.

Contents

With the exception of Fig. [5.12] and [8.7], all the figures are original by the author. Software used: "Wolfram Mathematica® 11.3," and "Inkscape® 0.92."

1

Coordinate Systems and Vectors

Geometry has two great treasures: one is the Theorem of Pythagoras; the other, the division of a line into extreme and mean ratio. The first we may compare to a measure of gold; the second we may name a precious jewel.

Johannes Kepler

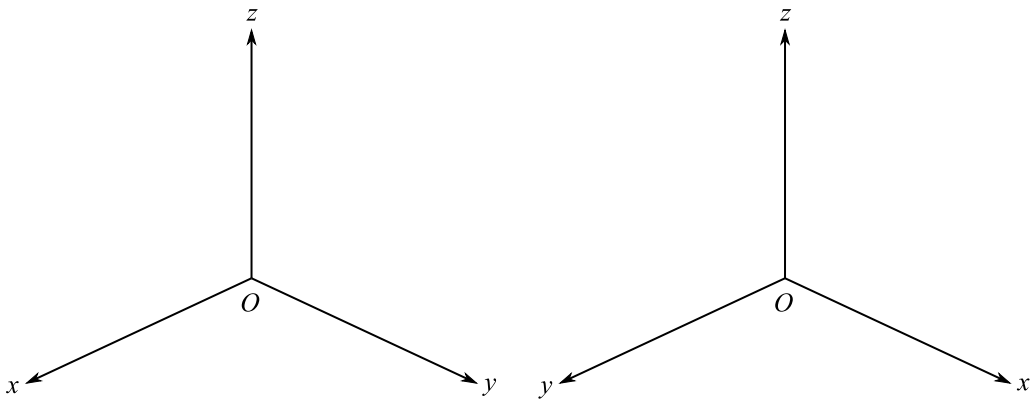
Introduction

Coordinate covariance is defined as the invariance of the laws of physics under change of coordinate systems, whether they differ from each other by translation, rotation, or both. In this chapter we begin by introducing the various two- and three-dimensional coordinate systems usually used in problems of physics and engineering. Some of these should be quite familiar to the reader, while others may be completely new. We specifically highlight the concept of "line element," also known as the "metric," as a defining principle of coordinate systems as well as a unifying theme amongst them. This should allow for a smoother transition into physics later. In principle, all that an experienced physicist needs in order to figure out what type of covariance is at play in a given situation is knowledge of the metric. Hence introducing the concept early on makes sense and allows us time to get used to it before things get more interesting. Next we will turn to defining the basics of vector mathematics and discussing their formulation in a new language simply known as the "index" or "component" notation.[1] This notation by itself connects vectors to the coordinate systems they are defined in and highlights their covariance. Furthermore, the notation provides the majority of the mathematical material that we need to collect in our search for a fully covariant formulation of physics, relativistic or not. In fact, I claim that the first two chapters *by themselves* constitute at least 80 percent of the mathematics needed to study *both* theories of relativity. As the reader will see, none of this mathematics is too far beyond ordinary algebra and a bit of multivariable calculus.

[1] Also sometimes referred to as the "tensor notation;" however, this is somewhat misleading, so we will avoid using it too much.

1.1 Coordinate Systems

For the purposes of many science and engineering applications, the most commonly used system of coordinates is the **Cartesian**,[2] also known as **box** or **rectangular** coordinates. By their very nature, Cartesian coordinates emphasize the intuitive concepts of width, depth, and height. It is usually the coordinate system of choice for **Euclidean** geometry; i.e. the ordinary geometry we all studied in school. It is, however, only *one* choice among many, as we will see. Euclidean geometry describes flat space, where parallel lines stay parallel and the angles of a triangle sum up to exactly 180°. If one happens to be interested in curved or **non-Euclidean** geometries, like cartographers doing geometry on the surface of the spherical globe, then not only do Cartesian coordinates stop making sense, but they actually become impossible to define, except approximately on a very small scale. To clarify this, consider the surface of planet Earth as an example, assumed a perfect sphere; one can draw straight lines in a grid on the floor in a room or in the street without trouble, but if you want to draw longer "straight" lines between cities, countries, or even continents, you find that you cannot do so and are forced to work with circles, like those of latitude or longitude. In a sense, then, most curved surfaces of interest are approximately flat on small scales; one says they are *locally* flat, as opposed to *globally* so.



(a) A right-handed coordinate system. (b) A left-handed coordinate system.

Fig. 1.1: Cartesian coordinates.

Traditionally, we use the symbols x, y, and z to denote the three Cartesian axes in three-dimensional space, ordered in the so-called "right-handed" system, as shown in Fig. [1.1a]. The right-handedness convention is particularly useful in physics, as it is the basis of various mathematical operations such as vector cross products. For comparison a left-handed system is shown, but we will not be using that. Generally,

[2]Invented in the seventeenth century by the French philosopher and mathematician René Descartes (1596-1650), whose name in Latinized form is Cartesius.

coordinate systems are defined via the geometric shapes they seem to trace. So in the case of Cartesian coordinates, surfaces of constant x, y, and z are infinite flat planes as shown in Fig. [1.2], hence the alternative title "box coordinates." Now the conventional choice of the letters x, y, and z to denote the Cartesian axes is very familiar to most people; however, a more useful notation, and one that we might as well start using right away, employs *indices* to denote the different axes. So x becomes x^1 (i.e. the *first* axis), y, being the second axis, becomes x^2, and $z \to x^3$, where the reader is warned *not* to confuse indices with powers, so x^2 is not x squared, but is simply pronounced "ex two" and similarly for "ex one" and "ex three." If we wish to denote exponentiation, then we can simply use parentheses; for example, the square of "ex three" is $\left(x^3\right)^2 = x^3 \times x^3 = z^2$. If there is a possibility of confusion, we can temporarily go back to the old (x, y, z) system, depending on the context. This "numbering" system, also known as the **index notation** or the **component notation**, is a much more powerful choice than the conventional one for a variety of reasons, as we will see.

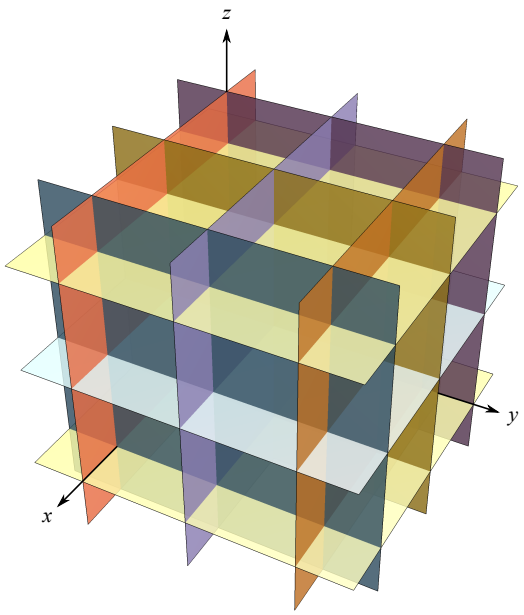


Fig. 1.2: The Cartesian grid: planes of constant x, y, and z.

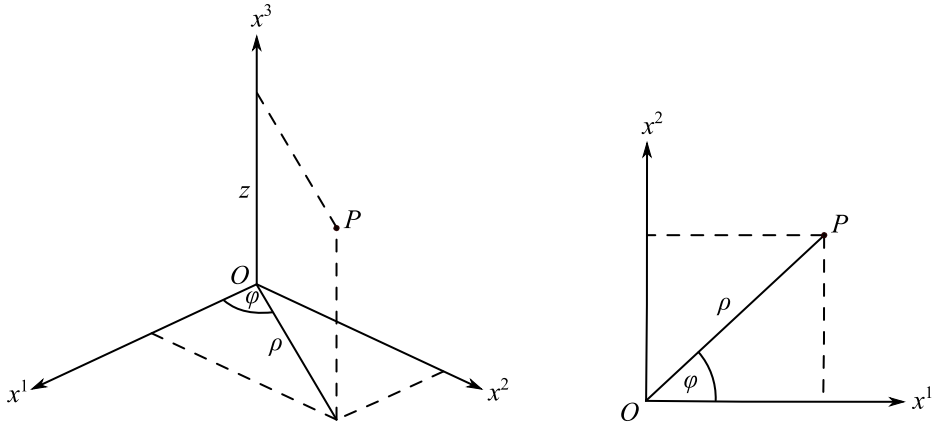
The next most commonly used system of coordinates in three dimensions is **cylindrical coordinates** (which become **polar coordinates** in the x–y plane), also known as **cylindrical polar coordinates**. They are denoted by (ρ, φ, z) and are related to the Cartesian coordinates via

$$\begin{aligned} x^1 &= \rho\cos\varphi \\ x^2 &= \rho\sin\varphi \\ x^3 &= z, \end{aligned} \tag{1.1}$$

and their inverse

$$\begin{aligned} \rho^2 &= \left(x^1\right)^2 + \left(x^2\right)^2 \\ \tan\varphi &= \frac{x^2}{x^1} \\ z &= x^3, \end{aligned} \tag{1.2}$$

as is easy to confirm by inspecting Fig. [1.3]. Cylindrical coordinates are a particularly useful choice in physics when the problem under study has cylindrical symmetry, i.e. studies a situation with cylindrical shapes or a reasonable approximation thereof. The name comes from the fact that surfaces of constant ρ describe infinite cylinders spanning the entire z direction. Surfaces of constant z are of course still infinite planes, and so are surfaces of constant φ (Fig. [1.4]).



(a) Cylindrical polar coordinates.

(b) Planar polar coordinates.

Fig. 1.3: Two- and three-dimensional cylindrical polar coordinates.

An important difference between the cylindrical (as well as two-dimensional polar) and Cartesian coordinate systems is that the cylindrical system does *not* cover all of three-dimensional space. What this means is that while any point in space can be described by a specific set of numbers in Cartesian coordinates $\left(x^1, x^2, x^3\right)$, in the cylindrical coordinate system there is a single point, out of the infinity of points in space, that *cannot* be described by a unique set of the three numbers (ρ, φ, z). This is the point located exactly at the origin of coordinates, which we usually denote by O. While the coordinates of O in the Cartesian system are simply $(0, 0, 0)$, in the cylindrical system only $\rho = 0$ and $z = 0$ can be defined but φ remains ambiguous, because *any* value can be used for it. In other words, there is no unique set of three

numbers describing the location of O; the set of numbers $(\rho = 0, \varphi = 0°, z = 0)$ does imply the origin, but so do the sets $(\rho = 0, \varphi = 10°, z = 0)$, $(\rho = 0, \varphi = 20°, z = 0)$, $(\rho = 0, \varphi = 312.56°, z = 0)$, and so on. There is an infinite number of possibilities, all describing a single point! This phenomenon is a *defect* in the coordinate system itself and cannot be remedied; the system simply fails at that specific point. One then says that while Cartesian coordinates *cover*, *map*, or *chart* all three-dimensional space, cylindrical polar coordinates do not; exactly one point defies definition. Coordinate systems in this regard can be thought of as a grid covering three-dimensional space; the Cartesian grid or map is infinite, smooth, and complete, while the polar one has a "hole" in it located at O. If one then wishes to do a problem that studies the point at the origin using such a "defective" coordinate system, one has no choice but to shift the coordinate system itself such that the point in question no longer coincides with the origin. There is simply no other way. This issue also further emphasizes that spatial points and coordinate maps are *not* the same thing; an obvious issue, perhaps, but one that is sometimes confused in the minds of beginners.

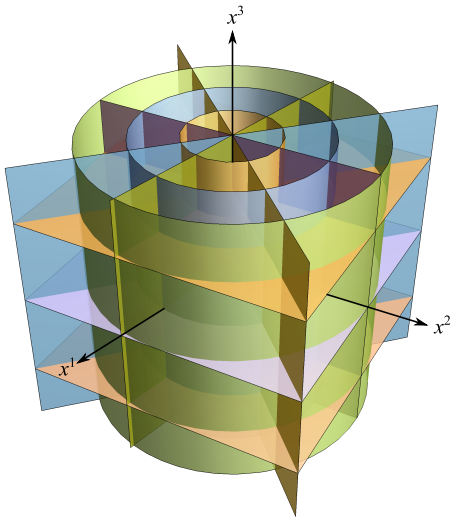


Fig. 1.4: The cylindrical coordinates grid: cylinders of constant ρ and planes of constant z and φ.

The remaining most commonly used system of coordinates in three-dimensional space is the **spherical coordinate system**, particularly useful in physics when problems exhibit spherical symmetry. As shown in Fig. [1.5], we define the numbers (r, θ, φ). They are related to the Cartesian coordinates as follows:

$$\begin{aligned} x^1 &= r \sin\theta \cos\varphi \\ x^2 &= r \sin\theta \sin\varphi \\ x^3 &= r \cos\theta \end{aligned} \tag{1.3}$$

$$\begin{aligned} r^2 &= \left(x^1\right)^2 + \left(x^2\right)^2 + \left(x^3\right)^2 \\ \tan\varphi &= \frac{x^2}{x^1} \\ \cos\theta &= \frac{x^3}{r}. \end{aligned} \tag{1.4}$$

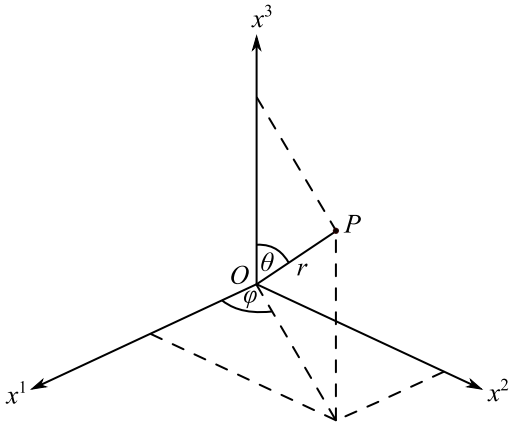


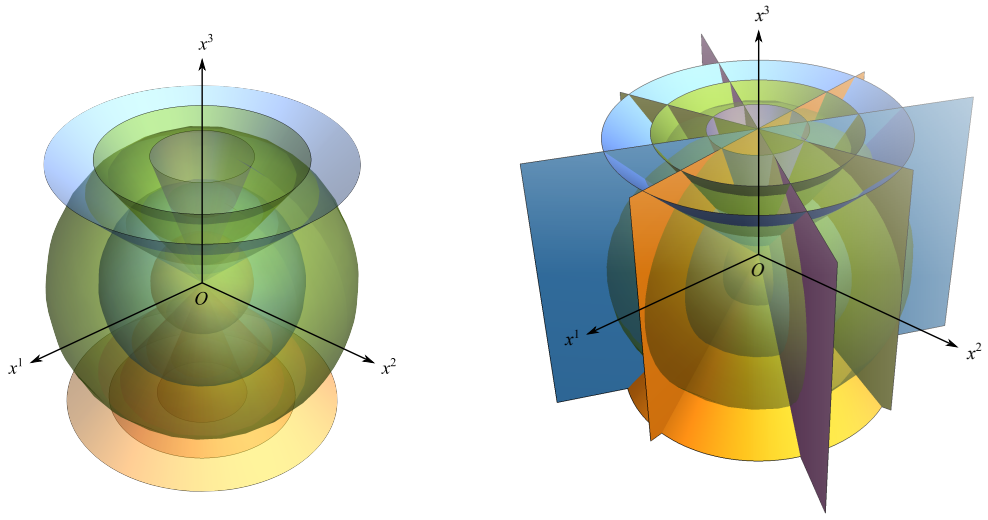
Fig. 1.5: Spherical coordinates.

It is also possible to define transformation relations between cylindrical coordinates and spherical coordinates. These are:

$$\begin{aligned} \rho &= r\sin\theta \\ z &= r\cos\theta \\ r^2 &= \rho^2 + z^2 \\ \tan\theta &= \frac{\rho}{z}. \end{aligned} \tag{1.5}$$

Note that the spherical coordinate system also suffers from the same defect we noted for cylindrical coordinates. Once again, the point at the origin cannot be specified by three unique numbers: only $r = 0$ can be used, while *both* θ and φ are ambiguous. As before, the only way to deal with this in calculations is to avoid addressing the origin altogether or just shift the coordinate system. Now the defining property of spherical coordinates is that surfaces of constant r describe spheres, concentric at the origin, while surfaces of constant θ and φ define cones and flat surfaces respectively, as shown in Fig [1.6].

As noted, coordinate systems are defined based on geometric shapes: flat planes for Cartesian, cylinders for cylindrical (or circles for polar), and spheres for spherical. We note here that these are *not* the only options; other shapes can also be used to



(a) Spheres of constant r and cones of constant θ in spherical coordinates.

(b) The complete spherical coordinates grid.

Fig. 1.6: Spherical coordinates.

define coordinate systems. For example, consider the so-called **elliptic cylindrical coordinates** (u, v, z), where surfaces of constant u define concentric elliptic cylinders and surfaces of constant v are hyperbolas, as shown in Fig. [1.7]. They are related to the Cartesian coordinates via

$$\begin{aligned} x^1 &= a \cosh u \cos v \\ x^2 &= a \sinh u \sin v \\ x^3 &= z, \end{aligned} \tag{1.6}$$

where a is an arbitrary constant.

There are coordinate systems based on the parabolic shape, such as **parabolic cylindrical coordinates**, as shown in Fig. [1.8]:

$$\begin{aligned} x^1 &= \xi\eta \\ x^2 &= \frac{1}{2}\left(\eta^2 - \xi^2\right) \\ x^3 &= z. \end{aligned} \tag{1.7}$$

A final example is the **parabolic coordinates**, as shown in Fig. [1.9]:

$$\begin{aligned} x^1 &= uv \cos\theta \\ x^2 &= uv \sin\theta \\ x^3 &= \frac{1}{2}\left(u^2 - v^2\right). \end{aligned} \tag{1.8}$$

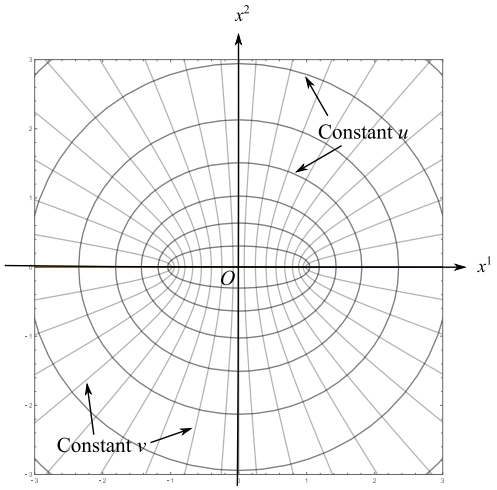


Fig. 1.7: The elliptic cylindrical coordinates grid in the x–y plane: curves of constant u (ellipses) and v (hyperbolas) in the $z = 0$ plane. The extension to three dimensions is straightforward.

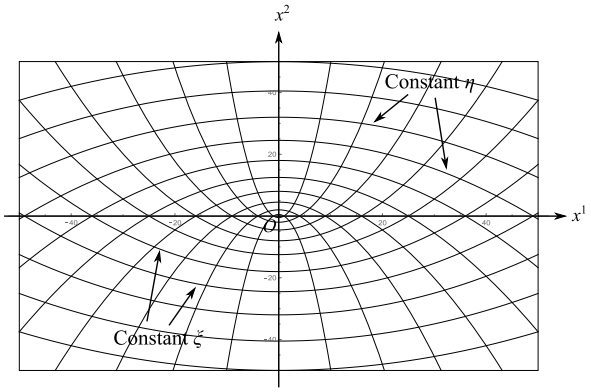


Fig. 1.8: The parabolic cylindrical coordinates grid in the x–y plane: parabolic curves of constant ξ (upright) and η (inverted) in the $z = 0$ plane. The extension to three dimensions is straightforward.

All of the systems of coordinates we discussed, and many others that exist, are collectively known as **orthogonal coordinate systems**, i.e. perpendicular. In other words, their respective grid lines are perpendicular to each other at every point. There do exist non-orthogonal, or **skew**, coordinate systems, a simple example of which is given in the exercises.

Exercise 1.1 Using a computer, see if you can reproduce some of the coordinate grids in this section, such as Fig. [1.4], [1.7], or [1.9]. Once you have had enough practice, plot the

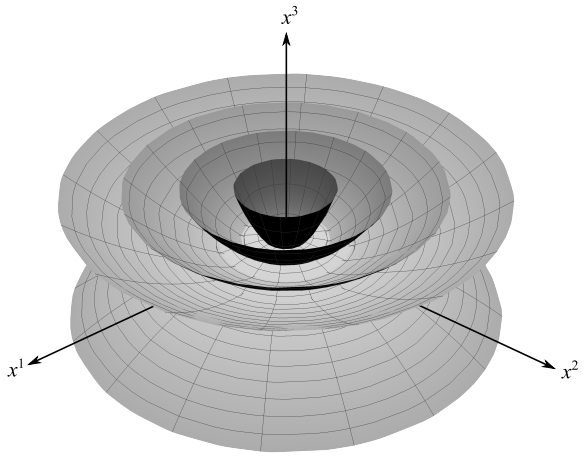


Fig. 1.9: The parabolic coordinate grid: The paraboloidal surfaces of constant v (upright) and u (inverted) are shown. They can be thought of as the surfaces of revolution of the parabolas of constant ξ and η respectively in the parabolic cylindrical coordinates.

complete coordinates grids for the following systems. For the numerical values of the given constants, you may wish to experiment with a few and see which ones give you the best plots.

1. Paraboloidal coordinates $(\mu,\, \nu,\, \lambda)$ defined by

$$\begin{aligned} x^2 &= \frac{4}{a-b}\,(\mu-a)\,(a-\nu)\,(a-\lambda) \\ y^2 &= \frac{4}{a-b}\,(\mu-b)\,(b-\nu)\,(\lambda-b) \\ z &= \mu+\nu+\lambda-a-b, \end{aligned} \tag{1.9}$$

   where the constants a and b set the following limits on the coordinates:

$$\mu > a > \lambda > b > \nu > 0.$$

2. Oblate spheroidal coordinates $(\mu,\, \nu,\, \varphi)$ defined by

$$\begin{aligned} x &= a\cosh\mu\cos\nu\cos\varphi \\ y &= a\cosh\mu\cos\nu\sin\varphi \\ z &= a\sinh\mu\sin\nu, \end{aligned} \tag{1.10}$$

   where a is an arbitrary constant.
3. Toroidal coordinates $(\sigma,\, \tau,\, \varphi)$ defined by

$$\begin{aligned} x &= \frac{a\sinh\tau}{\cosh\tau-\cos\sigma}\cos\varphi \\ y &= \frac{a\sinh\tau}{\cosh\tau-\cos\sigma}\sin\varphi \\ z &= \frac{a\sin\sigma}{\cosh\tau-\cos\sigma}, \quad \text{where } a \text{ is an arbitrary constant.} \end{aligned} \tag{1.11}$$

Exercise 1.2 As an example of non-orthogonal, or skew, coordinates, consider a Cartesian coordinate system where the z axis has been rotated in the y–z plane by an angle α, as shown in Fig. [1.10]. Derive the transformation equations relating the original orthogonal x, y, and z coordinates to the skew coordinates x', y', and z'.

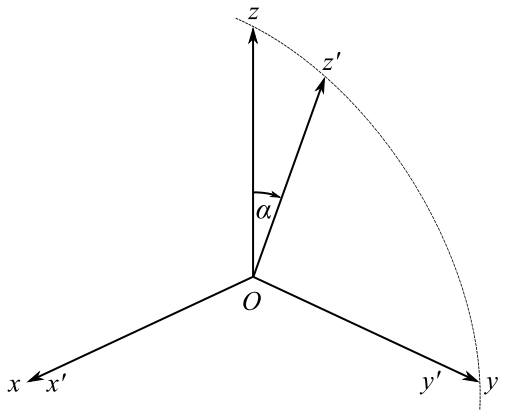


Fig. 1.10: The simplest possible example of a non-orthogonal, or skew, coordinate system.

1.2 Measurements and the Metric

A fundamental question in geometry is the question of spatial measurement—specifically: How does one measure a distance in space and relate it to the coordinate system in use? In Cartesian coordinates this is fully described by exploiting the familiar **Pythagorean theorem**.[3] Consider, to start, the two-dimensional plane with Cartesian coordinates shown in Fig. [1.11]. We define Δl as the distance between two specific points with coordinates (x^1, x^2) and $(x^{1'}, x^{2'})$. The question is: How is Δl related to the changes in coordinates $\Delta x^1 = x^{1'} - x^1$ and $\Delta x^2 = x^{2'} - x^2$ between the points in question? This is clearly a problem for Pythagoras:

$$\Delta l^2 = \left(\Delta x^1\right)^2 + \left(\Delta x^2\right)^2 . \tag{1.12}$$

Typically, we would be more interested in *infinitesimal* changes of coordinates (if we wanted to do calculus), so making the distance Δl arbitrarily small; i.e. $\Delta l \to dl$ gives

$$dl^2 = \left(dx^1\right)^2 + \left(dx^2\right)^2 . \tag{1.13}$$

[3] Pythagoras: Greek philosopher and mathematician (571–495 BCE).

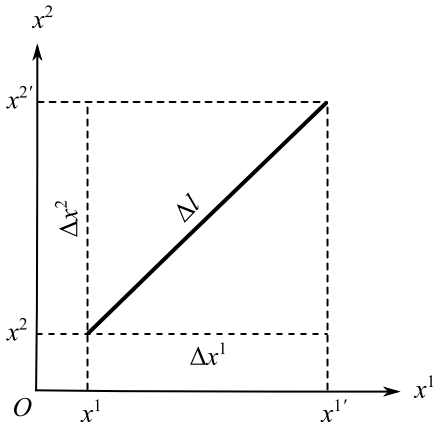


Fig. 1.11: The Pythagorean theorem.

The explicit use of indices in (1.13) is clearly becoming cumbersome. But this is exactly where the index notation becomes truly powerful. To see this, let us begin by using the summation symbol Σ in (1.13):

$$dl^2 = \sum_{i=1}^{2} \left(dx^i\right)^2. \tag{1.14}$$

We can easily generalize to three dimensions by simply adding an extra term:

$$dl^2 = \left(dx^1\right)^2 + \left(dx^2\right)^2 + \left(dx^3\right)^2 \tag{1.15}$$

$$= \sum_{i=1}^{3} \left(dx^i\right)^2. \tag{1.16}$$

Interestingly, if, for whatever reason, one imagines a higher-dimensional space, in, say, arbitrary N dimensions, then (1.15) is easily generalized by just adding more terms x^4, x^5, all the way up to x^N,

$$dl^2 = \left(dx^1\right)^2 + \left(dx^2\right)^2 + \left(dx^3\right)^2 + \left(dx^4\right)^2 + \left(dx^5\right)^2 + \cdots + \left(dx^N\right)^2, \tag{1.17}$$

or just changing the summation's upper limit to whatever the value of N is:

$$dl^2 = \sum_{i=1}^{N} \left(dx^i\right)^2. \tag{1.18}$$

So the same formula (1.18) can be used to represent the Pythagorean theorem in *any* number of dimensions. All one needs to do is set the value of N. This simplicity

is one reason why the index notation is a preferred choice. Now, expressions such as $\left(dx^i\right)^2$ may be a bit confusing. A common method used to avoid such confusion is to write (1.18) as follows:

$$dl^2 = \sum_{i=1}^{N} dx^i dx^i, \tag{1.19}$$

which should help a little. However, let us take this quest for simplification even further: Summations such as (1.19) arise all the time—sometimes double, triple, or more summations (as we will see). They are used to compactly write what would otherwise be very long equations (sometimes *pages* long). This is very common in both non-relativistic and relativistic physics. As he was pondering his theory of relativity, Albert Einstein himself came up with another simplification to expressions such as (1.19). This is now known as the **Einstein summation convention**, a particularly simple idea that goes a long way in simplifying the form of the equations. This is done by removing the summation symbol entirely:

$$dl^2 = dx^i dx^i. \tag{1.20}$$

The summation is hidden but still exists; eqn (1.20) is identical in meaning to (1.19).[4] But when one encounters such an expression, how does one know whether or not there is a summation if it is not explicitly written? The rule is this: If there are two *identical* indices per term, then a summation over them exists (*exactly* two; no more, no less). Either the limits of the summation are known beforehand or one can simply express them like this:

$$dl^2 = dx^i dx^i, \qquad i = 1, 2, \ldots N, \tag{1.21}$$

where N can be 2 (1.13), 3 (1.15), or just arbitrary. This language is very commonly used in theoretical physics and has more rules to it that we will explore later. On first encounter, an equation such as (1.21) may cause a great deal of confusion to the reader, so it may be advisable to go back and read from the beginning of this section just to remind oneself that (1.21) is nothing more than a shorthand for the Pythagorean theorem.

The quantity dl is a distance between two (infinitesimally close) points, so as far as pure geometry is concerned it is always **positive definite**, as a distance should be. Generally, dl is known as **the line element**. It is also sometimes referred to as the **metric** (from Latin *metricus*, and Greek *metrikos* or *metron*, meaning "measure"). In differential geometry and multivariable calculus dl represents the basic element for measuring distances on curves. A familiar equation that arises from it is the formula for distance between two points on a given curve in the x–y plane which one encounters in calculus:

$$l = \int_{x=a}^{x=b} dx \sqrt{1 + \left(\frac{dy}{dx}\right)^2}, \tag{1.22}$$

[4] I imagine Einstein simply got tired of writing summation symbols all the time.

where in this context $y\,(x)$ is the function describing the curve whose length is being measured. For our purposes, however, the quantities l or dl are unessential; it is dl^2 that is more important, for reasons that will become clear later. Hence, despite the possible notational confusion this might cause, in physics textbooks the terms "line element" and "metric" are more commonly used in reference to dl^2, *not* dl.

Now, what about other coordinate systems? What does the Pythagorean theorem look like in, say, two-dimensional polar coordinates? This is easily found; simply consider Fig. [1.12]. Infinitesimally, the shaded area approaches a right triangle. Applying the Pythagorean theorem to it gives

$$dl^2 = d\rho^2 + \rho^2 d\varphi^2. \tag{1.23}$$

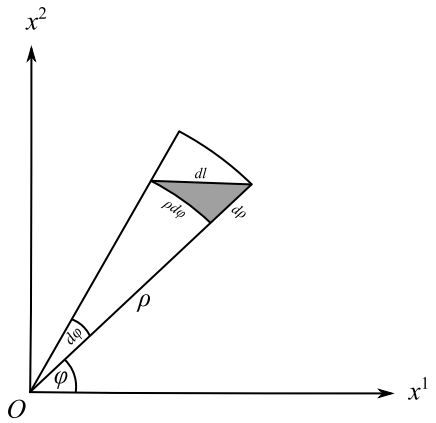


Fig. 1.12: The Pythagorean theorem applied to dl in planar polar coordinates.

Now, we could have found eqn (1.23) analytically if we had considered how Cartesian coordinates are related to polar coordinates, i.e. by using the transformations (1.1). Simply put, take the differentials of both sides of (1.1), then plug the result into (1.13). Explicitly,

$$\begin{aligned} dx^1 &= d\rho\cos\varphi - \rho d\varphi\sin\varphi \\ dx^2 &= d\rho\sin\varphi + \rho d\varphi\cos\varphi, \end{aligned} \tag{1.24}$$

leading to

$$\begin{aligned} dl^2 &= \left(dx^1\right)^2 + \left(dx^2\right)^2 \\ &= d\rho^2\cos^2\varphi + \rho^2 d\varphi^2\sin^2\varphi - 2\rho\, d\rho\, d\varphi\sin\varphi\cos\varphi \\ &\quad + d\rho^2\cos^2\varphi + \rho^2 d\varphi^2\sin^2\varphi + 2\rho\, d\rho\, d\varphi\sin\varphi\cos\varphi \\ &= \left(d\rho^2 + \rho^2 d\varphi^2\right)\left(\sin^2\varphi + \cos^2\varphi\right) \\ &= d\rho^2 + \rho^2 d\varphi^2, \end{aligned} \tag{1.25}$$

where we have used Euler's[5] extremely useful identity $\sin^2\varphi+\cos^2\varphi=1$, itself another form of Pythagoras' theorem. Cylindrical coordinates are a straightforward extension; just add dz^2:

$$dl^2 = d\rho^2 + \rho^2 d\varphi^2 + dz^2. \tag{1.26}$$

A similar calculation, albeit a bit longer, leads to the equivalent spherical coordinates expression by either plugging (1.3) into (1.15) or (1.5) into (1.26) to find the line element:

$$dl^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\varphi^2. \tag{1.27}$$

Exercise 1.3 Find the metric dl^2 using the analytical method for:

1. Spherical coordinates using (1.3).
2. Spherical coordinates using (1.5).
3. Elliptic cylindrical coordinates (1.6).
4. Parabolic cylindrical coordinates (1.7).
5. The skew coordinates system of Exercise 1.2.

The analytical method just described is extremely useful because it provides a systematic way of transforming or generating metrics from one system of coordinates to another without having to draw sketches and agonizing over the trigonometry. We can easily generalize the method to any arbitrary system of coordinates (u, v, w) whose relationship to the Cartesian (or any other coordinate system for that matter) is known; i.e. one can write the Cartesian coordinates as functions in, or transformations to, the new coordinates, $x^i(u, v, w)$, such as (1.1) or (1.3). To find the metric in the new coordinate system all one needs to do is take the appropriate derivatives and plug into (1.15). For this purpose let's define the so-called **scale factors** h_u, h_v, and h_w:

$$\begin{aligned} h_u^2 &= \left(\frac{\partial x^1}{\partial u}\right)^2 + \left(\frac{\partial x^2}{\partial u}\right)^2 + \left(\frac{\partial x^3}{\partial u}\right)^2 \\ h_v^2 &= \left(\frac{\partial x^1}{\partial v}\right)^2 + \left(\frac{\partial x^2}{\partial v}\right)^2 + \left(\frac{\partial x^3}{\partial v}\right)^2 \\ h_w^2 &= \left(\frac{\partial x^1}{\partial w}\right)^2 + \left(\frac{\partial x^2}{\partial w}\right)^2 + \left(\frac{\partial x^3}{\partial w}\right)^2. \end{aligned} \tag{1.28}$$

If we define a parameter k that can be either u, v, or w, we can more compactly write

$$h_k^2 = \left(\frac{\partial x^i}{\partial k}\right)\left(\frac{\partial x^i}{\partial k}\right), \tag{1.29}$$

[5]Leonhard Euler: Swiss mathematician, physicist, astronomer, geographer, logician, and engineer (1707–1783).

with the summation convention on i assumed. These scale factors lead directly to the metric/line element in the (u, v, w) coordinate system:

$$dl^2 = h_u^2 du^2 + h_v^2 dv^2 + h_w^2 dw^2. \tag{1.30}$$

For example, for spherical coordinates (1.3)

$$\begin{aligned} h_r^2 &= \sin^2\theta\cos^2\varphi + \sin^2\theta\sin^2\varphi + \cos^2\theta = 1 \\ h_\theta^2 &= r^2\cos^2\theta\cos^2\varphi + r^2\cos^2\theta\sin^2\varphi + r^2\sin^2\theta = r^2 \\ h_\varphi^2 &= r^2\sin^2\theta\sin^2\varphi + r^2\sin^2\theta\cos^2\varphi = r^2\sin^2\theta, \end{aligned} \tag{1.31}$$

which leads to (1.27) as required. Note that the scale factors method, as described, works for orthogonal coordinates *only*.

Exercise 1.4 Find the metric dl^2 using the scale factors method for:

1. Parabolic coordinates (1.8).
2. Paraboloidal coordinates (1.9).
3. Oblate spheroidal coordinates (1.10).
4. Toroidal coordinates (1.11).

Exercise 1.5 Consider the following metric:

$$dl^2 = du^2 + dv^2 + dw^2 - \left(\frac{3}{13}du + \frac{4}{13}dv + \frac{12}{13}dw\right)^2. \tag{1.32}$$

It appears that this is a three-dimensional metric in u, v, and w. This is not true! This metric is in fact two-dimensional in disguise. Prove this by finding a transformation between the "coordinates" u, v, and w and the Cartesian coordinates x and y, similar to (1.1), for example, that reduces (1.32) to

$$dl^2 = dx^2 + dy^2. \tag{1.33}$$

1.3 Vectors in Cartesian Coordinates

In introductory physics courses, vectors are defined as physical quantities that have both magnitude and direction. They are abstractedly represented by directional arrows whose lengths signify their magnitudes. This description, while correct, suffers from the defect that the visual image that goes along with it is only useful in two or three dimensions. In higher-dimensional physics, which we will consider soon, one can still define vectors in the same way, but the visual is lost. This is not because it doesn't exist, but because our brains are not built to visualize more than three dimensions. It is also true that the concept of vectors generalizes into a variety of more exotic mathematical objects that go by the names "tensors," "differential forms," and "spinors," all of

which are quite useful in theoretical physics.[6] These objects, by their very nature, resist simple visualization except in the most restricted of special cases or analogies. Given all of this, we need to redefine the language used to describe vectors to make the transition to higher dimensions, as well as to the related objects, as smooth as possible, while at the same time preserving the coordinate-invariant nature of vector mathematics. We can still imagine, as well as sketch, vectors as "arrows" in space, but the emphasis shifts toward the calculational over the graphical.

To start, let's define vectors and their components in Cartesian coordinates in the usual way:

$$\mathbf{V} = V_x\hat{\mathbf{x}} + V_y\hat{\mathbf{y}} + V_z\hat{\mathbf{z}}, \tag{1.34}$$

where $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ are called the **Cartesian basis vectors**, usually defined as unit vectors pointing in the directions of the x, y, and z axes respectively.[7] Addition and subtraction of vectors is defined pictorially as usual by completing the triangle formed by two vectors to find the result, as shown in Fig. [1.13].

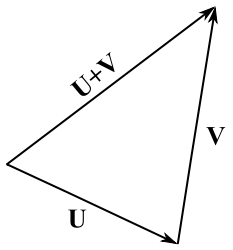


Fig. 1.13: Addition of vectors.

So, given two vectors $\mathbf{V}$ and $\mathbf{U} = U_x\hat{\mathbf{x}} + U_y\hat{\mathbf{y}} + U_z\hat{\mathbf{z}}$, their sum is

$$\mathbf{V} + \mathbf{U} = (V_x + U_x)\,\hat{\mathbf{x}} + (V_y + U_y)\,\hat{\mathbf{y}} + (V_z + U_z)\,\hat{\mathbf{z}}. \tag{1.35}$$

Furthermore, the so-called **dot product** of two vectors $\mathbf{V}$ and $\mathbf{U}$, also known as the **scalar** or **inner** product, is defined as follows:

$$\mathbf{V}\cdot\mathbf{U} = V_xU_x + V_yU_y + V_zU_z, \tag{1.36}$$

where the result no longer carries the vector status but is rather a **scalar**; a quantity with magnitude only. Another equivalent definition of the dot product is

$$\mathbf{V}\cdot\mathbf{U} = |\mathbf{V}|\,|\mathbf{U}|\cos\theta, \tag{1.37}$$

where θ is the angle between $\mathbf{V}$ and $\mathbf{U}$, as shown in Fig. [1.14].

[6] Two of which, tensors and differential forms, we will study in this book.

[7] In introductory physics textbooks the notation $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$, and $\hat{\mathbf{k}}$ is more commonly used.

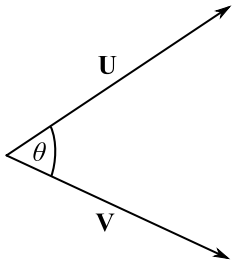


Fig. 1.14: The scalar or dot product of two vectors.

Now the magnitude of a vector, also known as its **norm**, is defined as the square root of its dot product with itself:

$$V = |\mathbf{V}| = \sqrt{\mathbf{V} \cdot \mathbf{V}} = \sqrt{V_x^2 + V_y^2 + V_z^2}. \tag{1.38}$$

Note that we define everything in the most general sense, so the vectors considered can be functions in space as well as time—in other words, they can be **vector fields**, i.e. $\mathbf{V}(t, \mathbf{r})$ with components $V_x(t, x, y, z)$, $V_y(t, x, y, z)$, and $V_z(t, x, y, z)$.

The main language needed to achieve our objectives in this chapter is exactly the index notation briefly introduced in §1.2. As pointed out earlier, it is also sometimes referred to as the tensor notation, since it becomes truly powerful when dealing with tensors, which we will introduce later. For now let's see how Cartesian vectors can be recast in indices. Consider the definition (1.34). Rename the basis vectors as follows:

$$\hat{\mathbf{x}} \to \hat{\mathbf{e}}_1, \qquad \hat{\mathbf{y}} \to \hat{\mathbf{e}}_2, \qquad \hat{\mathbf{z}} \to \hat{\mathbf{e}}_3. \tag{1.39}$$

Also rename the components of the vector

$$V_x \to V^1, \qquad V_y \to V^2, \qquad V_z \to V^3, \tag{1.40}$$

where once again it is extremely important *not* to confuse the superscripts with exponents. The numbers 1, 2, and 3 simply keep track of which component of the vector $\mathbf{V}$ we are talking about. We also note that we gave the basis vectors subscripts rather than superscripts. At this point this is just a matter of convention, but it will acquire a deeper meaning later.

It is now clearly possible to write the vector $\mathbf{V}$ in the following compact form:

$$\mathbf{V} = V^1\hat{\mathbf{e}}_1 + V^2\hat{\mathbf{e}}_2 + V^3\hat{\mathbf{e}}_3 = \sum_{i=1}^{3} V^i\hat{\mathbf{e}}_i. \tag{1.41}$$

Alternatively, by adopting the Einstein summation convention (i.e. dropping the summation symbol) we get the elegant shorthand

$$\mathbf{V} = V^i \hat{\mathbf{e}}_i. \tag{1.42}$$

Recall that the summation is still there, it is just hidden; once again, the rule is: ***If an index appears twice in one term, it is summed over***. Here is another example: the vector addition formula (1.35) can now be written as follows:

$$\mathbf{V} + \mathbf{U} = V^i \hat{\mathbf{e}}_i + U^i \hat{\mathbf{e}}_i = \left(V^i + U^i\right) \hat{\mathbf{e}}_i. \tag{1.43}$$

Note that which letter to use as an index in a summation is *completely* arbitrary. For example, $V^j \hat{\mathbf{e}}_j$ or $V^a \hat{\mathbf{e}}_a$ (or even $U^\heartsuit \hat{\mathbf{e}}_\heartsuit$ if one wishes) still means *exactly* the same thing. In fact, it is perfectly acceptable, as well as useful sometimes, to change the letters of the indices in mid calculation. Consider that we can write

$$\mathbf{V} + \mathbf{U} = V^i \hat{\mathbf{e}}_i + U^j \hat{\mathbf{e}}_j = \left(V^i + U^i\right) \hat{\mathbf{e}}_i \tag{1.44}$$

and it is still equivalent to (1.43), even though we changed the index j into i in the last step. Because of this freedom, summation indices are sometimes referred to as **dummy indices**. There is another type of index that we will introduce soon.

The process of performing summations using the Einstein convention is known as **index contraction**, so (1.42) may be described as "V^i contracted with $\hat{\mathbf{e}}_i$," which is exactly the same thing as saying multiply each individual V^i with its respective $\hat{\mathbf{e}}_i$ and sum them up. For ordinary two- or three-dimensional physics, the indices will count over either $1, 2$ or $1, 2, 3$ depending on whether we are looking at two-dimensional vectors in the plane or the full three-dimensional structure. For example, the position vector

$$\mathbf{r} = x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}} \tag{1.45}$$

can now be rewritten as

$$\mathbf{r} = x^i \hat{\mathbf{e}}_i, \tag{1.46}$$

where $x = x^1$, $y = x^2$, and $z = x^3$ as before. We now use this language to reconstruct everything we have ever learned about vectors. Note that we will be working purely in Cartesian coordinates for a while. Attention to vectors in curvilinear coordinates will be given eventually.

Suppose we want to find the dot product of the two vectors

$$\mathbf{V} = V^i \hat{\mathbf{e}}_i, \qquad \mathbf{U} = U^i \hat{\mathbf{e}}_i \tag{1.47}$$

by putting them together as usual:

$$\mathbf{V} \cdot \mathbf{U} = \left(V^i \hat{\mathbf{e}}_i\right) \cdot \left(U^i \hat{\mathbf{e}}_i\right). \qquad \text{Wrong!} \tag{1.48}$$

We remember that we are not allowed to have more than two repeating indices in any one term of an equation, since this should be a double sum: one summation is

for vector $\mathbf{V}$ and the other for $\mathbf{U}$. As it stands now, expression (1.48) implies a single sum, which is generally incorrect. To solve this problem we use the dummy property of the summation indices and give one of them a different name, signifying a different summation:

$$\mathbf{V} \cdot \mathbf{U} = \left(V^i \hat{\mathbf{e}}_i\right) \cdot \left(U^j \hat{\mathbf{e}}_j\right), \tag{1.49}$$

where the index j also runs from 1 to 3. In other words, this is equivalent to

$$\mathbf{V} \cdot \mathbf{U} = \sum_{i=1}^{3} \sum_{j=1}^{3} \left[\left(V^i \hat{\mathbf{e}}_i\right) \cdot \left(U^j \hat{\mathbf{e}}_j\right)\right]. \tag{1.50}$$

Now, let's continue by reordering:

$$\mathbf{V} \cdot \mathbf{U} = V^i U^j \left(\hat{\mathbf{e}}_i \cdot \hat{\mathbf{e}}_j\right). \tag{1.51}$$

If an equation such as (1.51), or any other subsequent equation for that matter, is not immediately obvious, then the reader is advised to write it out in detail to see how the summation convention works. In this case (1.51) is simply shorthand for

$$\begin{aligned} \mathbf{V} \cdot \mathbf{U} = {} & V^1 U^1 \left(\hat{\mathbf{e}}_1 \cdot \hat{\mathbf{e}}_1\right) + V^1 U^2 \left(\hat{\mathbf{e}}_1 \cdot \hat{\mathbf{e}}_2\right) + V^1 U^3 \left(\hat{\mathbf{e}}_1 \cdot \hat{\mathbf{e}}_3\right) \\ & + V^2 U^1 \left(\hat{\mathbf{e}}_2 \cdot \hat{\mathbf{e}}_1\right) + V^2 U^2 \left(\hat{\mathbf{e}}_2 \cdot \hat{\mathbf{e}}_2\right) + V^2 U^3 \left(\hat{\mathbf{e}}_2 \cdot \hat{\mathbf{e}}_3\right) \\ & + V^3 U^1 \left(\hat{\mathbf{e}}_3 \cdot \hat{\mathbf{e}}_1\right) + V^3 U^2 \left(\hat{\mathbf{e}}_3 \cdot \hat{\mathbf{e}}_2\right) + V^3 U^3 \left(\hat{\mathbf{e}}_3 \cdot \hat{\mathbf{e}}_3\right). \end{aligned} \tag{1.52}$$

Now we have known since introductory physics that the expression $\left(\hat{\mathbf{e}}_i \cdot \hat{\mathbf{e}}_j\right)$ can only be either 1 or zero, because these are orthogonal (i.e. perpendicular) unit vectors. So, if $i = j$, then $\left(\hat{\mathbf{e}}_i \cdot \hat{\mathbf{e}}_j\right) = 1$, while for $i \neq j$ we have $\left(\hat{\mathbf{e}}_i \cdot \hat{\mathbf{e}}_j\right) = 0$ (previously known as $\hat{\mathbf{x}} \cdot \hat{\mathbf{x}} = \hat{\mathbf{y}} \cdot \hat{\mathbf{y}} = \hat{\mathbf{z}} \cdot \hat{\mathbf{z}} = 1$ and $\hat{\mathbf{x}} \cdot \hat{\mathbf{y}} = \hat{\mathbf{y}} \cdot \hat{\mathbf{z}} = \hat{\mathbf{z}} \cdot \hat{\mathbf{x}} = 0$). So applying this to (1.52) leads to

$$\mathbf{V} \cdot \mathbf{U} = V^1 U^1 + V^2 U^2 + V^3 U^3, \tag{1.53}$$

as required. The orthogonality property of the basis vectors allows us to symbolically incorporate this information in (1.51) by introducing the following new symbol:

$$\delta_{ij} = \hat{\mathbf{e}}_i \cdot \hat{\mathbf{e}}_j \quad \text{such that} \quad \delta_{ij} \begin{cases} = 1 \text{ for } i = j \\ = 0 \text{ for } i \neq j \end{cases}. \tag{1.54}$$

The symbol δ_{ij} is known as the **Kronecker delta**,[8] and can be understood at this point as a "book keeping" method of deciding which expressions are vanishing and which aren't. We can then rewrite (1.51):

$$\mathbf{V} \cdot \mathbf{U} = V^i U^j \delta_{ij}. \tag{1.55}$$

[8]Leopold Kronecker: Prussian/German mathematician (1823–1891).

This is a double sum over i and j, which, if we perform faithfully, gives

$$\begin{aligned}\mathbf{V}\cdot\mathbf{U} &= V^iU^j\delta_{ij} = \sum_{i=1}^{3}\sum_{j=1}^{3} V^iU^j\delta_{ij}\\ &= V^1U^1\delta_{11} + V^1U^2\delta_{12} + V^1U^3\delta_{13} + V^2U^1\delta_{21} + V^2U^2\delta_{22} + V^2U^3\delta_{23}\\ &\quad + V^3U^1\delta_{31} + V^3U^2\delta_{32} + V^3U^3\delta_{33}\\ &= V^1U^1 + V^2U^2 + V^3U^3, \quad \text{since only } \delta_{11},\ \delta_{22}, \text{ and } \delta_{33} \text{ are non-vanishing.}\end{aligned} \tag{1.56}$$

The use of the Kronecker delta is extremely useful, as expressions tend toward more complicated index manipulations. In reference to eqn (1.20), for example, one can rewrite the Cartesian metric as follows:

$$dl^2 = \delta_{ij}dx^i dx^j. \tag{1.57}$$

The reader can verify that it means exactly the same thing as before.

Exercise 1.6 Expand (1.57) and show that it leads to (1.15).

Now, as briefly discussed earlier, the fact that we are using both upper and lower indices may be slightly confusing at this point. As far as summations are concerned, which is the reason the index language was invented in the first place, it apparently doesn't matter whether we use upper or lower indices. So the following expressions are essentially all equivalent:

$$V^iU^j\delta_{ij} = V_iU_j\delta^{ij} = V^iU^j\delta^{ij} = V_iU_j\delta_{ij}. \tag{1.58}$$

The Kronecker delta does not change meaning, either; for instance, $\delta^{22} = \delta_{22} = \delta_2^2 = 1$ and $\delta^{23} = \delta_{23} = \delta_3^2 = 0$, and so on. So if they are all the same, why are we making a distinction between quantities with upper indices and quantities with lower indices? The reason is that quantities with upper or lower indices are equivalent to each other *only* in Cartesian coordinates! When we make the transition to non-Cartesian coordinates we will find that vectors with upper indices and vectors with lower indices are no longer exactly the same. So we might as well get used to the concept of upper and lower indices from the start. At this point we can, however, define a relationship between vector components with upper indices and those with lower indices as follows. Let

$$\begin{aligned}U^1\delta_{11} &= U_1\\ U^2\delta_{22} &= U_2\\ U^3\delta_{33} &= U_3,\end{aligned} \tag{1.59}$$

or symbolically

$$U^j\delta_{ij} = U_i \quad \text{and} \quad V^i\delta_{ij} = V_j \tag{1.60}$$

for the components of any vectors **U** and **V**. Also,

$$U_j\delta^{ij} = U^i \quad \text{or} \quad V_i\delta^{ij} = V^j. \tag{1.61}$$

The process defined by (1.60) and (1.61) is usually known as the **lowering** or the **raising** of indices respectively. If we begin with an upper index component U^i, we can contract it with a Kronecker delta to define a lower index U_j. We may even take this further by defining the **hybrid Kronecker delta** as follows

$$\delta^i_j = \delta^{ik}\delta_{kj}, \tag{1.62}$$

which can be used to change the names of the indices *without* raising or lowering

$$U^j\delta^i_j = U^i, \quad U_i\delta^i_j = U_j. \tag{1.63}$$

Notice the appearance of a single non-repeating index in (1.60), (1.61), and (1.63). This is called a **free index**. ***The rule for free indices is that if there is one of them in a term of the equation, then there should be one by the same name in each other term***. They are obviously not used in summations, but rather to denote that an equation such as (1.60) is in fact three equations, one for each value of the free index; in other words, (1.60) means the same thing as all three in (1.59). Note that eqn (1.62) has *two* free indices and only one dummy index, which implies that it is in fact equivalent to nine separate equations:

$$\begin{aligned}
\delta^1_1 &= \delta^{1k}\delta_{k1} \\
\delta^2_1 &= \delta^{2k}\delta_{k1} \\
\delta^3_2 &= \delta^{3k}\delta_{k2}, \quad \text{and so on.}
\end{aligned} \tag{1.64}$$

Now in the literature a vector whose components carry upper indices is known as a **contravariant** vector, while a vector whose components carry lower indices is known as a **covariant** vector. Another terminology is in calling a lower-index vector the **dual vector**, while the upper-index vector is simply the **vector**. There is no need to further the reader's confusion, though, so we will just call them "upper and lower indices." Once again, as far as Cartesian vectors are concerned, the components are the same for both types of vectors anyway, so $U^1 = U_1$, $U^2 = U_2$, and $U^3 = U_3$. If it makes it clearer, here is how one reads a formula such as (1.60):

$$U^j\delta_{ij} = U^1\delta_{i1} + U^2\delta_{i2} + U^3\delta_{i3} = U_i. \tag{1.65}$$

Only one of the three terms in the last equation will be non-zero: the one where the free index i has the same value as the other index on the Kronecker delta, yielding U_i. Explicitly, (1.65) is equivalent to

$$\begin{aligned} U^j\delta_{1j} &= U^1\delta_{11} + U^2\delta_{12} + U^3\delta_{13} = U^1\delta_{11} = U_1 \\ U^j\delta_{2j} &= U^1\delta_{21} + U^2\delta_{22} + U^3\delta_{23} = U^2\delta_{22} = U_2 \\ U^j\delta_{3j} &= U^1\delta_{31} + U^2\delta_{32} + U^3\delta_{33} = U^3\delta_{33} = U_3. \end{aligned} \tag{1.66}$$

A similar process happens with (1.63):

$$U^j\delta^i_j = U^1\delta^i_1 + U^2\delta^i_2 + U^3\delta^i_3 = U^i. \tag{1.67}$$

So, we conclude this rather lengthy definition of the dot product by noting that, based on the preceding,

$$\mathbf{V}\cdot\mathbf{U} = V^iU^j\delta_{ij} = V_iU_j\delta^{ij} = V^iU_j\delta^j_i = V^iU_i = V_iU^i \tag{1.68}$$

are all equivalent, where the last two are derived by using either one of the equations in (1.60) and (1.61) and the dummy indices can carry any labels as usual.

The language and rules defined here are extremely important for the rest of the book. Yes, it is a lot to absorb all at once, but the reader is *strongly* advised not to move forward without deeply understanding everything mentioned in this section so far. In the hope of making all of this slightly more intuitive, let's take it one step further. The rather funny, and potentially confusing, upper and lower index notation is closely related to something you may already be familiar with.[9] In terms of matrices, a vector is usually defined as a column matrix:

$$\mathbf{V} = \begin{pmatrix} V_x \\ V_y \\ V_z \end{pmatrix}. \tag{1.69}$$

In order to, say, dot this vector with another,

$$\mathbf{U} = \begin{pmatrix} U_x \\ U_y \\ U_z \end{pmatrix}, \tag{1.70}$$

one must follow the rules of matrix multiplication which insist that the transpose of one of these vectors is found first, so:

$$\begin{aligned} \mathbf{V}\cdot\mathbf{U} &= \begin{pmatrix} V_x & V_y & V_z \end{pmatrix} \begin{pmatrix} U_x \\ U_y \\ U_z \end{pmatrix} \\ &= \begin{pmatrix} U_x & U_y & U_z \end{pmatrix} \begin{pmatrix} V_x \\ V_y \\ V_z \end{pmatrix} = V_xU_x + V_yU_y + V_zU_z \end{aligned}. \tag{1.71}$$

[9] If the words "matrix" and "matrix multiplication" do not mean anything to you, then feel free to *temporarily* skip the following to the end of the section, although eventually some knowledge of matrices will be needed. Good tutorials in the language of matrices may be easily found online.