Advanced General Relativity

Lecture notes by Sergei Winitzki

DRAFT February 5, 2026, version 1.1a

Lecture notes on topics in advanced General Relativity, including differential geometry, singularity theorems, variational principles, an introduction to spinors, and the vielbein formalism. This text is not yet in a final form.

Contents

Preface

This book is a revised and extended version of lecture notes for the course "Topics in Advanced General Relativity" taught by the author in the fall semester of 2005/06 at Ludwig-Maximilians University, Munich. The audience included advanced undergraduates and beginning graduate students. The choice of topics is intended to complement an introductory course in general relativity, which is assumed as a prerequisite. The present text covers differential geometry on manifolds, asymptotically flat spacetimes, singularity theorems, the use of variational principles in GR, the vielbein formalism, and spinor calculus. My goal is to provide a readable introduction to concepts that are covered in existing advanced texts, such as the classic books [12] and [36], and also explain a small number of newer results that are sufficiently tightly connected with the main topics of the lectures.

This book is intended a textbook rather than a research monograph. I give no overview of the history of the subject, and I do not intend to give complete references to first publications. I feel that this approach is justified because all the material (with a small number of exceptions) is well-established, and people who developed this subject already received the credit due to them. The scarcity of research-level references is also compensated by the fact that I derive all results in a self-contained fashion, unstead of referring readers to derivations in the literature. The bibliography contains[1] all the sources I consulted while preparing this book. After digesting the main ideas explained here, readers will be able to understand research-level papers and monographs listed in the bibliography.

I would like to make some comments regarding the organization of the text. I emphasize visual and conceptual explanations; however, I derive all the principal results in full detail. Definitions of useful mathematical notions are motivated and illustrated by examples. New terminology is shown in **boldface** within or near a definition; the *italics* type is reserved for emphasis. The symbol ■ marks the end of a derivation, a remark, or an example. These marks are intended to aid the reader in skipping unwanted material. Subsections marked with asterisks* may also be skipped at first reading.

The exposition in a previous version of this text was interspersed with unproved statements labeled "exercises" or "problems" that provided important additional information or even constituted an integral part of the development of the material. I call such statements **pseudo-exercises** since they are usually not as straightforward as genuine practice problems would be. It seems to me that the main reason for the existence of pseudo-exercises was my desire to avoid cluttering the text with derivations that may be omitted during a first reading. I feel that readers at an advanced level are insufficiently motivated and/or lack the time needed to solve pseudo-exercises. So I decided that this book will not contain any pseudo-exercises; every relevant statement[2] comes with a derivation.[3] The readers are certainly free to skip derivations that are not interesting to them. Many derivations and calculations are relegated to Appendix C to make the main text shorter. I also included a small number of short **Practice problems** intended as *straightforward* exercises for the readers. (Solutions to those are not given.)

A comment on the status of this text is in order. The book is "complete" in the sense that there are no gaps in the calculations. However, the text is presently being heavily revised. This preliminary version is now made freely available because I feel that it is useful even in the present, unfinished state. (Warning: the present text is a draft and *may* contain mistakes! If you *must* have a set of lecture notes in a finished form, please do not read this text until it is marked as a "finished" version. For example, it is presently not clear which sign convention for the Ricci tensor $R_{\mu\nu}$ I will finally adopt; because of this, signs might be inconsistent in some equations involving $R_{\mu\nu}$. Some comments regarding possible improvements are scattered in boldface throughout the text. These will disappear after the revision.)

 Comments and suggestions are welcome. The entire text was prepared by the author on computers running GNU/Linux, using the free document preparation systems LyX and LaTeX.

Sergei Winitzki, 2007

Suggested literature

There exist many textbooks on General Relativity written at every level. In the following table, I list some of the more advanced textbooks covering the main topics of the present text. For an introduction to GR, see e.g. the books [29, 10].

	[12]	[21]	[33]	[36]	[32]	[19]	[28]
Difficulty* (1-5):	5	2	4	4	3	2	1
Diff. geometry	+	+	+	+	+	+	+
Index-free						+*	
Action principle		+	+	+			+
Asympt. flatness	+	+		+	+*	+*	
Spinors in GR		+*		+	+		
Singularity thms.	+	+*		+*			+*
Hamiltonian GR		+		+			+
Tetrad formalism		+	+	+			

***Please note:**
A book's level of "difficulty" is my subjective estimate of how hard it would be for a well-prepared student to learn the contents of the book. (Higher level is harder.)

The book [21] discusses only some of the more elementary aspects of spinor calculus and singularity theorems.

The book [36] refers the reader to [12] for derivations of some key technical statements relevant to singularity theorems.

The book [32] treats asymptotic flatness exclusively in the spinor language.

The book [19] uses exclusively the abstract index notation. (Most other books use *non-abstract* index notation wherever, for instance, the non-tensorial Christoffel symbols $\Gamma^{\lambda}_{\mu\nu}$ are

[1] ***Will contain* when I finish the current revision of the book.**

[2] An **irrelevant statement** is something that seems generally interesting but has no direct relevance to the main subjects of the book. For example, the statement "Bianchi identities can be generalized to connections with nonzero torsion, see Ref. [xyz]" is irrelevant here because I *never* consider such connections in this book. Irrelevant statements, sometimes accompanied by references to additional literature, are confined to footnotes. The reader may safely ignore all the footnotes.

[3] **At present, some pseudo-exercises still remain, but almost all of them (except a very few) have full solutions. Pseudo-exercises are gradually converted to statements as the current major revision of the book progresses.**

used.) Also, the book [19] discusses only the basic facts relevant to asymptotic flatness.

The book [28] discusses only some basic aspects of singularity theorems relevant to black holes.

The present text aims to have the difficulty level 2. Right now the exposition is somewhat lacking in the coverage of asymptotic flatness, and there are very few applications of spinors and of the tetrad formalism. ■

1 Calculus in curved space

In this chapter I attempt to motivate and explain the main ideas and suggest images that illustrate the mathematical apparatus of differential geometry. The actual explanations start in Sec. 1.2.2. Additional literature on differential geometry from a mathematical viewpoint is [22, 11]. A more modern mathematical textbook (which I did not read) is [17].

1.1 Summary

General Relativity (**GR**) is a currently well-established physical theory of gravitation whose mathematical apparatus is based on differential geometry (more specifically, tensor calculus in curved spaces). In this chapter, I introduce the main concepts of this calculus: manifolds, tangent spaces, covariant derivatives, and curvature. I will explain and use the index-free notation. You (the reader) will perhaps appreciate this material more fully if you are already somewhat familiar with tensor calculus and GR, at least to the extent covered in Appendix A.

The pragmatic side of this chapter is to develop sufficiently powerful formalism for index-free calculations, which will be heavily used throughout this book. You may skip this chapter on first reading iff[1] you are familiar with the following concepts and notations.

$\mathcal{M}$	smooth manifold
$\mathbb{R}^n$	n-dimensional Euclidean space
$T_p\mathcal{M}$, $T_p^*\mathcal{M}$	(co)tangent spaces to $\mathcal{M}$ at point p
$T\mathcal{M}$	tangent bundle of a manifold $\mathcal{M}$
$\mathbf{v} \in T_p\mathcal{M}$	tangent vector at point p
$\mathbf{v} \circ f$	derivative of function f w.r.t. vector field $\mathbf{v}$
∂_t	vector field along a coordinate axis t
$[\mathbf{a}, \mathbf{b}] \equiv \mathcal{L}_\mathbf{a}\mathbf{b}$	commutator of two vector fields
$\mathcal{L}_\mathbf{v} T$	Lie derivative of tensor T w.r.t. vector $\mathbf{v}$
$\mathbf{u} \otimes \mathbf{v}$	tensor product of vectors
$\omega \circ \mathbf{v}$	1-form ω applied to vector $\mathbf{v}$
$\int \omega$, $\mathrm{d}\omega$	integral / exterior differential of n-form ω
$\omega_1 \wedge \omega_2$	exterior product of n-forms ω_1 and ω_2
$\iota_\mathbf{v}\omega$	insertion of vector $\mathbf{v}$ into n-form ω
$g(\mathbf{a}, \mathbf{b})$	scalar product of vectors using a metric g
$\hat{g}\mathbf{v}$	1-form corresp. to vector $\mathbf{v}$ through metric g
$\hat{g}^{-1}\omega$	vector corresp. to 1-form ω through metric g
$\boldsymbol{\varepsilon}$	volume n-form, Levi-Civita tensor
$\nabla_\mathbf{a}\mathbf{b}$	covariant derivative of vector $\mathbf{b}$ w.r.t. vector $\mathbf{a}$
$\mathrm{div}\mathbf{u}$	divergence of vector field $\mathbf{u}$
$R(\mathbf{a}, \mathbf{b})\mathbf{c}$	curvature tensor (Riemann tensor)
$R(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d})$	"fully covariant" Riemann tensor
$\mathrm{Ric}(\mathbf{a}, \mathbf{b})$	Ricci tensor

Concepts covered in this chapter also include: connecting vector fields; definition of the metric-compatible (Levi-Civita) connection on a manifold with a metric; torsion and torsion-freeness; properties of the Riemann tensor; geodesics and geodesic deviation; Riemann tensor for manifolds of constant curvature.

[1] If and only if.

1.1.1 Index-free notation

One approach to differential geometry is to treat vectors and tensors as multi-indexed arrays of numbers (called **components**), such as $T_{\mu\nu}$ or $R^\lambda{}_{\mu\alpha\beta}$, which depend on the choice of the coordinate system $\{x^\mu\}$. One manipulates expressions with many indices and implicit summations over indices, such as $g^{\mu\alpha} g_{\lambda\mu,\nu} g^{\nu\beta}$, and checks at the end of the calculations that the results are **covariant** (i.e. transform correctly under changes of coordinates). This approach was pioneered by Einstein and is still the way tensor calculus is presented in the current physical literature. I assume that you were already exposed to an introductory course of GR formulated using coordinates and components. If you are not sure, please scan Sec. A.2–A.7 in Appendix A for unfamiliar equations.

In the mathematical literature, the coordinate-free approach and the **index-free notation** are predominant. For example, the scalar product of vectors is denoted by $g(\mathbf{a}, \mathbf{b})$ or just $\langle \mathbf{a}, \mathbf{b} \rangle$ instead of $g_{\mu\nu} a^\mu b^\nu$ or $a_\mu b^\mu$. The preference for the coordinate-free approach is, in my opinion, due to the different character of the tasks typically undertaken by mathematicians and by physicists. Mathematicians study abstract relationships between objects and are more interested in finding objects that have rich properties and whose interconnections explain earlier results on a more general level and yield new concepts. On the other hand, physicists are mostly interested in making specific computations and deriving equations that will eventually give numerical values for quantities, even if the methods of computation are not particularly general, elegant, or illuminating.

In the coordinate-free approach, one avoids introducing a particular coordinate system $\{x^\mu\}$, and one does not talk about components of tensors. Instead, one uses geometrically defined operations, such as the covariant derivative or the Lie derivative, and the algebraic properties of tensors. In a calculation where a particular tensor is sought, especially when no symmetries are present, it may be helpful to introduce a suitable coordinate system and to determine the components of that tensor. However, often one needs to perform a "general" calculation that does not depend on a particular tensor (e.g., an investigation of a general property of the curvature tensor). In most cases, such "general" calculations are easier to perform in the coordinate-free approach because coordinate systems and components offer no computational or conceptual advantages. The experience of writing this book shows that the coordinate-free approach is well suited for applications such as studying the singularity theorems or the Hamiltonian formulation of GR.

It is natural[2] to use the index-free notation when one adopts the coordinate-free approach. Here is an example of a calculation in the index-free notation.

Statement: If a vector $\mathbf{u}$ is geodesic then $g(\mathbf{u}, \mathbf{u})$ is constant

[2] Although, strictly speaking, not necessary. Instead, one could use Penrose's "abstract index notation," where a^μ means a vector with an abstract label μ, rather than an array of components. See Sec. 1.7.1 for more details. One book that consistently uses the coordinate-free approach together with the abstract index notation is [19].

along the orbits of $\mathbf{u}$. *Proof*: The derivative of $g(\mathbf{u},\mathbf{u})$ along the flow of $\mathbf{u}$ is

$$\mathcal{L}_{\mathbf{u}} g(\mathbf{u},\mathbf{u}) = \nabla_{\mathbf{u}} g(\mathbf{u},\mathbf{u}) = 2g(\nabla_{\mathbf{u}}\mathbf{u},\mathbf{u}) = 0$$

since $\nabla_{\mathbf{u}}\mathbf{u} = 0$ for geodesic vector fields.

For comparison, the same calculation in the index notation looks like this: assuming that $u^\lambda u^\mu{}_{;\lambda} = 0$, we have

$$\frac{d}{d\tau}\left(g_{\mu\nu}u^\mu u^\nu\right) = u^\lambda \left(g_{\mu\nu}u^\mu u^\nu\right)_{;\lambda} = 2g_{\mu\nu}u^\lambda u^\mu{}_{;\lambda}u^\nu = 0.$$

The index notation is frequently very useful and even unavoidable in some calculations, especially when one needs to manipulate traces of high-rank tensors, as if often the case in GR. Nevertheless, the index-free notation has, in my view, significant advantages in the study of more mathematically advanced topics of field theory and GR. The index-free notation is concise, emphasizes the geometrical meaning of every quantity, and completely excludes non-tensorial quantities from consideration; for instance, the non-tensorial Christoffel symbols cannot even be defined. The index-free approach is well-suited for studying the general properties of various geometrical objects used in GR. Many general derivations (such as computing the second variation of the geodesic length) can be performed faster and more transparently in the index-free notation.

In this text, I adopt the coordinate-free approach; this leaves the choice of an index-free notation or the abstract index notation. I employ index-free calculations when it is practically more convenient than the abstract index notation. The index-free approach turns out to be almost always more convenient, except for certain cases, such as the variation of the Einstein-Hilbert action with respect to the metric, where index-free calculations become cumbersome (although still possible). In those cases I use the abstract index notation. Section 1.7 presents more discussion and shows how to convert expressions between the index-free and the index notations. In this book, I also introduce an admittedly nonstandard but unambiguous and adequate index-free notation for the trace of a tensor (Sec. 1.7.3).

To make the transition to the index-free approach easier for readers accustomed to the index notation, I will sometimes mention the index representation of tensor quantities used throughout the text. The formal correspondence between the index-free and the index notations is summarized in the following table.

$\mathbf{v}$	$v^\alpha\partial_\alpha \equiv v^\alpha \frac{\partial}{\partial x^\alpha}$	ω	$\omega_\alpha \mathrm{d}x^\alpha$ or ω_α
$\mathbf{v}\circ f$	$v^\alpha\partial_\alpha f \equiv v^\alpha f_{,\alpha}$	$\omega\circ\mathbf{v}$	$\omega_\alpha v^\alpha$
$[\mathbf{a},\mathbf{b}]$	$a^\mu\partial_\mu b^\nu - b^\mu\partial_\mu a^\nu$	$\mathrm{d}\omega$	$\omega_{\alpha,\beta} - \omega_{\beta,\alpha}$
$\nabla_{\mathbf{a}}\mathbf{b}$	$a^\nu\nabla_\nu b^\mu \equiv a^\nu b^\mu{}_{;\nu}$	$g(\mathbf{a},\mathbf{b})$	$g_{\mu\nu}a^\mu b^\nu \equiv a_\mu b^\mu$
$\operatorname{div}\mathbf{u}$	$\nabla_\alpha u^\alpha \equiv u^\alpha{}_{;\alpha}$	$\hat{g}\mathbf{v}$	$g_{\alpha\beta}v^\beta \equiv v_\alpha$
$\mathbf{u}\otimes\mathbf{v}$	$u^\alpha v^\beta$	$\hat{g}^{-1}\omega$	$g^{\alpha\beta}\omega_\beta \equiv \omega^\alpha$
$\omega\wedge\theta$	$\omega_\alpha\theta_\beta - \omega_\beta\theta_\alpha$	$\mathrm{Tr}_{\mathbf{x}}A(\mathbf{x},\mathbf{x})$	$g^{\alpha\beta}A_{\alpha\beta\ldots}{}^{\gamma\delta\ldots}$

Remark on the roman "d": Recently, certain science publishers started enforcing the rule of using a roman "d" in differentials (for instance, writing $\mathrm{d}^2x(\tau)/\mathrm{d}\tau^2$ or $\int \mathrm{d}^3x$) rather than using an italic "d," as was the common practice during the previous two centuries. I decided to use the roman symbol "d" to denote the exterior differential in the calculus of differential forms, but continue to use the italic "d" in other situations when differential forms are not involved. My intention is to use a roman "d" for rigorously defined objects, such as differential forms $\mathrm{d}x$, and to contrast them with "infinitesimals" dx, which are only heuristic notational tools that help the intuition. I write integrals, e.g. $\int \mathrm{d}^3x$, with a roman "d" when the integration of a differential form is implied, so that d^3x is a shorthand notation for the 3-form $\mathrm{d}x\wedge\mathrm{d}y\wedge\mathrm{d}z$. However, I write an italic "d" in derivatives, d/dt, and in integrals such as $\int f(x)dx$ where I need a simple one-dimensional integration rather than an integration of 1-form over a contour. I also write the "metric interval,"

$$ds^2 = dt^2 - dx^2 - dy^2 - dz^2,$$

with an italic "d" because this notation is essentially a jargon that bears only marginally on the calculus of differential forms; for instance, "ds" cannot be understood as a 1-form in this notation.

When using ordinary derivatives, I write $d/d\tau$ and $d^2/d\tau^2$ using the italic d, for the following reasons. The operators $d/d\tau$ and $d^2/d\tau^2$ are not essentially different from the partial derivative operators $\partial/\partial x$ or $\partial^2/\partial x^2$. Traditionally, one writes $\partial f/\partial x$ when f depends not only on x but also on other variables $y, z, \ldots$ that are held constant while $f(x,y,z,\ldots)$ is differentiated with respect to x. On the other hand, one writes df/dx to emphasize that $f(x)$ is a function of only one variable x, or that f is an expression that must be reduced, through suitable substitutions, to a function of x only. This notation is widely used to avoid cluttering the equations with explicit substitutions. For instance, the Euler-Lagrange equation of mechanics is written as

$$\frac{\partial L}{\partial x} - \frac{d}{dt}\frac{\partial L}{\partial \dot{x}} = 0,$$

implying that $\partial L/\partial\dot{x}$ must be expressed as a function of t before evaluating d/dt. However, the d's in the differential operators $\partial/\partial x$ and d/dx are merely historically developed notation and are not directly related to the exterior differential d. I use an italic "d" in "d/dx" to emphasize the difference between the rigorously defined operation "d" in an expression such as $\omega\wedge\mathrm{d}\omega$ and the heuristic "infinitesimal quantity" dx in d/dx. Readers who are comfortable with differential forms may ignore this typographical subtlety.

A related recent trend is to write a roman "e" for the base of the natural logarithm and a roman "i" for the imaginary unit ($\mathrm{i}^2 = -1$). One reason for switching to roman"d"s, "e"s, and "i"s is to avoid confusion when other quantities are denoted by the letters d, e, i. In this book, my principal intention is to make the notation unambiguous for a student who initially does not feel sure in this subject, while not excessively cluttering the writing. I use a roman "i" for the imaginary unit because the letter i is frequently used as an index, e.g. x_i. However, I use an italic $e = 2.718...$ to avoid confusion with a boldface roman $\mathbf{e}$ frequently used for frame basis vectors, $\{\mathbf{e}_a\}$. ■

1.1.2 Sample practice problems

You should be able to solve these problems if you know the material of Chapter 1.

Null vectors. Prove that two nonzero, non-parallel null vectors $\mathbf{l}, \mathbf{n}$ cannot be orthogonal. In other words, $g(\mathbf{l},\mathbf{n}) = 0$ is equivalent to $\mathbf{l} = \lambda\mathbf{n}$ when $g(\mathbf{l},\mathbf{l}) = g(\mathbf{n},\mathbf{n}) = 0$. Here g is a metric in four-dimensional space with the Lorentzian signature $(+,-,-,-)$.

Extremum of a functional on curves. Consider a functional

$$F[\gamma] = \int_{x_1}^{x_2} f(g(\dot\gamma,\dot\gamma))\mathrm{d}\tau,$$

where $f(x)$ is a given smooth, nonconstant function and $\dot\gamma$ is the tangent vector to the curve $\gamma(\tau)$. Derive the Euler-Lagrange equations describing the extremum of the functional $F[\gamma]$.

Symmetric geodesics. Consider a local coordinate system $\{x,y,z\}$ in a three-dimensional space. Suppose that the metric g has the form

$$g = \phi(x)dx^2 + A(y,z)dy^2 + C(y,z)dz^2,$$

where $\phi(x) > 0$, $A(y,z) > 0$, $C(y,z) > 0$ are smooth functions that depend on x,y,z as indicated. Show that lines $y = \text{const}, z = \text{const}$ (i.e. the coordinate lines of x) are geodesics. Obtain an explicit formula for a Killing vector $\mathbf{k}$ that the metric g admits. Is the vector field $\mathbf{k}$ geodesic?

Lie transport. Let $\mathbf{k}$ be a Killing vector for a metric g, and let two vector fields $\mathbf{a},\mathbf{b}$ be connecting for $\mathbf{k}$ and such that the triple $\{\mathbf{a},\mathbf{b},\mathbf{k}\}$ is orthogonal (but not necessarily orthonormal) at one point p. **a)** Can one choose the vectors $\mathbf{a},\mathbf{b}$ connecting for each other everywhere? **b)** Denote by γ the flow line of $\mathbf{k}$ that passes through p. Is the triple $\{\mathbf{a},\mathbf{b},\mathbf{k}\}$ orthogonal everywhere along γ?

Extrinsic curvature. A hypersurface with normal vector field $\mathbf{n}$ is embedded in a curved space with metric g. The extrinsic curvature $K(\mathbf{x},\mathbf{y})$ of the hypersurface is defined as a bilinear form

$$K(\mathbf{x},\mathbf{y}) = h(\nabla_{\mathbf{x}}\mathbf{n},\mathbf{y}),$$

where h is the induced metric on the hypersurface and $\mathbf{x},\mathbf{y}$ are arbitrary vectors tangent to the hypersurface. **a)** Starting from the definition and using $g(\mathbf{n},\mathbf{n}) = 1$, show that $K(\mathbf{x},\mathbf{y})$ is symmetric, $K(\mathbf{x},\mathbf{y}) = K(\mathbf{y},\mathbf{x})$. **b)** Compute the extrinsic curvature tensor explicitly for the surface $x^2 + y^2 - z^2 = -1$ defined in 3-dimensional Euclidean space with Cartesian coordinates $\{x,y,z\}$ at the point $\{0,0,1\}$, using $\{x,y\}$ as the local coordinates on the surface. ■

1.2 Basic notions: Manifolds and vector fields

In nongravitational physics, one describes events in four-dimensional Minkowski spacetime $\mathbb{R}^4$ with familiar coordinates $\{t,x,y,z\}$. Physical laws are expressed as partial differential equations for various fields in spacetime. General Relativity replaces Minkowski spacetime by a *curved* spacetime, and partial derivatives with respect to $\{t,x,y,z\}$ are replaced by more complicated differential operations. The mathematical concept of a "manifold" and accompanying notions are used to formalize the idea of a "curved space" and the ways of writing differential equations for fields in such spaces.

In the following subsection 1.2.1, I will list concise definitions of manifolds, tangent spaces, vector fields, etc. If you are satisfied by or already familiar with these definitions, you may skip the entire Sec. 1.2 and continue reading from Sec. 1.3. Otherwise, you may find it helpful to read the following sections (from Sec. 1.2.2 onwards) where I explain those definitions more visually.

1.2.1 Definitions

This section concisely summarizes the definitions and properties of manifolds and tangent vectors. If you prefer visual explanations to formal definitions, you may ignore this section and start reading from Sec. 1.2.2.

A **manifold** $\mathcal{M}$ is a set of points (which I informally call a "space") that locally looks like a subset of $\mathbb{R}^n$. In other words, every point $p \in \mathcal{M}$ has a small neighborhood around it, which is one-to-one mapped into a subset of $\mathbb{R}^n$. This mapping is called a **chart** and is equivalent to specifying a set of scalar functions, $x^\mu(p)$, defined in an open neighborhood $\mathcal{U}(p_0)$ of some point p_0 and giving the point $x^\mu \in \mathbb{R}^n$ corresponding to points $p \in \mathcal{M}$. These functions must uniquely identify each point p, so that for any two points p, p' from the neighborhood $\mathcal{U}(p_0)$ we have $x^\mu(p) = x^\mu(p')$ iff $p = p'$. Also, the image of the neighborhood $\mathcal{U}(p_0)$ under the map $p \mapsto x^\mu(p) \in \mathbb{R}^n$ must be an open set in $\mathbb{R}^n$. Under these conditions, the functions $x^\mu(p)$ constitute a **local coordinate system** defined on an open subdomain $\mathcal{U}(p_0) \subset \mathcal{M}$. Of course, different subdomains $\mathcal{U}(p_0)$, $\mathcal{U}(p_1)$, etc., may have different local coordinate systems. Whenever there is an intersection of two domains of $\mathcal{M}$ where local coordinate systems are defined, we get a **coordinate change** map $\mathbb{R}^n \to \mathcal{M} \to \mathbb{R}^n$ defined in a subset of $\mathbb{R}^n$. A manifold is **smooth** (**differentiable**) if all these coordinate change maps are smooth (differentiable) in the ordinary sense (as functions $\mathbb{R}^n \to \mathbb{R}^n$).

A particular manifold may be specified either by an explicit embedding into a larger space, or by giving a complete set of charts and re-charting maps, specifying how the charts are to be glued together.

A typical example of a manifold is the surface of a sphere in 3-dimensional space, usually called the **two-dimensional sphere** (or **2-sphere**) and denoted S^2. The 2-sphere S^2 is locally like $\mathbb{R}^2$, but globally is not (see Statement 1.2.2.1). It is useful to visualize a manifold as a "surface" embedded in a larger space, and to think that "we" are "flat" observers living entirely within the surface. So "we" cannot see the larger space and must describe the "surface" exclusively through its intrinsic properties.

As an example of a manifold described in a different way, consider the set of all 2×2 unitary complex matrices with unit determinant. These matrices form a group denoted $SU(2)$. (A matrix A is **unitary** if $A^\dagger A = 1$, where $A^\dagger$ is the Hermitian conjugate, i.e. the transposed and complex conjugate matrix.) It is perhaps not immediately clear how to introduce coordinates in this set of matrices. It can be shown (see Statement**??** on page ??) that all matrices $A \in SU(2)$ can be described explicitly as

$$A = \begin{pmatrix} a_0 - \mathrm{i}a_3 & -\mathrm{i}a_1 - a_2 \\ -\mathrm{i}a_1 + a_2 & a_0 + \mathrm{i}a_3 \end{pmatrix},$$

where a_0, ..., a_3 are real numbers satisfying $\sum_{j=0}^{3} a_j^2 = 1$. Therefore, $SU(2)$ is equivalent (as a manifold) to a three-dimensional sphere S^3.

A smooth, one-to-one map from one manifold to another is called a **diffeomorphism**. Two manifolds are called **diffeomorphic** if there exists a diffeomorphism between them. Diffeomorphic manifolds are "topologically equivalent." For example, a sphere can be diffeomorphically mapped onto the surface of a cube, but not onto a torus. Another example we have just seen is an explicit diffeomorphism (specified using coordinates) between $SU(2)$ and S^3.

We will always consider *smooth* finite-dimensional manifolds $\mathcal{M}$ and smooth functions $f(p)$, $p \in \mathcal{M}$, on these manifolds; so we omit the word "smooth" everywhere.

In our notation, $p \in \mathcal{M}$ is a point on a manifold, **scalar functions** (scalar fields) are functions $f : \mathcal{M} \to \mathbb{R}$, so that $f(p)$ is a number, and boldface letters $\mathbf{v}, \mathbf{w}, ...$ are vectors and vector fields (defined below). We use Greek letters $\alpha, \beta, ..., \lambda, \mu, ...$ for tensor indices and also for numbers (scalars).

A **curve** on a manifold is a set of points $\gamma(\tau)$ parametrized by a real number τ, i.e. a smooth map $\gamma : \mathbb{R} \to \mathcal{M}$, sometimes defined only on a subset of $\mathbb{R}$.

A **derivation** D **at a point** $p \in \mathcal{M}$ is a map from scalar functions to numbers, satisfying the "derivative-like" properties

$$\begin{aligned} \mathrm{D}(f+g) &= \mathrm{D}f + \mathrm{D}g, \text{ [linearity]} \\ \mathrm{D}(fg) &= \mathrm{D}(f)g + f\mathrm{D}(g), \text{ [Leibnitz rule]} \\ \mathrm{D}f &= 0 \text{ if } f = \text{const around } p. \end{aligned}$$

All the derivations at a point p form a vector space called the **tangent space** $T_p\mathcal{M}$ at the point p. Vectors from the tangent space are denoted by boldface letters, e.g. $\mathbf{u} \in T_p\mathcal{M}$, and their action on functions is denoted by $\mathbf{u} \circ f$. The number $\mathbf{u} \circ f$ is interpreted as the derivative of the function f at the point p in the direction $\mathbf{u}$. Using a local coordinate system $\{x^\alpha\}$, $\alpha = 1, ..., n$, one can prove that the tangent space is an n-dimensional vector space. An explicit basis in this space is given by the n coordinate derivatives $\mathbf{e}_\alpha \equiv \partial/\partial x^\alpha$. The components of a vector $\mathbf{v}$ in this basis can be found as $v^\alpha = \mathbf{v} \circ x^\alpha$, and then the vector $\mathbf{v}$ is decomposed as

$$\mathbf{v} = \sum_{\alpha=1}^{n} v^\alpha \mathbf{e}_\alpha = \sum_{\alpha=1}^{n} v^\alpha \frac{\partial}{\partial x^\alpha}.$$

The union of all tangent spaces $T_p\mathcal{M}$ for every $p \in \mathcal{M}$ is called the **tangent bundle** of $\mathcal{M}$ and is denoted $T\mathcal{M}$. The tangent bundle is itself a manifold with the same charts as $\mathcal{M}$ but with twice the dimension of $\mathcal{M}$. This is so because every chart will map a patch of $\mathcal{M}$ into a patch of $\mathbb{R}^n$ and tangent vectors map into vectors in $\mathbb{R}^n$, thus the patch of the tangent bundle will map into a patch of $\mathbb{R}^n \times \mathbb{R}^n$.

A **vector field** $\mathbf{v}(p)$ on a manifold $\mathcal{M}$ is a tangent vector $\mathbf{v} \in T_p\mathcal{M}$ chosen at every point $p \in \mathcal{M}$, i.e. a map $\mathbf{v} : \mathcal{M} \to T\mathcal{M}$ such that the image $\mathbf{v}(p)$ of every point p belongs to the tangent space $T_p\mathcal{M}$ at the same point p.

A **covector** is a linear map ω from vectors $\mathbf{v}$ into numbers. The space of covectors to tangent vectors at a point p is denoted by $T_p^*\mathcal{M}$ and is called the **cotangent space** at point p. A covector $\omega(p) \in T_p^*\mathcal{M}$ chosen at every point $p \in \mathcal{M}$ is called a **1-form**. The pointwise action of a 1-form ω on a vector field $\mathbf{v}$ is denoted by $\omega \circ \mathbf{v}$ and is a scalar function on $\mathcal{M}$. A **tensor field** on a manifold is defined similarly: it is a tensor based on the tangent space at a point p.

The **commutator** of vector fields $\mathbf{u}$ and $\mathbf{v}$ is the vector field $[\mathbf{u}, \mathbf{v}]$ defined by its action on functions f by

$$[\mathbf{u}, \mathbf{v}] \circ f \equiv \mathbf{u} \circ (\mathbf{v} \circ f) - \mathbf{v} \circ (\mathbf{u} \circ f).$$

(It can be shown that $[\mathbf{u}, \mathbf{v}]$ also satisfies the properties of a vector field; see Statement 1.2.10.1.) One says that the vector fields **commute** if their commutator is equal to zero. For example, coordinate basis vector fields ∂_x and ∂_y commute because $\partial_x \partial_y f = \partial_y \partial_x f$ for smooth functions $f(x, y)$. On the other hand, vector fields $x\partial_y$ and ∂_x do not commute because

$$x\partial_y (\partial_x f) - \partial_x (x\partial_y f) = -\partial_y f \neq 0.$$

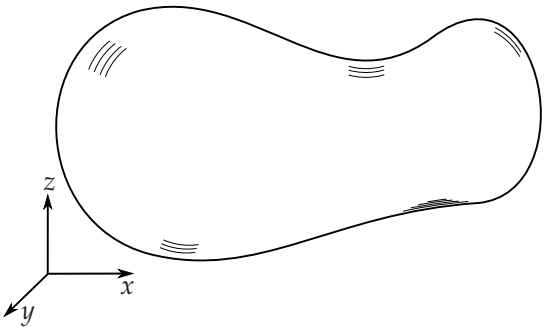

Figure 1.1: A visual example of a manifold is a curved surface embedded in the flat 3-dimensional space.

A vector field $\mathbf{c}$ is called a **connecting vector** for the vector field $\mathbf{v}$ if $[\mathbf{c}, \mathbf{v}] = 0$. (Since $[\mathbf{c}, \mathbf{v}] = -[\mathbf{v}, \mathbf{c}]$, it follows that $\mathbf{v}$ is also a connecting vector for $\mathbf{c}$.) All the coordinate basis vectors $\partial/\partial x^\alpha$ are connecting vectors for each other. Conversely, if one is given a set of n linearly independent vectors $\mathbf{e}_1$, ..., $\mathbf{e}_n$ all of which are connecting for each other within a domain, then there exists a local coordinate system $\{x^\alpha\}$ in that domain such that $\mathbf{e}_\alpha$ is equal to the α-th coordinate basis vector $\partial/\partial x^\alpha$. The orbits of $\mathbf{e}_1, ..., \mathbf{e}_n$ are the "coordinate grid" of that coordinate system.

1.2.2 Manifolds and coordinates

If you prefer formal definitions to visual explanations, or if the definitions in Sec. 1.2.1 are entirely familiar, you may skip Sec. 1.2.2 to 1.2.11.

One of the main assumptions underlying GR is that the geometry of our universe is curved rather than flat. In a curved geometry, initially parallel lines may move closer or further apart, and the sum of the angles of some triangles may be larger or smaller than π, depending on the triangle and its location in space. Let us examine the notion of "curved space," starting from an intuitive picture.

A simple visual image of a curved space is a 2-dimensional surface embedded in the 3-dimensional Euclidean (flat) space, as shown in Fig. 1.1. Since the surface is smooth, every sufficiently small patch of the surface looks like a small piece of a flat 2-dimensional space $\mathbb{R}^2$. However, different such patches are glued together in a nontrivial way, so that the entire surface has a shape we perceive as "curved."

According to GR, the geometry of the universe is similar to that of a curved surface. In a sufficiently small neighborhood of each point, the universe looks like the flat four-dimensional Minkowski spacetime $\mathbb{R}^4$, studied in Special Relativity. However, the *global* geometry of the universe (i.e. the geometry observed at sufficiently large distances and time intervals) is different from the Minkowski geometry.

This motivates us to consider the idea of a "general kind of curved space," that is, a space which is only *locally* similar to $\mathbb{R}^4$ but globally very different from $\mathbb{R}^4$. The mathematical object that formalizes this intuitive idea is called a **manifold**. A standard, formal definition of a manifold is given in Sec. 1.2.1. I will now motivate and explain that definition, using a two-dimensional surface as a main example.

To formalize the idea that a two-dimensional surface $\mathcal{M}$ "locally looks like" $\mathbb{R}^2$, one says that for each point p in $\mathcal{M}$ there is a small patch of $\mathcal{M}$ around p which is mapped bijectively (one-to-one) into some patch of $\mathbb{R}^2$. A patch of $\mathbb{R}^2$ can

be described by coordinates $\{x_1, x_2\}$ that vary in some finite ranges. Therefore, there is a one-to-one map assigning values x_1 and x_2 to points of $\mathcal{M}$ in a small patch (see Fig. 1.3). This map is called a **local coordinate system** or a **chart**. We can now formulate a first rough definition: A **manifold** is a set $\mathcal{M}$ such that there is a local coordinate system around every point of $\mathcal{M}$. A manifold can be visualized as a surface that is smooth near each point and does not have "corners" or "spikes."

Example: 2-sphere. A typical example of a smooth manifold is the **unit 2-sphere** specified by the equation

$$x^2 + y^2 + z^2 = 1,$$

where x, y, z are the ordinary Cartesian coordinates in a three-dimensional Euclidean space. The 2-sphere is usually denoted S^2. To establish that S^2 is a manifold, it is sufficient to find a local coordinate system near each point. This is straightforward; for instance, it is easy to see that the first two Cartesian coordinates, $\{x, y\}$, can be used as local coordinates $\{x_1, x_2\}$ in a neighborhood of the north pole $\{0, 0, 1\}$, even in the entire northern hemisphere (but not beyond the equator). The coordinates $\{x, y\}$ cannot be used near the point $\{x, y, z\} = \{0, 1, 0\}$ because, e.g., a part of the straight line segment $\{-1 < x < 1, y = 0.99\}$ in $\mathbb{R}^2$ that maps onto S^2 becomes a small circle in S^2. So the coordinates $\{x, y\}$ do not provide a one-to-one map from any subset of $\mathbb{R}^2$ into S^2 near the point $\{0, 1, 0\}$. However, $\{x, z\}$ can be used as a local coordinate system near $\{0, 1, 0\}$.

An example of a non-manifold is a 2-sphere with a line attached to it. No two-dimensional local coordinate system can adequately represent the geometry in the neighborhood of the point where the line is attached.

Example: 2-Torus. Another example of a two-dimensional manifold is a 2-torus. One way to describe the 2-torus is to specify it as the set of points $\{x, y, z\} \in \mathbb{R}^3$ satisfying the equation (see Fig. 1.2)

$$\left(\sqrt{x^2 + y^2} - A\right)^2 + z^2 = R^2, \quad A > R. \tag{1.1}$$

Another possibility is to describe the torus as a square with opposite sides identified. One can introduce local coordinates $\{\phi, \theta\}$ on the torus and the specify the embedding into $\mathbb{R}^3$ by $x = \rho\cos\phi$, $y = \rho\sin\phi$, $z = R\cos\theta$, where$\rho = A + R\sin\theta$. Since this embedding is manifestly periodic in both ϕ and θ with period 2π, one needs to restrict the range of ϕ and θ at most to $0 < \phi, \theta < 2\pi$ to satisfy the one-to-one mapping condition. Another choice of local coordinates is

$$x = \frac{a\cos\phi}{b - \sin\theta}, \quad y = \frac{a\sin\phi}{b - \sin\theta}, \quad z = c\frac{\cos\theta}{b - \sin\theta}, \tag{1.2}$$

with appropriately chosen constants a, b, c. ■

Practice problem: Determine the values of the constants a, b, c for Eq. (1.2) such that Eq. (1.1) is satisfied for all ϕ and θ. ■

As I attempted to illustrate in Fig. 1.3, the relationship between the local coordinates $\{x_1, x_2\}$ and the Cartesian coordinates $\{x, y, z\}$ may be complicated. However, in some cases one can simply select a subset of the Cartesian coordinates — for instance, $\{x, y\}$ — and obtain a local coordinate system covering some part of $\mathcal{M}$. The only requirement on the local coordinates is that they should provide a one-to-one correspondence between a patch $\mathcal{U} \subset \mathcal{M}$ and a patch $\mathcal{V} \subset \mathbb{R}^2$. For instance, any straight line segment situated within $\mathcal{V}$ must correspond to a non-self-intersecting line segment in $\mathcal{M}$.

torus-coords.eps

Figure 1.2: A schematic representation of a torus given by Eq. (1.1). The intersection with the x-z plane consists of two circles of radius R, centered at $x = \pm A$.

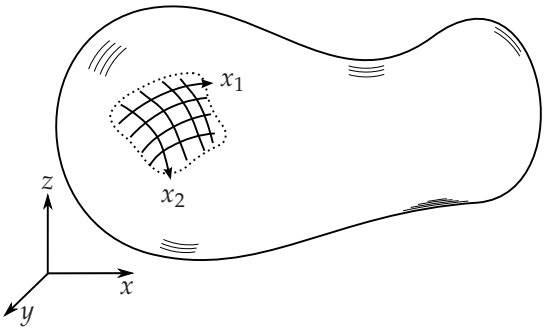

Figure 1.3: A local coordinate system $\{x_1, x_2\}$ covers a small patch of a manifold.

It is important to note that there may be *no global* coordinate system covering the entire manifold $\mathcal{M}$. This is the case, for instance, if the manifold $\mathcal{M}$ is a closed surface, such as a sphere (see Statement 1.2.2.1 below), or a donut-shaped surface (a torus). If a global coordinate system exists, $\mathcal{M}$ is an infinitely extended (perhaps curved) plane-like surface that never "closes on itself." Such a surface is *topologically* the same as $\mathbb{R}^2$.

It is easy to see that the global geometry of the sphere S^2 is different from that of $\mathbb{R}^2$, even though small patches of S^2 look like small patches of $\mathbb{R}^2$.

Statement 1.2.2.1: There exists no global coordinate system (chart) in S^2 that would map S^2 smoothly and one-to-one into $\mathbb{R}^2$. (Proof on page 175.) ■

The often used **spherical coordinates** $\{\theta, \phi\}$, defined by

$$x = r\cos\theta\cos\phi, \; y = r\cos\theta\sin\phi, \; z = r\sin\theta,$$

do not constitute a global coordinate system on S^2 because the mapping $S^2 \to \{\theta, \phi\}$ is not one-to-one at the poles of the sphere, $\theta = \pm\pi/2$. Nevertheless, this flaw is relatively insignificant since only two points (the poles) are problematic. Such points where the chart map is not one-to-one are called **coordinate singularities** of a local coordinate system. Despite the presence of coordinate singularities, the spherical coordinates constitute an "almost global" coordinate system, which is very convenient for many calculations. The singular points $\theta = \pm\pi/2$ sometimes need to be handled specially.

Practice problem: Consider the surface in $\mathbb{R}^3$ specified by the equation $x^2 + y^2 - z^2 = 1$, where $\{x, y, z\}$ are the standard coordinates. Prove that this surface is a two-dimensional manifold by showing how to find local coordinates $\{x_1, x_2\}$ in the neighborhood of an arbitrary point $\{x, y, z\}$ on the surface. Is there a global coordinate system for this manifold? *Answer:* No. ■

Although our considerations so far concerned two-dimensional surfaces embedded in a three-dimensional Euclidean space, it is clear that completely analogous arguments apply to higher-dimensional "hypersurfaces" embedded into an even higher-dimensional Euclidean space. This picture leads to the general definition of an ***n*-dimensional manifold**: a set of points $\mathcal{M}$ such that each point $p \in \mathcal{M}$ belongs to a chart that maps it into a subset of $\mathbb{R}^n$.

1.2.3 Manifolds: intrinsic picture

A very important idea is to imagine that there are "flat" observers living entirely within the surface $\mathcal{M}$. These observers cannot see the larger space and describe the surface $\mathcal{M}$ exclusively through the relationships between the points of $\mathcal{M}$, without referring to any embedding into a larger space. Such a description is called **intrinsic**. The main fact is that the intrinsic description is sufficient for all purposes relevant to the internal geometry of the manifold. One may certainly introduce an embedding of $\mathcal{M}$ into a larger space[3] — either as a way to describe a particular manifold $\mathcal{M}$ more concisely or as a guide for the intuition,— but the final results must be independent of any such embedding. In any case, it turns out that the intrinsic description of the geometry of $\mathcal{M}$ is more elegant than a description in terms of an embedding into a larger space. Another argument in favor of the intrinsic description is that according to General Relativity, *we are* observers living in a four-dimensional curved manifold (the spacetime), and we do not directly see any higher-dimensional space in which our spacetime is embedded as a "hypersurface." Therefore, we need a way to describe the properties of our spacetime in terms of *intrinsic* measurements, i.e. measurements performed entirely within the spacetime.

An example of an intrinsic property is smoothness. A function is **smooth** (of class C^∞) if it is infinitely many times differentiable. In the embedding picture, one can visualize a smooth n-dimensional manifold $\mathcal{M}$ as an n-dimensional hypersurface smoothly embedded into an N-dimensional Euclidean space with coordinates $\{X^a\}$, $a = 1, ..., N$. Using a fixed smooth embedding of a manifold $\mathcal{M}$ into $\mathbb{R}^N$, one could easily define smoothness of functions and curves on $\mathcal{M}$. These definitions may look like this. A **smooth scalar function** is a map $f : \mathcal{M} \to \mathbb{R}$ which is a smooth function of the coordinates $\{X^a\}$. A **smooth curve** $\gamma(\tau)$ on a manifold $\mathcal{M}$ is a map $\gamma : \mathbb{R} \to \mathcal{M}$ such that the curve coordinates $X^a(\tau)$ in the embedding space are smooth functions of τ. However, these definitions depend on the embedding of $\mathcal{M}$ into $\mathbb{R}^N$. This dependence can be avoided and equivalent definitions can be given *intrinsically*, using local charts on $\mathcal{M}$ that map patches of $\mathcal{M}$ into patches of $\mathbb{R}^n$. Here is how one can define a smooth scalar function. Within one chart, a function $f : \mathcal{M} \to \mathbb{R}$ becomes a function $\mathbb{R}^n \to \mathbb{R}$, defined on a patch of $\mathbb{R}^n$. Thus, smoothness of f can be defined as smoothness of each restriction of f to patches of $\mathbb{R}^n$. In brief, a function f is smooth in $\mathcal{M}$ if it is smooth in every chart. Similarly, a curve $\gamma : \mathbb{R} \to \mathcal{M}$ is smooth if its restriction $\gamma : \mathbb{R} \to \mathbb{R}^n$ to each chart is smooth.

If we define smoothness of functions through their restrictions to charts, we need to consider what happens within an intersection of two different charts. Could it happen that a function f is smooth according to one chart but not smooth according to the other chart? Let $\mathcal{U}_1, \mathcal{U}_2 \subset \mathcal{M}$ be two different chart domains and $\phi_1 : \mathcal{U}_1 \to \mathbb{R}^n$, $\phi_2 : \mathcal{U}_2 \to \mathbb{R}^n$ the corresponding local coordinate maps. Since ϕ_1 and ϕ_2 are one-to-one maps, we may define a map $\phi_1 \phi_2^{-1} : \mathbb{R}^n \to \mathbb{R}^n$. (This map is defined in a small patch of $\mathbb{R}^n$ which is the image of the intersection of $\mathcal{U}_1$ and $\mathcal{U}_2$ under ϕ_2.) We may call this map a **coordinate change map**. If *every* coordinate change map (for every pair of intersecting charts) is a smooth map $\mathbb{R}^n \to \mathbb{R}^n$ in the usual sense, i.e. defined by smooth functions in the standard coordinates $\{x_1, ..., x_n\}$, then all charts will agree about the smoothness or otherwise of a function $f : \mathcal{M} \to \mathbb{R}$ or a curve $\gamma : \mathbb{R} \to \mathcal{M}$. Therefore, one adds to the definition of a smooth manifold the requirement that every coordinate change map $\phi_1 \phi_2^{-1}$ is smooth, for every pair of intersecting charts. In this way, smoothness of the manifold itself is adequately described by the requirement that the coordinate change maps be smooth.

If there exists a smooth one-to-one map between two entire manifolds, these manifolds are called **diffeomorphic**. For example, the surface $z = x^2 + y^2$ is diffeomorphic to $\mathbb{R}^2$ because there is a one-to-one map between points $\{x, y, z\}$ on the surface and the plane $\{x, y\}$. Thus, just one chart (with $\{x, y\}$ as the coordinates) is sufficient to map this manifold into $\mathbb{R}^2$. On the other hand, a 2-sphere is not diffeomorphic to $\mathbb{R}^2$ (Statement 1.2.2.1).

1.2.4 Tangent spaces

Vectors are used in ordinary physics to specify "directed magnitudes," such as velocities and forces. Therefore, it is important to develop the concept of a "directed magnitude" in a curved space.

Embedding picture. To aid the intuition, let us visualize a manifold $\mathcal{M}$ as a surface embedded in a larger space $\mathbb{R}^3$. An immediate example of a "directed magnitude" comes from considering the velocity of a moving point. Hence, let us imagine a point that moves along a **curve** $\gamma(\tau)$ within a manifold, where γ is a map $\mathbb{R} \to \mathcal{M}$. It is clear that the velocity of the moving point can be viewed as a vector *in the larger space* $\mathbb{R}^3$. This velocity vector always remains tangent to the surface $\mathcal{M}$. Due to this picture, "directed magnitudes" defined within a manifold are called **tangent vectors**.

At every point $p \in \mathcal{M}$, there is a tangent plane to the surface. This tangent plane is a vector subspace of $\mathbb{R}^3$ that contains all the tangent vectors at point p. Considered as a two-dimensional vector space, it is denoted $T_p\mathcal{M}$ and is called the **tangent space** to the manifold $\mathcal{M}$ at point p (see Fig. 1.4). If the manifold $\mathcal{M}$ is an $(n-1)$-dimensional hypersurface embedded into an n-dimensional Euclidean space $\mathbb{R}^n$ then each tangent plane is a vector space of dimension $n-1$.

For instance, consider a hypersurface defined by the equation $f(\mathbf{x}) = 0$, where $\mathbf{x} \in \mathbb{R}^3$. The normal vector $\mathbf{n}$ at a point $\mathbf{x}_0$ has the following components,

$$n_j(\mathbf{x}_0) = \left.\frac{\partial f}{\partial x^j}\right|_{\mathbf{x}=\mathbf{x}_0}.$$

[3] It is shown in differential geometry (Whitney's embedding theorem [37]) that any manifold can be embedded in a sufficiently high-dimensional Euclidean space.

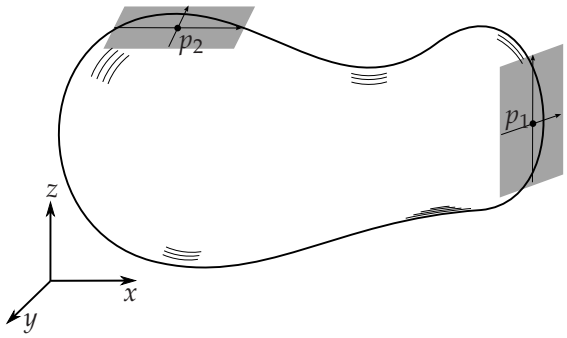

Figure 1.4: Tangent spaces at points p_1 and p_2 of a manifold, visualized as vector subspaces of $\mathbb{R}^3$.

Any vector $\mathbf{t}$ orthogonal to $\mathbf{n}$ is tangent to the hypersurface at the point $\mathbf{x}_0$. Hence, the tangent plane at the point $\mathbf{x}_0$ is the set of points $\mathbf{x}$ in $\mathbb{R}^3$ specified by

$$\mathbf{n} \cdot (\mathbf{x} - \mathbf{x}_0) \equiv \sum_{j=1}^{n} (x^j - x_0^j) n_j(\mathbf{x}_0) = 0.$$

The tangent space is made of vectors with components $x^j - x_0^j$ satisfying this equation. This is a two-dimensional vector space that depends on the point $\mathbf{x}_0$ (i.e. it is a different space for each $\mathbf{x}_0$).

It is important to realize that tangent spaces at different points are *not* naturally related to each other, even though every tangent space is two-dimensional and looks like $\mathbb{R}^2$. For instance, there is no natural way to add a tangent vector at a point p_1 to a tangent vector at another point p_2, or to subtract these vectors. This should be especially clear from Fig. 1.4, which shows different tangent planes. There is no obvious way to map vectors from one tangent plane into vectors from the other.

We will see such a map later when we define an additional structure on the manifold called a "connection." That structure literally provides a connection (i.e. a one-to-one map) between tangent spaces that are infinitesimally close (and thus, in principle, between any two tangent spaces). However, there are infinitely many possible "connections" on a manifold. In order to select a single "preferred" connection, one needs to introduce additional considerations.

Remark: If an embedding of $\mathcal{M}$ into a Euclidean space is given, there is a way to construct a geometrically preferred connection. One imagines that the manifold $\mathcal{M}$ and the tangent planes are solid surfaces, and that one can "roll" the first tangent plane without sliding and without turning around the surface $\mathcal{M}$ along a certain path in $\mathcal{M}$ until the location of the other plane is reached. This "mechanically" motivated procedure yields a one-to-one map from one tangent plane into another one. The mapping between tangent planes depends on the path along which we "roll" them, but is unambiguous for *infinitesimally close* points. (It is only the mapping between infinitesimally close tangent spaces that is necessary to determine a connection.) The resulting connection is called the **Levi-Civita connection**; it is defined in Sec. 1.6.6 without referring to the mechanical construction just described. Of course, this connection depends on the embedding of $\mathcal{M}$ into a Euclidean space. ■

Intrinsic picture. So far we considered tangent vectors using the embedding picture, i.e. by imagining that the $\mathcal{M}$ manifold is a hypersurface embedded into a larger space. However, it is much more useful to describe tangent vectors *intrinsically*, i.e. without such an embedding. We could try to define a tangent vector as the "directed velocity" $d\gamma/d\tau$ of some curve $\gamma(\tau)$ going through p. However, we cannot directly differentiate $\gamma(\tau)$ with respect to τ without an embedding because the expression $\gamma(\tau+\delta) - \gamma(\tau)$ is undefined: we cannot "subtract" one point from another. To circumvent this problem without introducing coordinates, we focus not on points but on *scalar-valued functions* of points. (Generally, shifting the attention from an abstract set to functions defined on the set is a powerful idea that has proven useful in many areas of mathematics.)

Let $f : \mathcal{M} \to \mathbb{R}$ be a scalar function defined on the manifold $\mathcal{M}$. (A **scalar** function means a function with a numeric value, as opposed to a matrix-valued or a vector-valued function.) Suppose that a curve $\gamma(\tau)$ goes through the point p_0 at $\tau = 0$, i.e. that $\gamma(\tau = 0) = p_0$, and consider the **directional derivative** of the function $f(p)$ **along the curve** γ at the point p_0. This directional derivative is defined as the derivative of the function $f(\gamma(\tau))$ with respect to τ,

$$\mathrm{D}_\gamma f(p_0) \equiv \left.\frac{d}{d\tau}\right|_{\tau=0} f(\gamma(\tau)).$$

This derivative is interpreted as the instantaneous rate of change of the function f at point p_0, as one moves along the curve γ passing through p_0.

To make the discussion specific, consider the manifold $\mathcal{M} = \mathbb{R}^n$. Since points of $\mathcal{M}$ are arrays of n coordinates $\{x^1, ..., x^n\}$, a curve $\gamma(\tau)$ is specified by n functions $\gamma^j(\tau)$. Then the directional derivative $\mathrm{D}_\gamma f$ of a function $f(x^1, ..., x^n)$ at a point p_0 is expressed by the formula

$$\mathrm{D}_\gamma f(p_0) = \sum_{k=1}^{n} \left.\frac{d\gamma^k}{d\tau}\right|_{\tau=0} \left.\frac{\partial f}{\partial x^k}\right|_{\{x^j\}=p_0}.$$

For our present purposes, it is important to view the directional derivative D_γ as an object that knows how to evaluate a derivative of any given function, i.e., as a *map* from functions to numbers. The directional derivative D_γ depends not only on the direction of the curve γ, but also on the "speed" at which the points $\gamma(\tau)$ are traversed as τ changes. For instance, the value of $\mathrm{D}_\gamma f$ will be multiplied by 2 if we keep the shape of the curve $\gamma(\tau)$ unchanged but replace the parameter τ by 2τ. Also, we note that $\mathrm{D}_\gamma f$ depends on the behavior of the curve γ in an infinitesimal neighborhood of p_0 but does not depend on the behavior of γ at other points; there are many curves γ that yield the same derivative operation D_γ at the point p_0. Therefore, the operation D_γ carries the information only about the magnitude and the direction of the "intantaneous velocity" along the curve at p_0. This is precisely the information we expect to be represented by a tangent vector at the point p_0. Thus we are motivated to view the *map* D_γ *itself* as the tangent vector.

Remark: Here is a visual motivation for considering directional derivatives. We would like to define the tangent vector as an object that carries information about the velocity of a moving point. The point is *moving* if some quantities are *changing* along its path. A scalar function f can represents a local quantity, or some property at a point (such as intensity of a physical field or a temperature). Then it is natural to consider the derivative of f along the path of a moving particle.

This directional derivative, $\mathrm{D}_\gamma f$, depends both on the motion of the particle and on the function f. So the part of $\mathrm{D}_\gamma f$ that is independent of f, i.e. the differential operator itself, is the object that carries information about the velocity of the motion. Hence, it is natural to identify the tangent vector as the differential operator D_γ. ■

We can now define the space of the tangent vectors at p_0, i.e. the **tangent space** $T_{p_0}\mathcal{M}$, as the vector space consisting of all the possible directional derivatives D_γ at p_0. This will be an *intrinsic* definition of the tangent space because this definition does not make use of any embedding of $\mathcal{M}$ into a larger space. This definition also does not rely on a choice of a local coordinate system because only the *curve* $\gamma(\tau)$ — a set of points in the manifold — that is used to construct the directional derivative.

It is perhaps difficult for the reader to visualize the "space of all possible directional derivatives." In order to help the intuition and to describe this "space of derivatives" more explicitly, let us temporarily introduce a local coordinate system $\{x^\alpha\}$ around the point p_0. If the curve $\gamma(\tau)$ is represented in the coordinates $\{x^\alpha\}$ by functions $\gamma^\alpha(\tau)$, we can express the directional derivative $\mathrm{D}_\gamma f$ of a function f as follows,

$$\mathrm{D}_\gamma f(p_0) = \sum_\alpha \left.\frac{\partial f}{\partial x^\alpha}\right|_{p_0} \left.\frac{d\gamma^\alpha}{d\tau}\right|_{\tau=0} = \left[\sum_\alpha \left.\frac{d\gamma^\alpha}{d\tau}\right|_{\tau=0} \left.\frac{\partial}{\partial x^\alpha}\right|_{p_0}\right] f.$$

It is clear from this formula that the derivative operation D_γ is represented by the differential operator

$$\mathrm{D}_\gamma \equiv \sum_\alpha \left.\frac{d\gamma^\alpha}{d\tau}\right|_{\tau=0} \left.\frac{\partial}{\partial x^\alpha}\right|_{p_0}$$

acting on functions f, where $\partial/\partial x^\alpha$ is the partial derivative with respect to one coordinate x^α. Moreover, we see that all the possible directional derivatives D_γ are described by the set of coefficients $\{d\gamma^\alpha/d\tau\}$ multiplying a fixed set of differential operators $\{\partial/\partial x^\alpha\}$. The coefficients $d\gamma^\alpha/d\tau$ depend on the curve γ, while the operators $\partial/\partial x^\alpha$ do not. Therefore, it is natural to regard $\partial/\partial x^\alpha$ as the basis in the vector space of directional derivatives and $d\gamma^\alpha/d\tau$ as the components of D_γ in that basis.

In this way, we have expressed directional derivatives D_γ at a point p_0 as differential operators in a local coordinate system. An arbitrary directional derivative can be written as $\sum_\alpha u^\alpha \partial/\partial x^\alpha$, where u^α are some coefficients. Now it is clear that all directional derivatives at point p_0 form a vector space spanned by the basis $\{\partial/\partial x^\alpha\}$. This vector space is *the same* as the tangent space $T_{p_0}\mathcal{M}$ defined above using the embedding picture. This is so because the tangent space $T_{p_0}\mathcal{M}$ in the embedding picture consists of all the possible vectors tangent to $\mathcal{M}$ at p_0, which is the same as the set of velocity vectors of all possible curves $\gamma(\tau)$ at p_0, i.e. the set of all possible values of the components $d\gamma^\alpha/d\tau$. So we may *identify* the space of tangent vectors to curves at p_0 with the space of directional derivatives at p_0. These are two equivalent descriptions of the tangent space $T_{p_0}\mathcal{M}$. A tangent vector can be understood both as a "direction within the manifold" and, at the same time, as a differential operator acting on scalar functions.

The natural basis $\{\partial/\partial x^\alpha\}$ in the tangent space is called the **coordinate basis** corresponding to a chosen coordinate system $\{x^\alpha\}$.

I will denote tangent vectors by boldface letters such as $\mathbf{v}$ rather than by D_γ, since the curve γ in D_γ plays only an auxiliary role. The action of a tangent vector $\mathbf{v}$ on a function f will be denoted by $\mathbf{v} \circ f$ rather than by $\mathrm{D}_\gamma f$. In local coordinates $\{x^\alpha\}$, one can express the value of a directional derivative (tangent vector) $\mathbf{v}$ at point p_0 applied to a function f as

$$\mathbf{v} \circ f = \sum_\alpha v^\alpha \left.\frac{\partial f}{\partial x^\alpha}\right|_{p_0}, \tag{1.3}$$

where v^α are the components of the vector $\mathbf{v}$ in the basis $\{\partial/\partial x^\alpha\}$. Hence, it is convenient to visualize the vector $\mathbf{v}$ as the derivative operator

$$\mathbf{v} \equiv \sum_\alpha v^\alpha \frac{\partial}{\partial x^\alpha}.$$

The derivative of a function f at a point p_0 along a curve $\gamma(\tau)$ can be written as

$$\mathrm{D}_\gamma f(p_0) \equiv \dot{\gamma}|_{p_0} \circ f,$$

where the notation $\dot{\gamma}|_{p_0}$ means the tangent vector to the curve $\gamma(\tau)$ at point p_0.

Remark on notation: The symbol "$\circ$" is used in general to denote the application of an operation to its argument(s). For instance, a tangent vector $\mathbf{v}$ is by definition a derivative operation, which can be applied to a function f and yields another function, $\mathbf{v} \circ f$. In this notation, "$X \circ Y$" reads "X is applied to Y." Below, I also occasionally use the symbol "$\circ$" to denote the application of a 1-form to a vector and of multilinear forms to sets of vectors. For instance, a bilinear form B applied to two vectors $\mathbf{x}$ and $\mathbf{y}$ yields a number denoted by $B \circ (\mathbf{x}, \mathbf{y})$.

Tangent vectors to curves $\gamma(\tau)$ are denoted $\dot{\gamma}$ where this does not cause confusion. I will not use the notation D_γ any more, and I will call tangent vectors simply **vectors** unless a vector is explicitly constructed as tangent to a curve or a surface within $\mathcal{M}$. Coordinate basis vectors such as $\partial/\partial x$ or $\partial/\partial y$ are denoted ∂_x and ∂_y for brevity; in that case, the subscript "x" in ∂_x is *not* an index but a local coordinate, and this will be made clear in the text. I will call attention to cases when the traditional index notation and the Einstein summation convention, such as $v^\alpha \partial_\alpha$, are used. These cases will not be common in the text. Usually, summation over indices will be written out explicitly. ■

If a vector $\mathbf{v}$ is given, the components v^α in any local coordinate system $\{x^\alpha\}$ can be found by substituting the n coordinate functions $f = x^1$, $f = x^2$, ..., $f = x^n$ into Eq. (1.3). Thus, the components v^α can be found as

$$v^\alpha = \mathbf{v} \circ x^\alpha.$$

This is a convenient way to compute the components of a vector in a coordinate system, if the vector is already known through another coordinate system or expressed in another way.

Example: Consider a two-dimensional manifold with local coordinates $\{x, y\}$. Suppose that a curve $\gamma(\tau)$ is specified as

$$\gamma(\tau) = \{x_0(\tau), y_0(\tau)\} = \{\cos\tau, 2\sin\tau\}$$

in these coordinates. Let us compute the tangent vector to the curve, $\mathbf{v}(\tau) \equiv \dot{\gamma}(\tau)$. By definition, the tangent vector $\mathbf{v}$ acts on functions $f(x, y)$ as

$$\mathbf{v} \circ f \equiv \frac{d}{d\tau} f(x_0(\tau), y_0(\tau)).$$

The tangent vectors ∂_x and ∂_y are a basis in the tangent space at any point. The components of $\mathbf{v}$ in the basis $\{\partial_x, \partial_y\}$ are found by using the coordinates x and y instead of f in the above equation. We find

$$v^x = \mathbf{v} \circ x = \frac{d}{d\tau} x_0(\tau) = -\sin\tau,$$
$$v^y = \mathbf{v} \circ y = \frac{d}{d\tau} y_0(\tau) = 2\cos\tau.$$

Thus, $\mathbf{v} = -(\sin\tau)\,\partial_x + 2(\cos\tau)\,\partial_y$. By inspection, we can express $\mathbf{v}$ directly through x and y as

$$\mathbf{v} = -\frac{1}{2} y \partial_x + 2x\partial_y.$$

Using this formula, we can define $\mathbf{v}$ as a vector *field* everywhere (not only on the curve γ). However, there are infinitely many different vector fields $\mathbf{v}$ that coincide with $\dot{\gamma}$ on the curve γ; the vector field $\mathbf{v}$ shown above is just one such field that was, by a coincidence, easy to write down in the present case.

Let us also compute the components of the vector field $\mathbf{v}$ in a different coordinate system. For example, consider a new local coordinate system $\{a, b\}$ defined by

$$a = x + y, \quad b = e^{x-y}.$$

The inverse functions are

$$x = \frac{1}{2}(a + \ln b), \quad y = \frac{1}{2}(a - \ln b).$$

The components of $\mathbf{v}$ are found as

$$v^a = \mathbf{v} \circ a = \left(-\frac{1}{2} y \partial_x + 2x\partial_y\right)(x + y) = -\frac{1}{2} y + 2x,$$
$$v^b = \mathbf{v} \circ b = \left(-\frac{1}{2} y \partial_x + 2x\partial_y\right) e^{x-y} = \left(-\frac{1}{2} y - 2x\right) e^{x-y}.$$

Substituting x, y through a, b, we can express the vector $\mathbf{v}$ in the new coordinate system as

$$\mathbf{v} = \left(\frac{3}{4} a + \frac{5}{4} \ln b\right) \partial_a + \left(-\frac{5}{4} a - \frac{3}{4} \ln b\right) b \partial_b.$$

■

Practice problems: **(a)** Determine the components of the vector $\mathbf{v} = x\partial_x + y\partial_y$ in the coordinate system $\{a, b\}$, where $a = x + y$, $b = 3xy$.
(b) Given the vector $\mathbf{u} = \partial_x + x\partial_y$, find a coordinate system $\{X, Y\}$ (where X and Y are functions of x, y) such that $\mathbf{u} = \partial_X$.
Answers: **(a)** $\mathbf{v} = a\partial_a + 2b\partial_b$. **(b)** One suitable coordinate system is $X = x$, $Y = x^2 - 2y$, and there are infinitely many others. ■

In most calculations involving directional derivatives, we will not actually need to introduce local coordinates $\{x^\alpha\}$ and the coordinate basis $\{\partial/\partial x^\alpha\}$. Instead, we will use the following properties of vectors, which follow immediately from the fact that vectors are directional derivatives at a point:

$$\mathbf{v} \circ (f + g) = \mathbf{v} \circ f + \mathbf{v} \circ g, \text{ [linearity]} \tag{1.4}$$
$$\mathbf{v} \circ (fg) = (\mathbf{v} \circ f)\, g + f\, \mathbf{v} \circ g, \text{ [Leibnitz rule]} \tag{1.5}$$
$$\mathbf{v} \circ f = 0 \text{ if } f = \text{const around } p_0. \tag{1.6}$$

For example, one can prove that the **chain rule** follows from the properties (1.4)-(1.6).

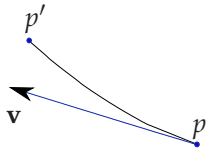

Figure 1.5: A short curve segment between points p and p' can be approximated by a tangent vector $\mathbf{v}$.

Statement 1.2.4.1: If f is a smooth function $f : \mathbb{R} \to \mathbb{R}$; g is a smooth function $\mathcal{M} \to \mathbb{R}$; and $\mathbf{v}$ is a derivation, then

$$\mathbf{v} \circ (f(g)) = \frac{df(g)}{dg} \mathbf{v} \circ g.$$

(Proof on page 175). ■

1.2.5 Tangent vectors as short curve segments

An intuitive picture of an ordinary vector is that of a piece of a straight line with an arrow at one end. In a curved space, there are in general no naturally defined "straight lines," but a sufficiently short line segment can be considered "almost straight." So one can visualize a tangent vector $\mathbf{v} \in T_p\mathcal{M}$ heuristically as a short line segment going from a point p to a neighbor point p'. Let us now explore this heuristic picture, since it provides an important intuitive link between the ordinary geometric notions in flat space and the corresponding notions in the curved space calculus.

To build a correspondence between tangent vectors and line segments, consider a curve $\gamma(\tau)$ that starts at p and goes through p'. Let us assume that p corresponds to $\tau = \tau_0$, that is, $p = \gamma(\tau_0)$, and likewise $p' = \gamma(\tau_0 + \sigma)$, where σ is heuristically a "very small" number. (In other words, we will take the limit $\sigma \to 0$ at the end). Now we would like to define a tangent vector $\mathbf{v}$ that corresponds to the short segment of the curve γ between the points p and p'.

A tangent vector is, by definition, a derivative along a curve (the derivative is applied to scalar functions). So it is natural to consider the derivative of an arbitrary function f along the curve γ at point p,

$$\dot{\gamma} \circ f \equiv \lim_{\delta\tau \to 0} \frac{f(\gamma(\tau_0 + \delta\tau)) - f(\gamma(\tau_0))}{\delta\tau}.$$

While finding the exact value of $\dot{\gamma} \circ f$ requires taking the limit $\delta\tau \to 0$, we will compute $\dot{\gamma} \circ f$ sufficiently precisely if we use the finite value $\delta\tau = \sigma$, provided that σ is sufficiently small. Then we obtain the approximate expression

$$\sigma\dot{\gamma} \circ f \approx f(p') - f(p).$$

The error of this approximation is of order σ^2.

Thus it is natural to identify the tangent vector $\sigma\dot{\gamma}$ as a representation of the short curve segment between the points p and p' (see Fig. 1.5). Let us temporarily denote this vector by $\mathbf{v} \equiv \sigma\dot{\gamma}$. The vector $\mathbf{v}$ does not depend (up to terms of order σ) on the choice of the parameter τ or the curve γ.

Given two "sufficiently close" points p and p', how can we determine the tangent vector $\mathbf{v}$ that represents the "almost straight" line between p and p'? We can choose some curve $\gamma(\tau)$ passing through these points, such that $\gamma(\tau_0) = p$ and $\gamma(\tau_1) = p'$. We can compute the vector $\dot{\gamma}$ at the point p, and then the formula for $\mathbf{v}$ is

$$\mathbf{v} = (\tau_1 - \tau_0)\,\dot{\gamma}(\tau_0) \in T_p\mathcal{M}.$$

The tangent vector $\mathbf{v}$ represents the "arrow between the points" p and p' in the sense of the formula

$$f(p') - f(p) \approx \mathbf{v} \circ f, \tag{1.7}$$

which is approximately valid for *any* smooth function f up to second-order corrections.

1.2.6 *Tangent space as space of derivations

In this section we will see that the properties (1.4)-(1.6) are a minimal necessary set of properties describing *all* the possible directional derivatives. Therefore, one may forget that D_γ was defined using some curve $\gamma(\tau)$, drop the subscript γ, call any map D from functions to numbers satisfying the above properties a **derivation** at p_0, and define the tangent space $T_{p_0}\mathcal{M}$ as the set of all possible derivations. This somewhat abstract definition of $T_p\mathcal{M}$ is more elegant from the purely mathematical point of view but entirely equivalent to the earlier definitions, because every derivation D satisfying the properties (1.4)-(1.6) is a directional derivative D_γ along some curve γ. In other words, the only way to differentiate a function is to take a derivative along a curve in some direction.

To prove these statements, let us perform explicit calculations in a local coordinate system $\{x^\alpha\}$. If D is a derivation satisfying Eqs. (1.4)-(1.6) and $f(p)$ is a smooth function, then we can apply the chain rule (see Statement 1.2.4.1) and compute

$$\mathrm{D} \circ f(p) = \sum_\alpha \frac{\partial f}{\partial x^\alpha} \mathrm{D} \circ x^\alpha(p),$$

where $\mathrm{D} \circ x^\alpha$ is the application of the derivation D to the scalar coordinate function $x^\alpha(p)$. Hence,

$$\mathrm{D} \circ f(p) = \left[\sum_\alpha (\mathrm{D} \circ x^\alpha) \frac{\partial}{\partial x^\alpha} \right] f(p) \equiv \sum_\alpha u^\alpha \frac{\partial}{\partial x^\alpha},$$

where u^α are coefficients defined by $u^\alpha \equiv \mathrm{D} \circ x^\alpha$. It is always possible to find a curve $\gamma(\tau)$ with the derivatives $d\gamma^\alpha/d\tau = u^\alpha$ at the point p. Thus any derivation D is equivalent to a tangent vector in the previous sense.

1.2.7 Vector fields and flows

A **vector field** on a manifold $\mathcal{M}$ is obtained by choosing a tangent vector $\mathbf{v}|_p$ at each point $p \in \mathcal{M}$. Most of the time, we will only need a vector field defined in a neighborhood of some point, rather than on the entire manifold $\mathcal{M}$.

Another way to visualize a vector field is to imagine that the manifold is filled by nonintersecting curves going through all its points (or at least through every point in some neighborhood); see Fig. 1.6. Such a collection of curves is called a **congruence** of curves or a **flow** of a vector field. If a congruence of curves is given, derivatives along curves at various points p will determine tangent vectors $\mathbf{v}|_p$ at all points and thus define a vector field $\mathbf{v}$. This vector field $\mathbf{v}$ is called **tangent** to the flow or the **generator** of the flow. The curves in the congruence are also called the **flow lines** or **orbits** of the vector field $\mathbf{v}$. In our notation, a curve γ is an orbit of a vector field $\mathbf{v}$ iff $\dot\gamma(\tau) = \mathbf{v}|_{\gamma(\tau)}$ at every point $p \equiv \gamma(\tau)$ along the curve.

Orbits of a nonzero, smooth vector field always exist and are unique because they are solutions of ordinary differential equations. To see this, consider a local coordinate system $\{x^\alpha\}$ where a vector field $\mathbf{v}$ is specified by its components v^α as

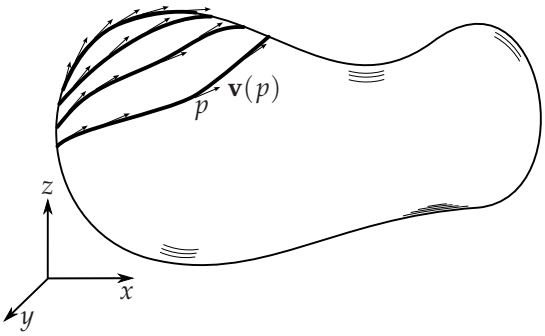

Figure 1.6: A vector field $\mathbf{v}$ tangent to a congruence of curves, which are the orbits of $\mathbf{v}$. For each point p, the vector $\mathbf{v}(p)$ belongs to the tangent space $T_p\mathcal{M}$ at p.

$\mathbf{v} = \sum_\alpha v^\alpha \mathbf{e}_\alpha$ (the components v^α can be found as $v^\alpha = \mathbf{v} \circ x^\alpha$). A curve $\gamma(\tau)$ is specified by n coordinate functions $\gamma^\alpha(\tau)$. The condition $\dot\gamma(\tau) = \mathbf{v}|_{\gamma(\tau)}$ translates into a system of differential equations,

$$\frac{d\gamma^\alpha(\tau)}{d\tau} = v^\alpha|_{p=\gamma(\tau)}.$$

We may choose the initial condition as $\gamma^\alpha(0) = x^\alpha_{(0)}$, where $x^\alpha_{(0)}$ are the local coordinates of an initial point of the curve. Since the vector field $\mathbf{v}$ is smooth, such a system of equations always has a unique solution $\gamma^\alpha(\tau)$ by the well-known theorem of differential calculus. This solution $\gamma^\alpha(\tau)$ represents the required curve $\gamma(\tau)$ in the local coordinate system. Of course, this system of equations may be complicated and so it is not always possible to obtain an *explicit* formula for the solution $\gamma^\alpha(\tau)$. In most cases, an approximate solution can be found numerically. Since solutions $\gamma^\alpha(\tau)$ exist and are unique starting from any point p, the orbits of a nonzero, smooth vector field always form a congruence that (at least locally) fills space.

Example: Given a vector field $\mathbf{v}$, consider the orbits of $\mathbf{v}$ going through each point of some neighborhood. Let τ be the parameter along each of the orbits $\gamma(\tau)$, such that $\dot\gamma = \mathbf{v}$ everywhere. This defines τ as a scalar function on the manifold. We can compute the scalar function $\mathbf{v} \circ \tau$. Along every orbit, we have

$$\mathbf{v} \circ \tau = \frac{d}{d\tau}\tau = 1,$$

thus $\mathbf{v} \circ \tau = 1$ everywhere. ■

If $\{x^\mu\}$ are local coordinates around a point p_0, we can consider a congruence of curves of varying x^0 and constant x^1, x^2, ... The tangent vector field to this congruence is $\partial/\partial x^0$. This congruence consists of "curved coordinate axes" of the coordinate x^0. Similarly, the other coordinate basis vectors $\partial/\partial x^1$, $\partial/\partial x^2$, etc., can be seen as tangent vectors to a congruence of curves made by other coordinate axes.

Diffeomorphism flow. The flow of a vector field can be visualized as a collection of diffeomorphisms, i.e. of smooth maps $\mathcal{M} \to \mathcal{M}$ that move points in the manifold. For each value of the parameter τ, there is a diffeomorphism, denoted by $\exp(\tau\mathbf{v})$. This diffeomorphism, by definition, moves each point p to a point p' which is at parameter distance τ along the orbit of $\mathbf{v}$ that starts at p. ■

Practice problem: A diffeomorphism flow is given explicitly by the equations

$$\{x, y, z\} \mapsto \{x + \tau y, y, z - \tau x\}$$

in a local coordinate system $\{x, y, z\}$. Compute the vector field $\mathbf{v}$ that generates this flow.

Answer: $\mathbf{v} = y\partial_x - x\partial_z$. ■

1.2.8 *Tangent bundle

The set of all the tangent spaces (one tangent space $T_p\mathcal{M}$ for each point $p \in \mathcal{M}$) is denoted $T\mathcal{M}$ and called the **tangent bundle** of the manifold $\mathcal{M}$. The tangent bundle is itself a manifold with dimension $2n$ if $\mathcal{M}$ has dimension n. This manifold is perhaps not easy to visualize as a whole, but we will not need to do anything complicated with the tangent bundle $T\mathcal{M}$ as a manifold. Let us only verify that $T\mathcal{M}$ has the structure of a smooth manifold.

By assumption, $\mathcal{M}$ is a smooth manifold and, as such, can be covered by a set of charts. Consider one such chart that maps a subset $\mathcal{U}$ of $\mathcal{M}$ into a subset $\mathcal{V}$ of $\mathbb{R}^n$. Tangent vectors at points $p \in \mathcal{U}$ are mapped into tangent vectors in $\mathcal{V}$, which are simply vectors in $\mathbb{R}^n$ since $\mathcal{V}$ is a portion of the flat Euclidean space $\mathbb{R}^n$. Therefore, the part of $T\mathcal{M}$ consisting of tangent planes to points $p \in \mathcal{U}$ is one-to-one mapped into $\mathcal{V} \times \mathbb{R}^n$. Since the set $\mathcal{V} \times \mathbb{R}^n$ is a subset of $\mathbb{R}^n \times \mathbb{R}^n = \mathbb{R}^{2n}$, we have established that $T\mathcal{M}$ can be covered by patches, and one can find charts that map these patches into $\mathbb{R}^{2n}$. In other words, the tangent bundle $T\mathcal{M}$ is a smooth $2n$-dimensional manifold.

Remark: It is important to realize that the tangent bundle, as a whole manifold, is not necessarily equivalent (i.e. diffeomorphic) to the direct product $\mathcal{M} \times \mathbb{R}^n$, even though every small patch of $T\mathcal{M}$ is diffeomorphic to a patch of $\mathcal{M} \times \mathbb{R}^n$. The reason is the same as that for the manifold $\mathcal{M}$ not being diffeomorphic to $\mathbb{R}^n$, even though small patches of $\mathcal{M}$ are diffeomorphic to patches of $\mathbb{R}^n$. Namely, the patches may be glued together in a nontrivial way such that the global geometry of $\mathcal{M}$ is different from that of $\mathbb{R}^n$. A detailed consideration would lead us too far into differential topology, so I will give only a qualitative example. Suppose a diffeomorphism $\phi : \mathcal{M} \times \mathbb{R}^n \to T\mathcal{M}$ exists, such that every vector space $p \times \mathbb{R}^n$, where $p \in \mathcal{M}$, is mapped one-to-one onto the vector space $T_p\mathcal{M}$ via an invertible linear transformation. Then for every $p \in \mathcal{M}$ one may apply ϕ to $p \times \{1, 0, 0, ..., 0\} \in \mathcal{M} \times \mathbb{R}^n$ and obtain a nonzero vector $\mathbf{e}_1(p) \in T_p\mathcal{M}$ at every point p. However, such a vector field $\mathbf{e}_1$ sometimes cannot exist; a typical example is the 2-sphere S^2. It is known that there exists no everywhere nonvanishing and smooth tangent vector field on a 2-sphere. This property is called the "impossibility of combing a sphere" (see Statement 6.3.2.1). So it is impossible to choose even a single nonzero vector field $\mathbf{e}_1$ that smoothly varies throughout the sphere. Thus, a diffeomorphism $\phi : S^2 \times \mathbb{R}^2 \to TS^2$ does not exist. A tangent bundle which is diffeomorphic to $\mathcal{M} \times \mathbb{R}^n$ is called **trivial**. It follows that the 2-sphere S^2 has a nontrivial tangent bundle. ■

1.2.9 Tensor fields

A **covector** is a linear map ω from vectors $\mathbf{v}$ into numbers. I assume that you know this standard construction from linear algebra, and I will only summarize the concepts we will need. In the index notation, covectors are denoted by letters with a subscript index, ω_α. In our notation, a covector ω applied to a vector $\mathbf{v}$ gives a number $\omega \circ \mathbf{v}$ (read "ω applied to $\mathbf{v}$"); sometimes we also write $\omega(\mathbf{v})$ for the same thing. (In the index notation, this is $\omega_\alpha v^\alpha$.)

Covectors themselves naturally form a vector space called the **dual space**. The dual space to the tangent space $T_p\mathcal{M}$ is denoted by $T_p^*\mathcal{M}$ and consists of covectors acting on tangent vectors at point p. The space $T_p^*\mathcal{M}$ is also called the **cotangent space** at point p.

A **covector field**, also called a **1-form**, is a choice of a covector $\omega|_p$ from the dual space (cotangent space) $T_p^*\mathcal{M}$ at each point p. A 1-form ω can be seen as a linear map from vector fields to functions on $\mathcal{M}$, by acting pointwise on a vector field $\mathbf{v}$ and producing a scalar function f,

$$\omega : \mathbf{v} \mapsto f; \quad \omega|_p \circ \mathbf{v}|_p \equiv f(p).$$

An example of a 1-form is the mapping of a vector $\mathbf{v}$ into the derivative of a *fixed* function f in the direction $\mathbf{v}$, that is, the map $\mathbf{v} \mapsto \mathbf{v} \circ f$ for a fixed f. Clearly, this map is linear in $\mathbf{v}$ and hence, by definition, it is a 1-form. This 1-form is called the **gradient** of the function f and is denoted by $\mathrm{d}f$. Thus, for a fixed function f we have defined the 1-form $\mathrm{d}f$ by its action on a vector field $\mathbf{v}$ as follows,

$$(\mathrm{d}f) \circ \mathbf{v} \equiv \mathbf{v} \circ f.$$

Let us recall that tangent vectors can be understood as an approximate representation of short curve segments (see Sec. 1.2.5). If a vector $\mathbf{v}$ represents a short curve segment between points p and p' then the 1-form $\mathrm{d}f$ acting on $\mathbf{v}$ yields approximately $f(p') - f(p)$, according to Eq. (1.7).

Example: Consider a two-dimensional manifold with a local coordinate system $\{x, y\}$. The gradient of the coordinate function x is the 1-form $\mathrm{d}x$. By definition, this 1-form acts on vector fields $\mathbf{u}$ as $(\mathrm{d}x) \circ \mathbf{u} = \mathbf{u} \circ x$. For instance, if $\mathbf{u} = \partial_x$ then obviously $\mathbf{u} \circ x = 1$. In this notation, we obtain the equation

$$(\mathrm{d}x) \circ \mathbf{u} = (\mathrm{d}x) \circ \partial_x = 1,$$

which may look unusual or even confusing at first glance. However, it is only the notation that is unusual; the results are always consistent.

Consider another vector field $\mathbf{v} = y^2\partial_x$ and a function $f(x, y) = x^2y^3$. The field $\mathbf{v}$ acts on f and produces the function

$$\mathbf{v} \circ f = y^2 \frac{\partial}{\partial x} f(x, y) = 2xy^5.$$

The 1-form $\mathrm{d}f$ can be computed using standard rules of calculus as

$$\mathrm{d}f = 2xy^3\mathrm{d}x + 3x^2y^2\mathrm{d}y.$$

However, note that now $\mathrm{d}f$, $\mathrm{d}x$, and $\mathrm{d}y$ are well-defined objects (1-forms) rather than heuristic "infinitesimals."

The action of the 1-form $\mathrm{d}f$ on the vector $\mathbf{v}$ is, by definition,

$$(\mathrm{d}f) \circ \mathbf{v} \equiv \mathbf{v} \circ f = 2xy^5.$$

Let us see how this result can be obtained using the explicit expressions for $\mathrm{d}f$ and $\mathbf{v}$. We have

$$\begin{aligned}(\mathrm{d}f) \circ \mathbf{v} &= \left(2xy^3\mathrm{d}x + 3x^2y^2\mathrm{d}y\right) \circ y^2\partial_x \\ &= \left(2xy^5\right)\mathrm{d}x \circ \partial_x + \left(3x^2y^4\right)\mathrm{d}y \circ \partial_x.\end{aligned}$$

By definition, $\mathrm{d}x \circ \partial_x = 1$ and $\mathrm{d}x \circ \partial_y = 0$, therefore $(\mathrm{d}f) \circ \mathbf{v} = 2xy^5$ as before. ■

Practice problems: **(a)** Given the vector $\mathbf{v} = \partial_x + x\partial_y$, find all 1-forms $\omega = a(x,y)\mathrm{d}x + b(x,y)\mathrm{d}y$ such that $\omega \circ \mathbf{v} = 0$. **(b)** Given the 1-form $\omega = x\mathrm{d}y + y\mathrm{d}x$, find all vectors $\mathbf{v} = a(x,y)\partial_x + b(x,y)\partial_y$ such that $\omega \circ \mathbf{v} = 0$. **(c)** The 1-form ω is defined through its action on an arbitrary vector $\mathbf{a}$ by $\omega \circ \mathbf{a} \equiv y^2\,(\mathbf{a} \circ x) - 2\mathbf{a} \circ y$; express ω through $\mathrm{d}x$, $\mathrm{d}y$. ■

Similarly to vector fields, one can define **tensor fields** of arbitrary rank. For example, if a linear transformation $L|_p$ is defined in each tangent space $T_p\mathcal{M}$, such that $L|_p\,\mathbf{v}|_p$ is another vector from $T_p\mathcal{M}$, then we have a tensor field L on the manifold $\mathcal{M}$. Also, one defines n-**forms** as totally antisymmetric tensor fields of rank (0,n). (The calculus of n-forms is explained in more detail in Sec. 1.4.) The operations of tensor product $L_1 \otimes L_2$, the exterior product $\omega_1 \wedge \omega_2$ of n-forms, and tensor contraction (for example, the contraction $\omega \circ \mathbf{v}$ of a 1-form and a vector) are naturally defined on tensor fields. These algebraic operations are performed on tensors in each tangent space $T_p\mathcal{M}$ separately.

1.2.10 Commutator of vector fields

Since a vector field $\mathbf{v}$ can be viewed as a (first-order) differential operator acting on functions as $\mathbf{v} : f \mapsto \mathbf{v} \circ f$, one may consider the **commutator** of two such operators,

$$[\mathbf{u}, \mathbf{v}] \circ f \equiv \mathbf{u} \circ (\mathbf{v} \circ f) - \mathbf{v} \circ (\mathbf{u} \circ f)\,. \tag{1.8}$$

It turns out (see Statement 1.2.10.1) that the result is *again a derivation*, i.e. a first-order differential operator acting on functions, even though it may appear at first glance that $[\mathbf{u}, \mathbf{v}] \circ f$ involves second-order derivatives of f. For this reason, the operation $[\mathbf{u}, \mathbf{v}] \circ f$ is equivalent to some vector field acting on f. This vector field is denoted by $[\mathbf{u}, \mathbf{v}]$ and is called the **commutator** of the fields $\mathbf{u}$ and $\mathbf{v}$.

Statement 1.2.10.1: **(a)** It follows from the definition (1.8) and Eqs. (1.4)-(1.5) that the commutator $[\mathbf{u}, \mathbf{v}]$ satisfies the defining properties (1.4)–(1.6) of a derivation. Hence, $[\mathbf{u}, \mathbf{v}]$ is itself a vector field. **(b)** The components c^α of the commutator $[\mathbf{u}, \mathbf{v}]$ in terms of the components u^α and v^α in a local coordinate system $\{x^\alpha\}$ with a coordinate basis $\{\mathbf{e}_\alpha\}$,

$$\mathbf{u} = \sum_\alpha u^\alpha \mathbf{e}_\alpha, \quad \mathbf{v} = \sum_\alpha v^\alpha \mathbf{e}_\alpha, \quad [\mathbf{u}, \mathbf{v}] \equiv \sum_\alpha c^\alpha \mathbf{e}_\alpha,$$

are given by the expression

$$([\mathbf{u}, \mathbf{v}])^\beta \equiv c^\beta = u^\alpha \frac{\partial v^\beta}{\partial x^\alpha} - v^\alpha \frac{\partial u^\beta}{\partial x^\alpha}. \tag{1.9}$$

(c) If $\mathbf{e}_\alpha \equiv \partial/\partial x^\alpha$ are the coordinate basis vector fields defined through a local coordinate system $\{x^\alpha\}$ then $\left[\mathbf{e}_\alpha, \mathbf{e}_\beta\right] = 0$.

Proof of Statement 1.2.10.1: **(a)** The properties (1.4) and (1.6) are obvious; the nontrivial one is the Leibnitz rule (1.5). To verify the Leibnitz rule, we compute $[\mathbf{u}, \mathbf{v}] \circ (fg)$, where f and g are smooth functions:

$$\begin{aligned}
[\mathbf{u}, \mathbf{v}] \circ (fg) &= \mathbf{u} \circ (g\mathbf{v} \circ f + f\mathbf{v} \circ g) - \mathbf{v} \circ (g\mathbf{u} \circ f + f\mathbf{u} \circ g) \\
&= g\mathbf{u} \circ (\mathbf{v} \circ f) + f\mathbf{u} \circ (\mathbf{v} \circ g) - g\mathbf{v} \circ (\mathbf{u} \circ f) - f\mathbf{v} \circ (\mathbf{u} \circ g) \\
&\quad + (\mathbf{u} \circ g)\,(\mathbf{v} \circ f) + (\mathbf{u} \circ f)\,(\mathbf{v} \circ g) \\
&\quad - (\mathbf{v} \circ g)\,(\mathbf{u} \circ f) - (\mathbf{v} \circ f)\,(\mathbf{u} \circ g) \\
&= ([\mathbf{u}, \mathbf{v}] \circ f)\,g + f\,[\mathbf{u}, \mathbf{v}] \circ g.
\end{aligned}$$

(b) A straightforward calculation using the representation of the basis fields $\mathbf{e}_\alpha$ as differential operators yields

$$\begin{aligned}
[\mathbf{u}, \mathbf{v}] &= \sum_{\alpha,\beta} \left(u^\alpha \frac{\partial}{\partial x^\alpha}\right) \left(v^\beta \frac{\partial}{\partial x^\beta}\right) - \sum_{\alpha,\beta} \left(v^\alpha \frac{\partial}{\partial x^\alpha}\right) \left(u^\beta \frac{\partial}{\partial x^\beta}\right) \\
&= \sum_{\alpha,\beta} u^\alpha v^\beta \frac{\partial}{\partial x^\alpha}\frac{\partial}{\partial x^\beta} + \sum_{\alpha,\beta} \left(u^\alpha \frac{\partial v^\beta}{\partial x^\alpha}\right) \frac{\partial}{\partial x^\beta} \\
&\quad - \sum_{\alpha,\beta} v^\alpha u^\beta \frac{\partial}{\partial x^\alpha}\frac{\partial}{\partial x^\beta} - \sum_{\alpha,\beta} \left(v^\alpha \frac{\partial u^\beta}{\partial x^\alpha}\right) \frac{\partial}{\partial x^\beta} \\
&= \sum_{\alpha,\beta} \left(u^\alpha \frac{\partial v^\beta}{\partial x^\alpha} - v^\alpha \frac{\partial u^\beta}{\partial x^\alpha}\right) \frac{\partial}{\partial x^\beta} \equiv \sum_\beta c^\beta \frac{\partial}{\partial x^\beta}.
\end{aligned}$$

This shows that $[\mathbf{u}, \mathbf{v}]$ is indeed a first-order differential operator, and also gives an explicit formula for c^β.

(c) For a smooth function $f(x^1, ..., x^n)$, we have

$$\frac{\partial}{\partial x^\alpha}\frac{\partial}{\partial x^\beta} f = \frac{\partial}{\partial x^\beta}\frac{\partial}{\partial x^\alpha} f$$

according to a well-known theorem of calculus. Therefore the basis vector fields $\mathbf{e}_\alpha$ commute with each other. ■

Discussion: Statement 1.2.10.1 presents two ways of proving the fact that $[\mathbf{a}, \mathbf{b}]$ is a first-order differential operation, in parts (a) and (b). The first proof (a) is performed purely algebraically, using the abstract definition of tangent vectors as (unspecified) operations $\mathbf{v} \circ (...)$ with certain properties. The other proof (b) uses a specific representation of tangent vectors as differential operators $\sum_\alpha v^\alpha \partial/\partial x^\alpha$ in a local coordinate system $\{x^\alpha\}$.

Both the proofs have certain advantages and disadvantages. The proof (a) is very general because it uses only some properties of $\mathbf{v} \circ (...)$. So this proof applies not only to directional derivatives but to *any* operation with the properties (1.4)-(1.5), such as the Lie derivative and the covariant derivative that we will use later. The proof (a) applies equally well to derivatives and to finite difference operators, to operators in infinite-dimensional spaces, and so on. The proof (a) is also more "elegant" because it is index-free, does not involve unnecessary structures (such as a local coordinate system and a basis), and does not depend on the explicit representation of vectors in coordinates. This is the kind of proof a mathematician would be looking for.

On the other hand, the proof (b) is conceptually easier to understand (especially for beginners) because it consists of a specific computation in a familiar and elementary context, namely, manipulations with partial derivatives of functions. Also, the result of the proof (b) is a directly usable formula for the components of the vector $[\mathbf{u}, \mathbf{v}]$, while the proof (a) merely shows that $[\mathbf{u}, \mathbf{v}]$ is a well-defined vector. In the text that follows, I will usually not need explicit formulas for components of vectors, and therefore index-free calculations will be preferable. However, in every case one can translate the final results into components in a local coordinate system. ■

It is important to understand the geometric interpretation of the commutator. Imagine drawing the orbits of two vector fields $\mathbf{a}$ and $\mathbf{b}$ (Fig. 1.7). Let us start from a point p_0, follow the orbit of $\mathbf{a}$ for a small interval $\delta\tau$ of the parameter τ, and then follow the orbit of $\mathbf{b}$ for another interval $\delta\tau$; we will arrive at a point p. If we again start from p_0 but first follow the lines of $\mathbf{b}$ and then the lines of $\mathbf{a}$ for the same parameter distances $\delta\tau$, we will arrive, in general, at a different point p'. In the

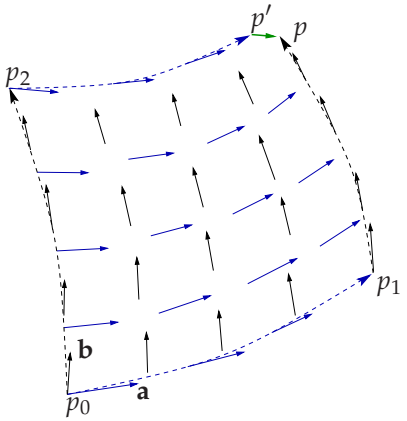

Figure 1.7: The commutator $[\mathbf{a}, \mathbf{b}]$ of two vector fields measures the discrepancy between the points p and p' obtained by following the orbits of the vector fields in different order, starting at a point p_0, for small intervals $\delta\tau$ of the parameter. The blue and the black arrows are the vector fields $\mathbf{a}$ and $\mathbf{b}$, while the dotted lines are the orbits of these vector fields. The point p is obtained by following the line $p_0 \to p_1$ and then the line $p_1 \to p$, while the point p' is obtained by following the line $p_0 \to p_2$ and then $p_2 \to p'$. The small green vector between the points p and p' is equal to $[\mathbf{a}, \mathbf{b}]\,\delta\tau^2$ in the limit of small distances.

limit $\delta\tau \to 0$, the points p and p' will become infinitesimally close to each other and to p_0. As shown in Statement 1.2.10.2, the line between p and p' specifies a well-defined vector in the limit $\delta\tau \to 0$, namely the vector $[\mathbf{a}, \mathbf{b}]\,\delta\tau^2$.

Statement 1.2.10.2: Consider an arbitrary smooth function f, vector fields $\mathbf{a}, \mathbf{b}$, and the points p_0, p, p' in the notation of Fig. 1.7. The commutator $[\mathbf{a}, \mathbf{b}]$ describes the difference between $f(p)$ and $f(p')$ in the following way,

$$\lim_{\delta\tau\to 0} \frac{f(p) - f(p')}{\delta\tau^2} = ([\mathbf{a}, \mathbf{b}] \circ f)|_{p_0}\,.$$

(Proof on page 175.) ■

Practice problems: **a)** Assuming a local coordinate system $\{x, y, z\}$, determine the commutators $[\mathbf{a}, \mathbf{b}]$, $[\mathbf{a}, \mathbf{c}]$, and $[\mathbf{b}, \mathbf{c}]$, where $\mathbf{a} = x\partial_x + y\partial_y + z\partial_z$, $\mathbf{b} = x\partial_y - y\partial_x$, $\mathbf{c} = \partial_z$.
b) A surface is specified as $f(p) = 0$, where f is a given scalar function. A vector field $\mathbf{v}$ is tangent to the surface if $\mathbf{v} \circ f = 0$ on the surface. Show that $[\mathbf{u}, \mathbf{v}]$ is also tangent to the surface if $\mathbf{u}$ and $\mathbf{v}$ are tangent.

Answers: **a)** $[\mathbf{a}, \mathbf{b}] = 0$, $[\mathbf{a}, \mathbf{c}] = -\partial_z$, $[\mathbf{b}, \mathbf{c}] = 0$. ■

1.2.11 Connecting vectors

If two vector fields $\mathbf{a}$ and $\mathbf{b}$ are such that $[\mathbf{a}, \mathbf{b}] = 0$, the vector field $\mathbf{a}$ is called a **connecting vector** for the field $\mathbf{b}$. (Since $[\mathbf{a}, \mathbf{b}] = -[\mathbf{b}, \mathbf{a}]$, the field $\mathbf{b}$ is then also a connecting vector for the field $\mathbf{a}$. One also says that the vector fields $\mathbf{a}$ and $\mathbf{b}$ **commute** with eath other.) The notion of a connecting vector turns out to be very useful in index-free calculations because of the following geometric interpretation.

Consider a vector field $\mathbf{v}$ and a congruence γ of its orbits (see Fig. 1.8). A connecting vector field $\mathbf{c}$ for the field $\mathbf{v}$ represents arrows between nearby points with equal parameter

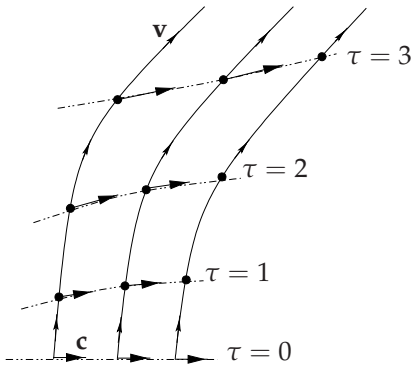

Figure 1.8: Connecting vectors $\mathbf{c}$ (thick arrows) point in the direction of equal parameter τ along a congruence of orbits of $\mathbf{v}$ (thin lines). Dotted lines are the orbits of $\mathbf{c}$, which are also the lines of equal τ.

τ. This construction can be described more formally as follows. For the convenience of our consideration, let us label a subset of the orbits by an additional real-valued parameter s, so that $\gamma(\tau; s)$ is a smooth function $\mathbb{R} \times \mathbb{R} \to \mathcal{M}$ and the tangent vector $\mathbf{v}$ is expressed as $\mathbf{v} = \text{“}\partial\gamma/\partial\tau\text{”}$. Using the function $\gamma(\tau; s)$, we may also consider the curves $\tilde{\gamma}(s) \equiv \gamma(\tau_0; s)$ for a fixed value of $\tau = \tau_0$. The curves $\tilde{\gamma}(s)$ go “straight across” the curves $\gamma(\tau)$, intersecting the latter at points of equal τ. We can then define the tangent vector field $\mathbf{c}$ to the curves $\tilde{\gamma}$ in the usual way, $\mathbf{c} = \text{“}\partial\gamma/\partial s\text{”}$. The geometric interpretation of the vector $\mathbf{c}$ is that it points “sideways” with respect to the curves $\gamma(\tau)$, connecting a point p on a curve $\gamma(\tau; s)$ to a nearby point p' on a neighbor curve, $\gamma(\tau; s')$, where s and s' are “infinitesimally close” while τ stays constant. In other words, the connecting vector $\mathbf{c}$ “connects” neighboring orbits $\gamma(\tau)$ of the field $\mathbf{v}$ at “corresponding” points, i.e. at points with the same value of the parameter τ. Thus, the orbits of $\mathbf{c}$ are also lines of constant τ. Now it is easy to check that $\mathbf{c}$ is a connecting vector for $\mathbf{v}$ according to the above definition.

Calculation 1.2.11.1: For the vector fields $\mathbf{v}$ and $\mathbf{c}$ defined in the preceding paragraph through the function $\gamma(\tau; s)$, we have $[\mathbf{c}, \mathbf{v}] = 0$. (Details on page 175.) ■

Since the parameter τ for different orbits of a given vector field $\mathbf{v}$ can be started with different values at different points, there are infinitely many connecting vector fields for a given field $\mathbf{v}$. The freedom of selecting a connecting field for $\mathbf{v}$ is the same as the freedom of choosing the initial values of the parameter τ on each orbit of $\mathbf{v}$.

If a vector $\mathbf{c}$ is fixed at one point p_0, say $\mathbf{c}(p_0) = \mathbf{c}_0$, one can always select $\mathbf{c}(p)$ in the neighborhood of p_0 in such a way that the field $\mathbf{c}$ commutes with $\mathbf{v}$ everywhere. To find such $\mathbf{c}(p)$, one needs to solve the equation $[\mathbf{c}, \mathbf{v}] = 0$. It is easy to see from the coordinate representation

$$[\mathbf{c}, \mathbf{v}] = \sum_\beta \mathbf{e}_\beta \left(\sum_\alpha c^\alpha \frac{\partial v^\beta}{\partial x^\alpha} - \sum_\alpha v^\alpha \frac{\partial c^\beta}{\partial x^\alpha} \right)$$

that the equation $[\mathbf{c}, \mathbf{v}] = 0$ is a linear differential equation for the unknown vector field $\mathbf{c}$ that involves only the derivative of $\mathbf{c}$ along the orbits of $\mathbf{v}$. Therefore, it is always possible to solve that equation with any given initial condition $\mathbf{c}_0$.

Remark: Since $[\mathbf{v}, \mathbf{v}] = 0$, every field $\mathbf{v}$ can be formally considered as a connecting field for itself, although there seems to be no useful geometric interpretation of that statement. ■

Vector fields ∂_{x^μ} defined through a coordinate system $\{x^\mu\}$ always commute. In general, any vector field $\mathbf{v}$ can be considered as a coordinate derivative ∂_x in *some* local coordinate system that includes x as one of the coordinates. To show this, it is sufficient to demonstrate that there exists a basis of connecting fields that includes $\mathbf{v}$; then the coordinate system can be constructed by using the orbits of these vector fields.

Statement 1.2.11.2: For any vector field $\mathbf{v} \neq 0$ given on an N-dimensional manifold $\mathcal{M}$, there exist connecting fields $\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_{N-1}$ such that $\{\mathbf{v}, \mathbf{c}_1, \mathbf{c}_2, ...\mathbf{c}_{N-1}\}$ is a basis at any point p in a neighborhood of an initial point p_0. (Proof on page 175.) ■

The geometric interpretation of a connecting basis can be understood from Fig. 1.7. If every pair of basis vector fields $\mathbf{e}_a, \mathbf{e}_b$ are connecting, there is no discrepancy between points obtained by following the coordinate lines in different order. The absence of this discrepancy is necessary for the existence of a coordinate system $\{x^\mu\}$ such that $\mathbf{e}_a = \partial/\partial x^a$.

Practice problems: **(a)** In a local coordinate system $\{x, y\}$, determine all connecting vector fields $\mathbf{c}$ for the vector $\mathbf{u} = x\partial_y$, i.e. all $\mathbf{c} = a(x,y)\partial_x + b(x,y)\partial_y$ such that $[\mathbf{c}, \mathbf{u}] = 0$ everywhere.
(b) Find at least two nonparallel connecting vectors for $\mathbf{v} = x\partial_y - x\partial_z$.
(c) In a local coordinate system $\{x, y, z\}$, find an explicit basis of connecting vector fields $\mathbf{c}_j$, $j = 1, 2, 3$, for the vector $\mathbf{v} = x\partial_y + y\partial_z$.

Answers: **(a)** $\mathbf{c} = f(x)\mathbf{u} + g(x)\left(x\partial_x + y\partial_y\right)$, where f and g are arbitrary functions of x. **(b)** For instance, ∂_y and ∂_z. **(c)** One such basis is $\mathbf{c}_1 = x\partial_y + y\partial_z$, $\mathbf{c}_2 = \partial_z$, $\mathbf{c}_3 = x^2\partial_x + xy\partial_y + \left(y^2 - xz\right)\partial_z$. ■

1.3 Lie derivative

A vector field $\mathbf{v}$ acts as a directional derivative on scalar functions. The **Lie derivative** $\mathcal{L}_\mathbf{v}$ with respect to a vector field $\mathbf{v}$ is an important differential operation that can be applied not only to scalars but also to all tensor fields. We now deduce the formulas for $\mathcal{L}_\mathbf{v}$ in an **axiomatic way**, i.e. by assuming some desirable properties of this operation. Later we will give a geometric interpretation of $\mathcal{L}_\mathbf{v}$.

We would like $\mathcal{L}_\mathbf{v}$ to act as the usual directional derivative on scalar functions f,

$$\mathcal{L}_\mathbf{v} f = \mathbf{v} \circ f. \tag{1.10}$$

Additionally, $\mathcal{L}_\mathbf{v}$ should have the following properties, fully analogous to Eqs. (1.4)-(1.5), but applicable to arbitrary tensors A, B:

$$\mathcal{L}_\mathbf{v}(A + B) = \mathcal{L}_\mathbf{v}(A) + \mathcal{L}_\mathbf{v}(B), \tag{1.11}$$
$$\mathcal{L}_\mathbf{v}(A \otimes B) = \mathcal{L}_\mathbf{v}(A) \otimes B + A \otimes \mathcal{L}_\mathbf{v}(B), \tag{1.12}$$
$$\mathcal{L}_\mathbf{v}(A \circ B) = \mathcal{L}_\mathbf{v}(A) \circ B + A \circ \mathcal{L}_\mathbf{v}(B), \tag{1.13}$$

where $A \otimes B$ is the tensor product and $A \circ B$ here denotes any "pairing" of the tensors A and B (e.g. $\omega \circ \mathbf{v}$ and also $\mathbf{v} \circ f$ are "pairings" in this sense). As we will see, these properties uniquely define the action of $\mathcal{L}_\mathbf{v}$ on any tensor.

1.3.1 Commutator as Lie derivative

For instance, if f is a scalar and $\mathbf{u}, \mathbf{v}$ are vectors, we expect to have

$$\mathcal{L}_\mathbf{v}(\mathbf{u} \circ f) = [\mathcal{L}_\mathbf{v}(\mathbf{u})] \circ f + \mathbf{u} \circ \mathcal{L}_\mathbf{v}(f).$$

Using Eq. (1.10), this is rewritten as

$$\mathbf{v} \circ (\mathbf{u} \circ f) = [\mathcal{L}_\mathbf{v}(\mathbf{u})] \circ f + \mathbf{u} \circ (\mathbf{v} \circ f).$$

Thus,

$$[\mathcal{L}_\mathbf{v}(\mathbf{u})] \circ f = \mathbf{v} \circ (\mathbf{u} \circ f) - \mathbf{u} \circ (\mathbf{v} \circ f) = [\mathbf{v}, \mathbf{u}] \circ f.$$

In other words, the Lie derivative $\mathcal{L}_\mathbf{v}$ of a vector field $\mathbf{u}$ coincides with the commutator of $\mathbf{v}$ and $\mathbf{u}$,

$$\mathcal{L}_\mathbf{v}\mathbf{u} = [\mathbf{v}, \mathbf{u}] = -\mathcal{L}_\mathbf{u}\mathbf{v}.$$

In the index notation, we have, according to Eq. (1.9),

$$(\mathcal{L}_\mathbf{v}\mathbf{u})^\alpha = \sum v^\beta \frac{\partial u}{\partial x^\beta}^\alpha - u^\beta \frac{\partial v^\alpha}{\partial x^\beta}. \tag{1.14}$$

It is important to realize that the Lie derivative $\mathcal{L}_\mathbf{v}\mathbf{u}$ depends on the *derivatives* of $\mathbf{v}$ (not only on the value of $\mathbf{v}$ and the derivatives of $\mathbf{u}$).

Calculation 1.3.1.1: The Lie derivative $\mathcal{L}_\mathbf{v}\mathbf{u}$ involves derivatives of $\mathbf{v}$. In particular, if $\mathbf{u}, \mathbf{v}$ are vector fields and λ is a scalar function then

$$\mathcal{L}_{\lambda\mathbf{v}}\mathbf{u} = \lambda\mathcal{L}_\mathbf{v}\mathbf{u} - (\mathbf{u} \circ \lambda)\,\mathbf{v}.$$

To derive this expression, we use the antisymmetry of the commutator and the Leibnitz rule for $\mathcal{L}_\mathbf{v}$:

$$\begin{aligned}\mathcal{L}_{\lambda\mathbf{v}}\mathbf{u} &= [\lambda\mathbf{v}, \mathbf{u}] = -[\mathbf{u}, \lambda\mathbf{v}] = -\mathcal{L}_\mathbf{u}(\lambda\mathbf{v}) \\ &= -\mathcal{L}_\mathbf{u}(\lambda)\,\mathbf{v} - \lambda\mathcal{L}_\mathbf{u}(\mathbf{v}) \\ &= -(\mathbf{u} \circ \lambda)\,\mathbf{v} + \lambda\mathcal{L}_\mathbf{v}\mathbf{u}.\end{aligned}$$

■

1.3.2 Lie derivative of tensors

Let us continue to investigate the properties of the operation $\mathcal{L}_\mathbf{v}$. Consider a 1-form ω and a vector field $\mathbf{u}$. According to Eq. (1.13), we have

$$\mathcal{L}_\mathbf{v}(\omega \circ \mathbf{u}) = \mathcal{L}_\mathbf{v}(\omega) \circ \mathbf{u} + \omega \circ \mathcal{L}_\mathbf{v}(\mathbf{u}).$$

On the other hand, $\omega \circ \mathbf{u}$ is a scalar function, so Eq. (1.10) gives

$$\mathcal{L}_\mathbf{v}(\omega \circ \mathbf{u}) = \mathbf{v} \circ (\omega \circ \mathbf{u}).$$

Hence, the Lie derivative of a 1-form ω with respect to $\mathbf{v}$ is the 1-form $\mathcal{L}_\mathbf{v}\omega$ that acts on an arbitrary vector $\mathbf{u}$ as

$$(\mathcal{L}_\mathbf{v}\omega) \circ \mathbf{u} = \mathbf{v} \circ (\omega \circ \mathbf{u}) - \omega \circ [\mathbf{v}, \mathbf{u}].$$

In the index notation,

$$[\mathcal{L}_\mathbf{v}(\omega)]_\mu = v^\alpha \partial_\alpha \omega_\mu + \omega_\alpha \partial_\mu v^\alpha.$$

The action of the Lie derivative on an arbitrary tensor can be derived similarly, using the properties (1.10)–(1.13). For example, suppose A is a bilinear form with vector values,

i.e. $A(\mathbf{u},\mathbf{v})$ is a vector. (The index notation for A would be $A^{\lambda}_{\alpha\beta\cdot}$.) Then the Lie derivative $\mathcal{L}_{\mathbf{w}}A$ is defined as

$$[\mathcal{L}_{\mathbf{w}}A](\mathbf{u},\mathbf{v}) = [\mathbf{w}, A(\mathbf{u},\mathbf{v})] - A([\mathbf{w},\mathbf{u}],\mathbf{v}) - A(\mathbf{u},[\mathbf{w},\mathbf{v}]).$$

In the index notation,

$$(\mathcal{L}_{\mathbf{v}}A)^{\lambda}_{\alpha\beta} = v^{\mu}\partial_{\mu}A^{\lambda}_{\alpha\beta} + A^{\lambda}_{\alpha\mu}\partial_{\beta}v^{\mu} + A^{\lambda}_{\mu\beta}\partial_{\alpha}v^{\mu} - A^{\mu}_{\alpha\beta}\partial_{\mu}v^{\lambda}.$$

The Lie derivative with respect to a coordinate basis vector of a given coordinate system has special properties with respect to that coordinate system.

Statement 1.3.2: Let $\{x^{\mu}\}$ be local coordinates in a manifold, and consider the standard coordinate bases $\{\partial/\partial x^{\mu}\} \equiv \{\partial_{\mu}\}$ and $\{\mathrm{d}x^{\mu}\}$ in the tangent and the cotangent spaces respectively.

(a) The Lie derivatives with respect to a coordinate basis vector of another basis vector or of a basis 1-form are zero,

$$\mathcal{L}_{\partial_{\mu}}(\partial_{\nu}) = 0, \quad \mathcal{L}_{\partial_{\mu}}(\mathrm{d}x^{\nu}) = 0.$$

(b) Let us represent an arbitrary tensor of rank (m,n) in the coordinate basis by an array of components. For example, a tensor T of rank $(1,2)$ is represented by an array $T^{\alpha}{}_{\beta\gamma}$ as

$$T = \sum_{\alpha,\beta,\gamma} T^{\alpha}{}_{\beta\gamma}\frac{\partial}{\partial x^{\alpha}} \otimes \mathrm{d}x^{\beta} \otimes \mathrm{d}x^{\gamma}.$$

If it is known that $\mathcal{L}_{\partial/\partial x^1}T = 0$, it means that the component functions $T^{\alpha}{}_{\beta\gamma}$ are independent of the coordinate x^1.

Proof: (a) The coordinate basis vectors commute since

$$\mathcal{L}_{\partial_{\mu}}(\partial_{\nu})f = \frac{\partial}{\partial x^{\mu}}\frac{\partial}{\partial x^{\nu}}f - \frac{\partial}{\partial x^{\nu}}\frac{\partial}{\partial x^{\mu}}f = 0$$

for a smooth function f.

The Lie derivative of a 1-form is defined through the action of the 1-form on an arbitrary vector field $\mathbf{v}$,

$$\left(\mathcal{L}_{\partial_{\mu}}\mathrm{d}x^{\nu}\right)\circ\mathbf{v} = \mathcal{L}_{\partial_{\mu}}(\mathrm{d}x^{\nu}\circ\mathbf{v}) - \mathrm{d}x^{\nu}\circ\left(\mathcal{L}_{\partial_{\mu}}\mathbf{v}\right).$$

By definition of the 1-form $\mathrm{d}x^{\nu}$ we have

$$\mathrm{d}x^{\nu}\circ\mathbf{v} \equiv \mathbf{v}\circ x^{\nu}, \quad \mathrm{d}x^{\nu}\circ\left(\mathcal{L}_{\partial_{\mu}}\mathbf{v}\right) \equiv \left(\mathcal{L}_{\partial_{\mu}}\mathbf{v}\right)\circ x^{\nu},$$

where x^{ν} is understood as a scalar function representing the ν-th coordinate of a point. Hence

$$\begin{aligned}\left(\mathcal{L}_{\partial_{\mu}}\mathrm{d}x^{\nu}\right)\circ\mathbf{v} &= \partial_{\mu}(\mathbf{v}\circ x^{\nu}) - [\partial_{\mu},\mathbf{v}]\circ x^{\nu} \\ &= \mathbf{v}\circ(\partial_{\mu}\circ x^{\nu}) = \mathbf{v}\circ\delta^{\nu}_{\mu} = 0.\end{aligned}$$

(b) It is sufficient to show the proof on the given example involving a tensor T of rank $(2,1)$. Using the Leibnitz rule, we compute

$$\begin{aligned}0 = \mathcal{L}_{\partial_1}T &= \sum_{\alpha,\beta,\gamma}\left(\mathcal{L}_{\partial_1}T^{\alpha}{}_{\beta\gamma}\right)\partial_{\alpha}\otimes\mathrm{d}x^{\beta}\otimes\mathrm{d}x^{\gamma} \\ &= \sum_{\alpha,\beta,\gamma}\left(\frac{\partial}{\partial x^1}T^{\alpha}{}_{\beta\gamma}\right)\partial_{\alpha}\otimes\mathrm{d}x^{\beta}\otimes\mathrm{d}x^{\gamma},\end{aligned}$$

since $\mathcal{L}_{\partial_1}\partial_{\alpha} = 0$ and $\mathcal{L}_{\partial_1}\mathrm{d}x^{\beta} = 0$ for every α,β. Since the set of basic tensors of the form $\partial_{\alpha}\otimes\mathrm{d}x^{\beta}\otimes\mathrm{d}x^{\gamma}$ is linearly independent, it follows that every derivative $\partial_1 T^{\alpha}{}_{\beta\gamma}$ vanishes. ■

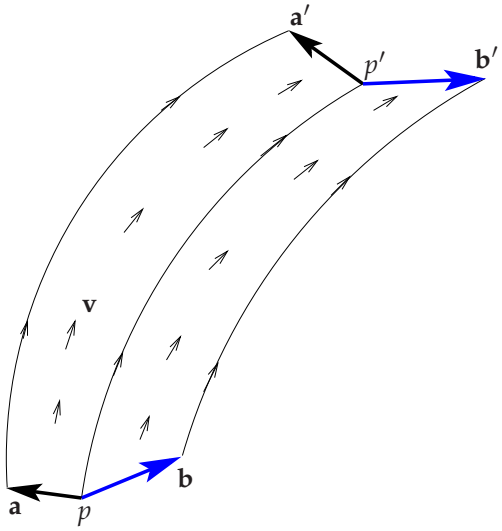

Figure 1.9: A neighborhood of points around p is transported along the orbits of a vector field $\mathbf{v}$ to a neighborhood around p'. Initial tangent vectors $\mathbf{a}$ and $\mathbf{b}$ from $T_p\mathcal{M}$ are visualized as directions between nearby points. These vectors are transported into vectors $\mathbf{a}'$ and $\mathbf{b}'$ from $T_{p'}\mathcal{M}$.

1.3.3 Geometric interpretation

To illustrate the geometric significance of the Lie derivative, we first note that the Lie derivative of a connecting vector field is zero. A connecting vector, such as $\mathbf{c}$ in Fig. 1.8, is transported from the initial point by the flow of the field $\mathbf{v}$. We can imagine that the entire neighborhood of points around the initial point p is transported along the orbits of $\mathbf{v}$ to another nearby point p'; this neighborhood is deformed in a certain way by the transport. The character of this deformation is determined by the vector field $\mathbf{v}$ in the neighborhood of the initial point p and along the way to p'. Tangent vectors at p can be visualized as directions towards points near p. So all tangent vectors $\mathbf{a},\mathbf{b},\ldots \in T_p\mathcal{M}$ are also transported to particular tangent vectors $\mathbf{a}',\mathbf{b}',\ldots \in T_{p'}\mathcal{M}$ (see Fig. 1.9). Thus the flow of $\mathbf{v}$ determines a map between tangent spaces $T_p\mathcal{M}$ and $T_{p'}\mathcal{M}$, as long as the points p and p' lie on a single orbit of $\mathbf{v}$.

We have already seen (Sec. 1.2.11) that a vector field $\mathbf{u}$ obtained by transporting an initial vector $\mathbf{u}|_p$ along the flow of $\mathbf{v}$ satisfies $\mathcal{L}_{\mathbf{v}}\mathbf{u} = 0$. (The transport from an initial point defines the vector field $\mathbf{u}$ along a single orbit of $\mathbf{v}$, so we need to use many orbits and many initial values to define the vector field $\mathbf{u}$ everywhere.) In general, for two vector fields $\mathbf{u}$ and $\mathbf{v}$ we will have $\mathcal{L}_{\mathbf{v}}\mathbf{u} \neq 0$. So we might say (heuristically) that the Lie derivative $\mathcal{L}_{\mathbf{v}}\mathbf{u}$ measures the extent to which a vector field $\mathbf{u}$ fails to be a connecting field for $\mathbf{v}$.

To make these considerations somewhat more precise, let us temporarily denote by $\varphi_{\tau} : T_p\mathcal{M} \to T_{p'}\mathcal{M}$ the map that transports tangent vectors along the orbits of $\mathbf{v}$ from point p to point p', where τ is the parameter distance between p and p'. (In other words, we introduce a curve γ parameterized by τ such that $\gamma(0) = p$ and $\gamma(\tau) = p'$, while $\dot{\gamma} = \mathbf{v}$ everywhere along γ.) In this notation, the vectors in Fig. 1.9 satisfy

$$\mathbf{a}' = \varphi_{\tau}(\mathbf{a}), \quad \mathbf{b}' = \varphi_{\tau}(\mathbf{b}).$$

Now we would like to compute the Lie derivative $\mathcal{L}_{\mathbf{v}}\mathbf{u}$, where $\mathbf{u}$ is some vector field. Let us compute $\mathcal{L}_{\mathbf{v}}\mathbf{u}$ at point p; let us assume that τ is very small and so $\mathbf{v}$ is a vector pointing from

p to p'. We might try to write the derivative of $\mathbf{u}$ along $\mathbf{v}$ at point p as

$$\lim_{\tau\to 0}\frac{\mathbf{u}|_{p'}-\mathbf{u}|_p}{\tau}, \quad ???$$

but this expression is meaningless since we cannot subtract vectors $\mathbf{u}$ at different points: these vectors belong to different tangent spaces. However, we can transport $\mathbf{u}|_{p'} \in T_{p'}\mathcal{M}$ back to the tangent space $T_p\mathcal{M}$ using the inverse map φ_τ^{-1}. So we write the derivative of $\mathbf{u}$ along $\mathbf{v}$ at point p as

$$\mathcal{L}_{\mathbf{v}}\mathbf{u} = \lim_{\tau\to 0}\frac{\varphi_\tau^{-1}\left(\mathbf{u}|_{p'}\right)-\mathbf{u}|_p}{\tau}. \tag{1.15}$$

The geometric interpretation of this expression is clear: $\mathcal{L}_{\mathbf{v}}\mathbf{u}$ is the extent to which the field $\mathbf{u}$ fails to be transported along $\mathbf{v}$.[4]

Similar considerations apply to the Lie derivative of a tensor A with respect to a vector field $\mathbf{v}$. Any tensor A can be defined as a multilinear function of tangent vectors and covectors, while vectors and covectors are defined through the points in an infinitesimal neighborhood of p. Thus, any tensor A is ultimately an object defined through combinations of points near p. When these points are transported by the flow of a vector field $\mathbf{v}$, the corresponding transport $\varphi_\tau(A)$ of the tensor A is also well-defined. Then the Lie derivative $\mathcal{L}_{\mathbf{v}}A$ is defined by a formula analogous to Eq. (1.15).

It follows from expressions derived above that the Lie derivative $\mathcal{L}_{\mathbf{v}}A|_p$ of a tensor A at a point p contains not only derivatives of A but also derivatives of $\mathbf{v}$. So the tensor $\mathcal{L}_{\mathbf{v}}A|_p$ depends not only on the value $\mathbf{v}|_p$ of the vector field $\mathbf{v}$ at the point p but also on the values of $\mathbf{v}$ in an entire neighborhood of p. Thus the Lie derivative $\mathcal{L}_{\mathbf{v}}A|_p$ is not a true directional derivative of A in the direction of $\mathbf{v}$ at the point p. A true directional derivative along a curve $\gamma(\tau)$, i.e. a derivative that depends only on the *value* of the vector $\dot{\gamma} \equiv \mathbf{v}(p)$ at the point p, must be independent of the *derivatives* of $\mathbf{v}$ at p. Such a true directional derivative cannot be defined without introducing some additional structures on the manifold $\mathcal{M}$. This is so because the tensors $A(p)$ and $A(p')$ at neighboring points belong to different tensor spaces, and thus there is no definition for a quantity such as $A(p')-A(p)$. Similarly, there is no possibility to integrate vectors or tensors over a domain. We might say, in figurative language, that the tangent spaces "are disconnected" and a directional derivative cannot be computed without a "connection" between these spaces. (The structure called the "connection" will be introduced in Sec. 1.6 below.) When we evaluate the Lie derivative $\mathcal{L}_{\mathbf{v}}$, the information about how to connect the tangent spaces comes from the flow of the vector field $\mathbf{v}$. This allows us to relate directions around p to directions around p', i.e. to "connect" the tangent spaces $T_p\mathcal{M}$ and $T_{p'}\mathcal{M}$. However, this procedure uses information about the vector field $\mathbf{v}$ in the entire neighborhood of p, not just at one point p.

1.4 Calculus of differential forms

The calculus of differential forms is a basic tool of differential geometry. Differential forms are not widely used in standard texts on GR because they do not provide a decisive computational advantage as long as one remains within the standard introductory material. A substantial computational advantage of differential forms is first seen when considering more advanced material, such as the Frobenius theorem or the tetrad formulation of GR. However, the calculus of forms is relatively simple, and in my view the learning effort involved is amply justified by the conceptual advantages and wide-ranging applicability of differential forms in theoretical physics and mathematics. In the following subsections I explain the motivation for introducing differential forms. Then (Sec. 1.4.3 and below) I give an overview of the basic properties of differential forms, mostly without proof. Since the calculus of forms is standard material and since it will be used in this book only as a computational tool, this brief introduction will suffice.[5]

1.4.1 Volume as antisymmetric tensor

The familiar notion of volume can be given a natural interpretation in terms of antisymmetric multilinear functions of vectors, i.e. antisymmetric tensors. This interpretation is a great help in calculations because antisymmetric tensors have rich properties.

We begin with the two-dimensional Euclidean plane $\mathbb{R}^2$. In two dimensions, the *area* plays the role of volume, so let us consider the area of a parallelogram spanned by two vectors $\mathbf{u}, \mathbf{v}$. According to standard formula of elementary geometry, the area of the parallelogram is

$$A = |\mathbf{u}|\,|\mathbf{v}|\sin\alpha,$$

where α is the angle between the vectors, which is a number between 0 and π determined by

$$\cos\alpha \equiv \frac{\mathbf{u}\cdot\mathbf{v}}{|\mathbf{u}|\,|\mathbf{v}|}.$$

Here $|\mathbf{u}| \equiv \sqrt{\mathbf{u}\cdot\mathbf{u}}$ and $\mathbf{u}\cdot\mathbf{v}$ is the standard Euclidean scalar product. Thus the area of a parallelogram spanned by $\mathbf{u}$ and $\mathbf{v}$ is a nonnegative number defined by

$$A(\mathbf{u},\mathbf{v}) \equiv \sqrt{|\mathbf{u}|^2\,|\mathbf{v}|^2-(\mathbf{u}\cdot\mathbf{v})^2}.$$

This number appears to be a complicated nonlinear function of the vectors $\mathbf{u}$ and $\mathbf{v}$. It is more convenient to work with the **oriented area** $\tilde{A}$, which is defined as $\tilde{A} = +A$ if the pair $\{\mathbf{u},\mathbf{v}\}$ has positive orientation and $\tilde{A} = -A$ otherwise. (We can fix the "positive orientation" to be that of the coordinate axes $\{x, y\}$ in the plane, taken in this order.) The decisive advantage of the oriented area $\tilde{A}$ over the conventional (unsigned) area A is that $\tilde{A}(\mathbf{u},\mathbf{v})$ turns out to be an *antisymmetric bilinear function* of $\mathbf{u}$ and $\mathbf{v}$. In mathematics, a number-valued bilinear function of two vectors is usually called a **bilinear form**, while an *antisymmetric* bilinear form is called a **2-form**.

The oriented area $\tilde{A}(\mathbf{u},\mathbf{v})$ is by definition antisymmetric in $(\mathbf{u},\mathbf{v})$, since the pair $\{\mathbf{v},\mathbf{u}\}$ has opposite orientation to $\{\mathbf{u},\mathbf{v}\}$. The fact that $\tilde{A}(\mathbf{u},\mathbf{v})$ is an bilinear form can be demonstrated quite straightforwardly. When one of the vectors $\mathbf{u}$ or $\mathbf{v}$ is multiplied by a positive number, the area of a parallelogram is multiplied by the same number. If one of the vectors is multiplied by a negative number, the orientation of the pair

[4] Although in this section I do not actually *prove* that Eq. (1.15) is equivalent to the previous definition of $\mathcal{L}_{\mathbf{v}}\mathbf{u}$, a proof can be straightforwardly filled in or found in differential geometry textbooks, albeit perhaps in a more formal or abstract language.

[5] The calculus of differential forms is explained, for example, in chapter 4 of [33] or in chapter 7 of [1].

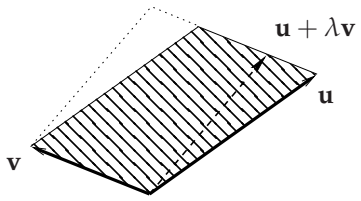

Figure 1.10: The area of the parallelogram spanned by a pair of vectors $\{\mathbf{u}, \mathbf{v}\}$ is equal to the area spanned by $\{\mathbf{u} + \lambda\mathbf{v}, \mathbf{v}\}$.

$\{\mathbf{u}, \mathbf{v}\}$ is reversed and thus $\tilde{A}(\mathbf{u}, \mathbf{v})$ changes sign. We also have $\tilde{A}(0, \mathbf{v}) = 0$. So it is easy to see that $\tilde{A}(\lambda\mathbf{u}, \mathbf{v}) = \lambda\tilde{A}(\mathbf{u}, \mathbf{v})$, where λ is an arbitrary real number (positive, zero, or negative). Since $\tilde{A}(\mathbf{u}, \mathbf{v}) = -\tilde{A}(\mathbf{v}, \mathbf{u})$, the same property holds for the argument $\mathbf{v}$. Further, one can show by an elementary geometric construction (see Fig. 1.10) that the area does not change when a multiple of $\mathbf{v}$ is added to $\mathbf{u}$, namely

$$\tilde{A}(\mathbf{u} + \lambda\mathbf{v}, \mathbf{v}) = \tilde{A}(\mathbf{u}, \mathbf{v}).$$

In two dimensions, two vectors $\{\mathbf{u}, \mathbf{v}\}$ are a basis if $\tilde{A}(\mathbf{u}, \mathbf{v}) \neq 0$. It follows that $\tilde{A}$ is linear in each argument: we can compute $\tilde{A}(\mathbf{u}_1 + \mathbf{u}_2, \mathbf{v})$ by decomposing $\mathbf{u}_1 = \lambda_1\mathbf{u} + \mu_1\mathbf{v}$, $\mathbf{u}_2 = \lambda_2\mathbf{u} + \mu_2\mathbf{v}$, and then it is easy to see that

$$\tilde{A}(\mathbf{u}_1 + \mathbf{u}_2, \mathbf{v}) = \tilde{A}(\mathbf{u}_1, \mathbf{v}) + \tilde{A}(\mathbf{u}_2, \mathbf{v}).$$

The same property holds for the argument $\mathbf{v}$. Thus, $\tilde{A}(\mathbf{u}, \mathbf{v})$ is an antisymmetric bilinear form that we may call the **area 2-form**.

To make the above discussion less abstract, let us express all quantities in Cartesian coordinates $\{x, y\}$. The vectors $\mathbf{u}$ and $\mathbf{v}$ have components $\{u_1, u_2\}$ and $\{v_1, v_2\}$; the unoriented area is

$$A(\mathbf{u}, \mathbf{v}) = \sqrt{(u_1^2 + u_2^2)\,(v_1^2 + v_2^2) - (u_1v_1 + u_2v_2)^2}.$$

After some algebraic simplification, this becomes

$$A(\mathbf{u}, \mathbf{v}) = \sqrt{(u_1v_2 - u_2v_1)^2} = |u_1v_2 - u_2v_1|. \qquad (1.16)$$

Now it is clear that $A(\mathbf{u}, \mathbf{v})$ is the absolute value of a 2-form $\tilde{A}$ defined by

$$\tilde{A}(\mathbf{u}, \mathbf{v}) = u_1v_2 - u_2v_1 = \begin{vmatrix} u_1 & v_1 \\ u_2 & v_2 \end{vmatrix}.$$

The sign of $\tilde{A}$ is chosen such that the area of the unit square spanned by the unit basis vectors $\{1, 0\}$ and $\{0, 1\}$ (in this order) is equal to 1. We note that the oriented area $\tilde{A}$ is more closely related to linear algebra than the unoriented area A.

A similar argument (with a suitable generalization of Fig. 1.10 to three dimensions) shows that the oriented 3-volume $\tilde{V}(\mathbf{a}, \mathbf{b}, \mathbf{c})$ of a parallelepiped spanned by three vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$ in a three-dimensional Euclidean space is a totally antisymmetric trilinear form (3-form). The same conclusion holds in any dimensions. The n-dimensional oriented volume of a parallelepiped spanned by vectors $\mathbf{a}_1$, ..., $\mathbf{a}_n$ in Euclidean space is given by

$$\tilde{V}(\mathbf{a}_1, ..., \mathbf{a}_n) = \begin{vmatrix} a_1^1 & \cdots & a_n^1 \\ \vdots & \ddots & \vdots \\ a_1^n & \cdots & a_n^n \end{vmatrix},$$

where a_j^i is the i-th component of the vector $\mathbf{a}_j$.

Remark: determinants. The last formula suggests a close connection between antisymmetric forms and the calculus of determinants. Indeed, such a close connection exists. For instance, one can show that the oriented volume of an n-dimensional parallelepiped transformed by a linear map T will change by a numerical factor, which depends on T but does not depend on the parallelepiped. This numerical factor is equal to the determinant of T. In Sec. 1.4.5, we use this property as a convenient *definition* of the determinant of a linear transformation T. ■

A somewhat more complicated concept is the area of a parallelogram spanned by two vectors $\{\mathbf{u}, \mathbf{v}\}$ in a higher-dimensional Euclidean space, for example, in $\mathbb{R}^3$. The ordinary (unoriented) area $A(\mathbf{u}, \mathbf{v})$ is a nonlinear function of $\mathbf{u}, \mathbf{v}$ given by Eq. (1.16). In three (or more) dimensions, one cannot define an "oriented" area $\tilde{A}(\mathbf{u}, \mathbf{v})$ linear in $\mathbf{u}, \mathbf{v}$. Instead, it turns out that one can define a *vector-valued* antisymmetric, bilinear function $\mathbf{A}(\mathbf{u}, \mathbf{v})$ whose absolute value is equal to the area $A(\mathbf{u}, \mathbf{v})$. This function is the familiar vector product of the vectors $\mathbf{u}$ and $\mathbf{v}$,

$$\mathbf{A}(\mathbf{u}, \mathbf{v}) = \mathbf{u} \times \mathbf{v}.$$

Indeed, it is well known that

$$|\mathbf{u} \times \mathbf{v}| = |\mathbf{u}|\,|\mathbf{v}| \sin\alpha = A(\mathbf{u}, \mathbf{v}).$$

Thus, one could say that in three dimensions the area has an orientation *vector*. Thus, it is natural to regard the fundamental area as vector-valued; the ordinary area is then computed as the absolute value of the "area vector."

Remark: The same idea can be generalized to n-dimensional parallelepipeds embedded in N-dimensional space (where $N > n$). This time, it turns out that the "orientation" of the n-dimensional volume is not a vector but an antisymmetric tensor; so the "oriented" n-dimensional volume is naturally tensor-valued. The ordinary volume is a suitable "absolute value" of that tensor. More explanations can be found in Sec. 1.4.4. One may even change the point of view and regard the *oriented* volume as a basic quantity — which is a certain totally antisymmetric tensor — and the ordinary volume as a quantity *defined* as the absolute value of the oriented volume. ■

1.4.2 Motivation for differential forms

In standard courses of multivariate calculus, one studies integrals such as

$$\iiint_{-\infty}^{+\infty} e^{-x^2-y^2-z^2}\,dxdydz = \pi^{3/2}.$$

Expressions such as "$dxdydz$" occurring under the sign of multiple integration are usually explained as a purely symbolic notation. It is perhaps not obvious to a student of calculus that "$dxdydz$" above can be interpreted, quite rigorously, as an *antisymmetric* product of "dx," "dy," and "dz," rather than an ordinary, symmetric product. In fact, the antisymmetric interpretation of "$dxdydz$" is necessary if one wishes to obtain a consistent calculus where objects such as "dx" or "$dxdy$" are well-defined and transform correctly under a change of variables. Differential forms can be understood as the rigorous mathematical objects that make precise sense out of expressions such as "$dxdydz$." In this section we explore these motivational ideas.

Let us begin by considering an example familiar from mechanics. Suppose a particle moves in three-dimensional space along a curve $\mathbf{x}(\tau) \equiv \{X(\tau), Y(\tau), Z(\tau)\}$ and experiences a space-dependent external force $\mathbf{F} \equiv \{F_1, F_2, F_3\}$. The work done by the force can be computed as the **line integral**,

$$W = \int_{\mathbf{x}(\tau)} (F_1 dx + F_2 dy + F_3 dz) \,. \tag{1.17}$$

This expression is computed as the ordinary integral

$$W = \int_{\tau_1}^{\tau_2} \left(F_1 \dot{X} + F_2 \dot{Y} + F_3 \dot{Z}\right) d\tau.$$

However, this form of the expression requires a parameterization $\mathbf{x}(\tau)$, while the work W is actually independent of the choice of the parameter τ along the curve: it depends only on the force $\mathbf{F}$ and on the layout of the given curve in space. So it is more natural to work with the line integral (1.17). We can interpret the integrand in Eq. (1.17) directly as a 1-form

$$\omega \equiv F_1 \mathrm{d}x + F_2 \mathrm{d}y + F_3 \mathrm{d}z.$$

Note that the tangent vector $\dot{\mathbf{x}} d\tau$ is an approximate representation of a short curve segment (see Sec. 1.2.5) between the points $\mathbf{x}(\tau)$ and $\mathbf{x}(\tau + d\tau)$. The expression

$$\left(F_1 \dot{X} + F_2 \dot{Y} + F_3 \dot{Z}\right) d\tau = \omega \circ \dot{\mathbf{x}} d\tau$$

is interpreted as the 1-form ω evaluated on the tangent vector $\dot{\mathbf{x}} d\tau$. This expression represents a small amount of work done along a short curve segment. So the line integral (1.17) is rewritten as

$$W = \int_{\tau_1}^{\tau_2} (\omega \circ \dot{\mathbf{x}}) \, d\tau \equiv \int_\gamma \omega,$$

where γ denotes the path along which the particle is moving in space (regardless of the choice of the parameter τ).

Remark: One does not write any "d"s after the integral sign in the last expression because all the "d"s are hidden in the 1-form ω. I write the roman "d" in 1-forms to emphasize the fact that "dx" is a separate, rigorously defined object, while "dx" stands for a heuristic "infinitesimal change of x." ■

In general, one defines the integral of a 1-form ω over a curve γ as follows. One splits the curve into a large number N of short segments. Each segment is almost straight and thus can be approximately represented by a tangent vector (see Sec. 1.2.5). Thus we represent the N segments by N tangent vectors $\mathbf{v}_1$, ..., $\mathbf{v}_N$. Each segment contributes to the integral a small amount computed as $\omega(\mathbf{v}_j)$. The integral $\int_\gamma \omega$ is then defined as

$$\int_\gamma \omega = \lim_{N\to\infty} \sum_{j=1}^{N} \omega(\mathbf{v}_j).$$

The line integral $\int_\gamma \omega$ can be reduced to an ordinary integral by choosing a parameterization $\gamma(\tau)$ for the curve γ and writing

$$\int_\gamma \omega = \int_{\tau_1}^{\tau_2} \omega(\dot{\gamma}) d\tau. \tag{1.18}$$

It is easy to see that the value of $\int_\gamma \omega$ actually does not depend on the choice of the parameter τ along the curve γ. (A redefinition $\tilde{\tau} = f(\tau)$ will introduce a factor of f' into $\dot{\gamma}$, while a compensating factor $1/f'$ will come from $d\tilde{\tau}$.) The reason is that the line integral $\int_\gamma \omega$ is defined purely geometrically through splitting of the curve γ into small segments. Thus $\int_\gamma \omega$ is a geometrically defined quantity that is independent of the coordinate representation of γ and of the choice of the parameter τ.

The calculus of 1-forms is designed to be compatible with the change of variables under the integral. For instance, consider an ordinary integral $\int f(x)dx$. After a change of variable $x = X(\phi)$, the integral becomes

$$\int f(x)dx = \int f(X(\phi)) \frac{dX}{d\phi} d\phi.$$

This corresponds precisely to the chain rule for the differential operator d,

$$f(x)\mathrm{d}x = f(x)\mathrm{d}X(\phi) = f(x)X'(\phi)\mathrm{d}\phi.$$

So it is consistent to interpret the expression $f(x)dx$ within an ordinary integral as a 1-form $f(x)\mathrm{d}x$.

Let us now consider two-dimensional integrals (**surface integrals**). An integral over a surface with local coordinates $\{x, y\}$ is computed as a repeated ordinary integral,

$$\iint F(x,y)dxdy = \int_{x_1}^{x_2} dx \int_{y_1}^{y_2} dy \, F(x,y). \tag{1.19}$$

Now we would like to interpret the expression $F(x,y)dxdy$ as an analog and a generalization of the 1-form $f(x)\mathrm{d}x$ seen in the previous example.

As in the previous example, we can divide the surface of integration into a large number N of small rectangles with sides δx and δy. The sides of each rectangle are approximately represented by a pair of tangent vectors $\{\mathbf{u}_j, \mathbf{v}_j\}$, $j = 1, ..., N$. A rectangle number j located at the point $\{x_j, y_j\}$ gives a small contribution $F(x_j, y_j)\delta x \delta y$ to the integral. It is clear that the number $F(x_j, y_j)\delta x \delta y$ can be understood as the evaluation of a bilinear form ω acting on tangent vectors $\{\mathbf{u}_j, \mathbf{v}_j\}$, i.e.

$$F(x_j, y_j)\delta x \delta y \equiv \omega \left(\mathbf{u}_j, \mathbf{v}_j\right).$$

The bilinear form ω is called a **2-form** since it has *two* vector arguments. Then the surface integral is defined as the limit

$$\iint F dxdy = \lim_{N\to\infty} \sum_{j=1}^{N} \omega \left(\mathbf{u}_j, \mathbf{v}_j\right).$$

The 2-form ω can be temporarily written as $\omega = F\mathrm{d}x\,\mathrm{d}y$. Let us now explore its properties under the change of variables in the integral.

We first perform a simple change of variables, $\{x, y\} \to \{x, \phi\}$, where we replace y by $y = Y(x, \phi)$ and Y is a fixed function. To transform the integral, it is convenient to use the formula (1.19). Since the variable x is held fixed within the integration over y, we find

$$\int_{x_1}^{x_2} dx \int_{y_1}^{y_2} dy \, F(x,y) = \int_{x_1}^{x_2} dx \int_{\Phi_1(x)}^{\Phi_2(x)} d\phi \, F(x,Y) \frac{\partial Y(x,\phi)}{\partial \phi}.$$

Hence, if we wish to interpret the expression $F\mathrm{d}x\,\mathrm{d}y$ consistently as a bilinear form, we must adopt the rule

$$\mathrm{d}x\,\mathrm{d}Y(x,\phi) = \frac{\partial Y(x,\phi)}{\partial \phi} \mathrm{d}x\,\mathrm{d}\phi.$$

However, according to the calculus of 1-forms, we have

$$\mathrm{d}Y(x,\phi) = \frac{\partial Y}{\partial x}\mathrm{d}x + \frac{\partial Y}{\partial \phi}\mathrm{d}\phi$$

and (assuming that the product $\mathrm{d}x\,\mathrm{d}y$ is bilinear)

$$\mathrm{d}x\,\mathrm{d}Y(x,\phi) = \mathrm{d}x\left(\frac{\partial Y}{\partial x}\mathrm{d}x + \frac{\partial Y}{\partial \phi}\mathrm{d}\phi\right).$$

Therefore, this calculus is consistent with the properties of two-dimensional integration only if we assume that

$$\mathrm{d}x\,\mathrm{d}x = 0.$$

Now it is clear that the 2-form "$\mathrm{d}x\,\mathrm{d}y$" must be a rather special kind of product of 1-forms $\mathrm{d}x$ and $\mathrm{d}y$. This product is called the **exterior product** (also **wedge product**) and is denoted by the symbol $\wedge$; so one writes $\mathrm{d}x \wedge \mathrm{d}y$ rather than "$\mathrm{d}x\,\mathrm{d}y$." The property

$$\mathrm{d}x \wedge \mathrm{d}x = 0$$

together with bilinearity means that the exterior product is *antisymmetric*. To see this, consider

$$\mathrm{d}(x+y) \wedge \mathrm{d}(x+y) = 0 \;\Rightarrow\; \mathrm{d}x \wedge \mathrm{d}y + \mathrm{d}y \wedge \mathrm{d}x = 0.$$

Let us now check the consistency of a more general change of variables, $\{x,y\} \to \{\phi,\psi\}$ where $x = X(\phi,\psi)$ and $y = Y(\phi,\psi)$. Calculations are straightforward because the only new rule is the antisymmetry of the exterior product. We find

$$\begin{aligned}\mathrm{d}x \wedge \mathrm{d}y &= \mathrm{d}X(\phi,\psi) \wedge \mathrm{d}Y(\phi,\psi)\\ &= \left(\frac{\partial X}{\partial \phi}\mathrm{d}\phi + \frac{\partial X}{\partial \psi}\mathrm{d}\psi\right) \wedge \left(\frac{\partial Y}{\partial \phi}\mathrm{d}\phi + \frac{\partial Y}{\partial \psi}\mathrm{d}\psi\right)\\ &= \left(\frac{\partial X}{\partial \phi}\frac{\partial Y}{\partial \psi} - \frac{\partial X}{\partial \psi}\frac{\partial Y}{\partial \phi}\right)\mathrm{d}\phi \wedge \mathrm{d}\psi, \qquad (1.20)\end{aligned}$$

where we used the antisymmetry properties

$$\mathrm{d}\phi \wedge \mathrm{d}\phi = \mathrm{d}\psi \wedge \mathrm{d}\psi = 0, \quad \mathrm{d}\phi \wedge \mathrm{d}\psi + \mathrm{d}\psi \wedge \mathrm{d}\phi = 0,$$

to get the last line. Now we may recognize Eq. (1.20) as the standard formula for the change of variables, involving the two-dimensional Jacobian of the transformation $\{x,y\} \to \{\phi,\psi\}$. In this way, we can see that the unusual rule of the exterior product is a natural consequence of the known properties of multiple integration.[6]

The geometric formulation of the surface integral is therefore the integral of a 2-form over a surface, written as

$$\int_{\mathcal{A}} \omega,$$

where $\mathcal{A}$ is a surface (or part of a surface) and ω is a 2-form.

The same formulation applies to integrals over surfaces embedded in a higher-dimensional space. Consider another example familiar from physics: the flux of a magnetic field $\mathbf{B}$ through a surface $\mathcal{A}$ is computed as a surface integral

$$\Phi = \int_{\mathcal{A}} \mathbf{B} \cdot d\mathbf{S},$$

where $d\mathbf{S}$ is understood as a vector-valued "infinitesimal surface element." The flux integral is usually written in components, $\mathbf{B} \equiv \{B_x, B_y, B_z\}$, as follows:

$$\int_{\mathcal{A}} \left(B_x dydz + B_y dxdz + B_z dxdy\right).$$

However, this notation needs to be supplemented by a complicated prescription for the orientation of the surface elements. Integrals of this kind become more transparent in the notation of 2-forms. One introduces the 2-form

$$\boldsymbol{\beta} \equiv B_x \mathrm{d}y \wedge \mathrm{d}z + B_y \mathrm{d}z \wedge \mathrm{d}x + B_z \mathrm{d}x \wedge \mathrm{d}y$$

and expresses the flux Φ as

$$\Phi = \int_{\mathcal{A}} \boldsymbol{\beta}.$$

The (somewhat complicated) rules for changing variables in such integrals, as well as the orientation prescriptions, are automatically reproduced by the calculus of differential forms. The rules of this calculus are simple: one just needs to use the chain rule for d and the antisymmetry of $\wedge$.

Similarly, one can treat a three-dimensional integral,

$$\iiint F(x,y,z)dxdydz,$$

as an integral of the 3-form $F(x,y,z)\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z$ over a three-dimensional manifold (or part of a manifold). The same rules will provide the correct formulas under any changes of variables in the integral. Collectively, n-forms ($n = 1,2,3,...$) are called **differential forms** since they represent expressions involving the differentials d.

After this motivation, we proceed to summarize the main definitions and properties of differential forms. We begin by studying totally antisymmetric tensors.

1.4.3 Antisymmetric tensors

An **n-form** ω is a totally antisymmetric map from sets of n vector fields to numbers, $\omega(\mathbf{v}_1,...,\mathbf{v}_n) \in \mathbb{R}$, which is linear in each of its vector arguments $\mathbf{v}_j$. In other words, an n-form is a totally antisymmetric tensor of rank $(0,n)$. The equivalent notations $\omega(\mathbf{v}_1,...\mathbf{v}_n) \equiv \omega \circ (\mathbf{v}_1,...,\mathbf{v}_n)$ will be used for convenience or clarity.

When considering vector fields on manifolds, one deals with vectors $\mathbf{v}|_p$ defined in each tangent space $T_p\mathcal{M}$. So it is natural to consider n-forms $\omega(\mathbf{v}_1,...,\mathbf{v}_n;p)$ defined in each tangent space $T_p\mathcal{M}$, i.e. "n-form fields." These "n-form fields" are called **differential forms** or simply n-forms, just as we sometimes call vector fields simply vectors.

In a local coordinate system $\{x^\alpha\}$, tangent vectors are represented by components v^α and n-forms are represented by multi-indexed arrays of components, $\omega_{\alpha\beta...\gamma}$, such that

$$\omega(\mathbf{v}_1,...,\mathbf{v}_n) = \omega_{\alpha\beta...\gamma} v_1^\alpha v_2^\beta ... v_n^\gamma.$$

The array of components $\omega_{\alpha\beta...\gamma}$ is totally antisymmetric in all indices. We use the index-free notation in which n-forms are represented explicitly through basis 1-forms such as $\mathrm{d}x$. To express the antisymmetry of an n-form, one writes $\mathrm{d}x \wedge \mathrm{d}y$, where $\wedge$ denotes the special antisymmetric product to be introduced now.

[6] However, note that we obtained the Jacobian itself rather than its absolute value, which one has in formulas involving the area of domain. This is consistent because integrals of 2-forms are defined on *oriented* surfaces and are not merely integrals of the ordinary, unsigned area. Reversing the orientation of the surface will reverse the sign of the integral.

The **exterior product** (also called the **wedge product**) of two forms, say an m-form ω_1 and an n-form ω_2, is an $(m+n)$-form $\omega_1 \wedge \omega_2$, defined by a total antisymmetrization of products of ω_1 and ω_2,

$$(\omega_1 \wedge \omega_2) \circ (\mathbf{v}_1, ..., \mathbf{v}_{m+n}) \equiv \frac{1}{m!n!} \sum_{\sigma} (-1)^{|\sigma|} \times \omega_1 \left(\mathbf{v}_{\sigma(1)}, ..., \mathbf{v}_{\sigma(m)} \right) \omega_2 \left(\mathbf{v}_{\sigma(m+1)}, ..., \mathbf{v}_{\sigma(m+n)} \right), \quad (1.21)$$

where the sum goes over all the permutations σ of the set $\{1, 2, ..., m+n\}$, and $|\sigma| = 0, 1$ is a function showing whether the permutation σ is even or odd. Note that the factors $m!n!$ in Eq. (1.21) will cancel due to the total antisymmetry of ω_1 and ω_2. For instance, if θ is a 1-form and ω is a 2-form then

$$(\theta \wedge \omega) \circ (\mathbf{x}, \mathbf{y}, \mathbf{z}) = \theta(\mathbf{x})\omega(\mathbf{y}, \mathbf{z}) + \theta(\mathbf{y})\omega(\mathbf{z}, \mathbf{x}) + \theta(\mathbf{z})\omega(\mathbf{x}, \mathbf{y}).$$

Similarly, if η, θ, and ω are 1-forms then

$$(\eta \wedge \theta) \circ (\mathbf{x}, \mathbf{y}) = \eta(\mathbf{x})\theta(\mathbf{y}) - \eta(\mathbf{y})\theta(\mathbf{x}) = \det \begin{vmatrix} \eta(\mathbf{x}) & \eta(\mathbf{y}) \\ \theta(\mathbf{x}) & \theta(\mathbf{y}) \end{vmatrix},$$

$$(\eta \wedge \theta \wedge \omega) \circ (\mathbf{x}, \mathbf{y}, \mathbf{z}) = \det \begin{vmatrix} \eta(\mathbf{x}) & \eta(\mathbf{y}) & \eta(\mathbf{z}) \\ \theta(\mathbf{x}) & \theta(\mathbf{y}) & \theta(\mathbf{z}) \\ \omega(\mathbf{x}) & \omega(\mathbf{y}) & \omega(\mathbf{z}) \end{vmatrix}.$$

It follows that the components of $\eta \wedge \theta$ in a local coordinate system $\{x^\alpha\}$ are

$$(\eta \wedge \theta)_{\alpha\beta} = \eta_\alpha \theta_\beta - \theta_\alpha \eta_\beta,$$

where η_α and θ_α are the components of the 1-forms η and θ.

Example: Consider a local coordinate system $\{x, y, z\}$ and the 1-forms $\eta = y\mathrm{d}x - 2\mathrm{d}z$, $\theta = x\mathrm{d}y + y\mathrm{d}z$. The exterior product of η and θ is computed as follows,

$$\begin{aligned} \eta \wedge \theta &= (y\mathrm{d}x - 2\mathrm{d}z) \wedge (x\mathrm{d}y + y\mathrm{d}z) \\ &= xy\mathrm{d}x \wedge \mathrm{d}y + y^2 \mathrm{d}x \wedge \mathrm{d}z + 2x\mathrm{d}y \wedge \mathrm{d}z \end{aligned}$$

(we used $\mathrm{d}y \wedge \mathrm{d}z = -\mathrm{d}z \wedge \mathrm{d}y$). Let us now compute how the 2-form $\eta \wedge \theta$ acts on a pair of vectors $\mathbf{u} = 5\partial_x - y\partial_z$, $\mathbf{v} = \partial_x + 3z\partial_y$. We have $\mathrm{d}y \circ \mathbf{u} = 0$ and $\mathrm{d}z \circ \mathbf{v} = 0$, which simplifies the calculations:

$$\begin{aligned} (\mathrm{d}x \wedge \mathrm{d}y) \circ (\mathbf{u}, \mathbf{v}) &= (\mathrm{d}x \circ \mathbf{u})\,(\mathrm{d}y \circ \mathbf{v}) = 15z, \\ (\mathrm{d}x \wedge \mathrm{d}z) \circ (\mathbf{u}, \mathbf{v}) &= -(\mathrm{d}z \circ \mathbf{u})\,(\mathrm{d}x \circ \mathbf{v}) = y, \\ (\mathrm{d}y \wedge \mathrm{d}z) \circ (\mathbf{u}, \mathbf{v}) &= -(\mathrm{d}z \circ \mathbf{u})\,(\mathrm{d}y \circ \mathbf{v}) = 3yz. \end{aligned}$$

Finally, we compute

$$(\eta \wedge \theta) \circ (\mathbf{u}, \mathbf{v}) = xy \cdot 15z + y^2 \cdot y + 2x \cdot 3yz = 21xyz + y^3.$$

■

A useful property is that the exterior product of linearly dependent 1-forms vanishes.

Statement 1.4.3.1: If ω_1, ..., ω_k are some 1-forms, then $\omega_1 \wedge ... \wedge \omega_k \neq 0$ iff the set of the 1-forms $\{\omega_j\}$ is linearly independent.

Proof of Statement 1.4.3.1: If the set $\{\omega_j\}$ is linearly dependent then we can express e.g. ω_1 through other ω_j, and the exterior product vanishes due to antisymmetry ($\omega_j \wedge \omega_j = 0$). If the set $\{\omega_j\}$ is linearly independent then it is either already a basis or can be completed to a basis $\{\omega_1, ..., \omega_k, ..., \omega_n\}$ in the n-dimensional space of 1-forms. So the exterior product is (by definition) an alternating linear combination of tensor products of basis 1-forms,

$$\omega_1 \wedge ... \wedge \omega_k = \sum_{\sigma} (-1)^{|\sigma|} \omega_{\sigma(1)} \otimes ... \otimes \omega_{\sigma(k)},$$

where the sum is over all permutations σ of the set $\{1, ..., k\}$. Since the set of all basis tensors $\{\omega_{j_1} \otimes ... \otimes \omega_{j_k}\}$ is by definition linearly independent, the linear combination cannot vanish. ■

1.4.4 *Oriented volume and n-vectors

In Sec. 1.4.1, we have introduced the notion of oriented area and oriented volume. Now we develop a more general picture in which certain antisymmetric tensors represent n-dimensional volumes embedded in N-dimensional space, where dimensions are arbitrary and $N \geq n$.

By analogy with the exterior product of forms, one can define the exterior product of vectors. For instance, the exterior product of two vectors is the antisymmetric tensor defined by

$$\mathbf{x} \wedge \mathbf{y} \equiv \mathbf{x} \otimes \mathbf{y} - \mathbf{y} \otimes \mathbf{x}; \quad (\mathbf{x} \wedge \mathbf{y})^{\alpha\beta} = x^\alpha y^\beta - x^\beta y^\alpha.$$

One can consider the vector space of linear combinations of such exterior products; this is the space of antisymmetric tensors of rank (2,0). Similarly, the exterior product of n vectors is an antisymmetric tensor of rank (n,0). Such totally antisymmetric tensors are sometimes called n-**vectors** or **multivectors**; for instance, $\mathbf{x} \wedge \mathbf{y}$ is called a **bivector**.

Using this construction, one can reinterpret an n-form as a linear map from n-vectors into numbers. For example, a 2-form ω acts on a bivector $\mathbf{x} \wedge \mathbf{y}$ as follows,

$$\omega(\mathbf{x} \wedge \mathbf{y}) = \omega(\mathbf{x} \otimes \mathbf{y} - \mathbf{y} \otimes \mathbf{x}) = \omega(\mathbf{x}, \mathbf{y}) - \omega(\mathbf{y}, \mathbf{x}) = 2\omega(\mathbf{x}, \mathbf{y}).$$

Notice the appearance of an extra factor 2. In the case of n-forms, this extra factor will be $n!$ since we will need to sum over $n!$ possible permutations of n vectors.

Remark: Some textbooks use extra factors of $n!$ when defining the exterior product of n 1-forms; for instance, one could define $(\eta \wedge \theta) \circ (\mathbf{x}, \mathbf{y}) = \frac{1}{2}(\eta(\mathbf{x})\theta(\mathbf{y}) - \eta(\mathbf{y})\theta(\mathbf{x}))$. Then extra factors of $n!$ will appear in some equations but disappear from some other equations. These factors play no essential role, i.e. they are **cosmetic** (they merely make some equations better looking). Of course, one must keep track of the extra factors when one uses formulas from different books. ■

An n-vector $\mathbf{a}_1 \wedge ... \wedge \mathbf{a}_n$ can serve as a representation of the n-dimensional volume of a parallelepiped spanned by a set of n given vectors $\{\mathbf{a}_1, ..., \mathbf{a}_n\}$ in an N-dimensional space (where $N \geq n$). Statement 1.4.4.1 shows that the (oriented) volumes of two such parallelepipeds spanned by the sets $\{\mathbf{a}_1, ..., \mathbf{a}_n\}$ and $\{\mathbf{b}_1, ..., \mathbf{b}_n\}$ are equal iff the two corresponding n-vectors, $\mathbf{a}_1 \wedge ... \wedge \mathbf{a}_n$ and $\mathbf{b}_1 \wedge ... \wedge \mathbf{b}_n$, are equal. This is proved by geometric arguments similar to those in Sec. 1.4.1, and by using an n-dimensional generalization of Fig. 1.10. Note that an analogon of Statement 1.4.3.1 holds also for vectors: an n-vector is nonzero iff the set of its constituent vectors is linearly independent.

Statement 1.4.4.1: We consider an n-dimensional space $\mathbb{R}^n$. **(a)** Every n-vector $\mathbf{a}_1 \wedge ... \wedge \mathbf{a}_n$ is proportional to every other n-vector. **(b)** The ratio of the oriented volumes of two n-dimensional parallelepipeds spanned by the sets $\{\mathbf{a}_1, ..., \mathbf{a}_n\}$ and $\{\mathbf{b}_1, ..., \mathbf{b}_n\}$ are related by the same proportionality factor as the two corresponding n-vectors, $\mathbf{a}_1 \wedge ... \wedge \mathbf{a}_n$ and $\mathbf{b}_1 \wedge ... \wedge \mathbf{b}_n$. (Proof on page 176.) ■

The ordinary (scalar) volume of the parallelepiped spanned by $\{\mathbf{a}_1, ..., \mathbf{a}_n\}$ can be calculated as a suitably defined "absolute value" or the **norm** of the n-vector $\mathbf{a}_1 \wedge ... \wedge \mathbf{a}_n$,

$$\text{Vol}\,(\mathbf{a}_1, ..., \mathbf{a}_n) = \sqrt{|\mathbf{a}_1 \wedge ... \wedge \mathbf{a}_n|^2}.$$

The norm $|\mathbf{a}_1 \wedge ... \wedge \mathbf{a}_n|$ is defined in the usual manner through the scalar product on the space of n-vectors, which is in turn defined through the scalar product in the vector space. For instance, the area of a parallelogram spanned by two vectors $\{\mathbf{a}, \mathbf{b}\}$ in an N-dimensional Euclidean space $\mathbb{R}^N$ is calculated as follows. The "oriented area" of the parallelogram is represented by the bivector $\mathbf{a} \wedge \mathbf{b}$. Using the standard Cartesian basis $\{\mathbf{e}_1, ..., \mathbf{e}_N\}$ in $\mathbb{R}^N$, one can decompose this bivector into basis bivectors $\{\mathbf{e}_i \wedge \mathbf{e}_j\}$, where $1 \leq i < j \leq N$, with some coefficients A_{ij},

$$\mathbf{a} \wedge \mathbf{b} = \sum_{1 \leq i < j \leq N} A_{ij} \mathbf{e}_i \wedge \mathbf{e}_j.$$

The scalar product is then defined in the space of bivectors by postulating that the bivectors $\{\mathbf{e}_i \wedge \mathbf{e}_j\}$ constitute an orthonormal basis. The scalar product of two arbitrary bivectors is then computed as

$$\left(\sum_{i,j} A_{ij} \mathbf{e}_i \wedge \mathbf{e}_j \right) \cdot \left(\sum_{i,j} A_{ij} \mathbf{e}_i \wedge \mathbf{e}_j \right) = \sum_{i,j} A_{ij} B_{ij}$$

(all the sums are taken over $1 \leq i < j \leq N$). Hence, the ordinary (scalar) area of the parallelogram is

$$\text{Area}\,(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{1 \leq i < j \leq N} |A_{ij}|^2}.$$

1.4.5 Determinants

A standard definition of the determinant of a square $n \times n$ matrix A_{ij} is a formula involving the matrix elements:

$$\det\left(A_{ij}\right) \equiv \sum_{\sigma} (-1)^{|\sigma|} A_{1\sigma(1)} ... A_{n\sigma(n)},$$

where the sum is performed over all transpositions σ. A "transposition" is a one-to-one map from the set $\{1, 2, ..., n\}$ to the same set. The quantity $|\sigma|$ is defined as 0 when σ is an **even** transposition (equivalent to an even number of pair interchanges) and 1 if σ is an odd transposition. This formula is explicit but complicated. The geometric meaning and the properties of the determinant are much more apparent if one adopts a different (but equivalent) definition.

We define the **determinant** $\det \hat{T}$ of a linear transformation $\hat{T}$ in an n-dimensional Euclidean space $\mathbb{R}^n$ as the volume of the image of the unit cube under the transformation $\hat{T}$. In other words, we select an orthonormal basis $\{\mathbf{e}_1, ..., \mathbf{e}_n\}$ and transform it using $\hat{T}$, obtaining the vectors $\{\hat{T}\mathbf{e}_1, ..., \hat{T}\mathbf{e}_n\}$. By definition, $\det \hat{T}$ is equal to the oriented volume of the n-dimensional parallelepiped spanned by these n vectors.

A little work using the concept of multivectors shows that the the initial set of vectors $\{\mathbf{e}_1, ..., \mathbf{e}_n\}$ does not actually need to be an orthonormal basis; this is desirable, since we can then define the determinant of a transformation $\hat{T}$ without the necessity to have a scalar product in the vector space.

Statement 1.4.5.1: Let $\{\mathbf{v}_1, ..., \mathbf{v}_n\}$ be n arbitrary vectors; denote by $\text{Vol}\,(\mathbf{v}_1, ..., \mathbf{v}_n)$ the oriented volume of the parallelepiped spanned by these vectors. Consider also the parallelepiped spanned by the transformed vectors $\{\hat{T}\mathbf{v}_1, ..., \hat{T}\mathbf{v}_n\}$. Then the transformed volume is proportional to the original volume with the factor $\det \hat{T}$,

$$\text{Vol}(\hat{T}\mathbf{v}_1, ..., \hat{T}\mathbf{v}_n) = \text{Vol}\,(\mathbf{v}_1, ..., \mathbf{v}_n) \cdot \det \hat{T}.$$

This factor is independent of the choice of the vectors $\{\mathbf{v}_j\}$. A similar relationship holds for the corresponding multivectors,

$$\hat{T}\mathbf{v}_1 \wedge \hat{T}\mathbf{v}_2 \wedge ... \wedge \hat{T}\mathbf{v}_n = (\mathbf{v}_1 \wedge \mathbf{v}_2 \wedge ... \wedge \mathbf{v}_n) \det \hat{T}.$$

(Proof on page 176.) ■

Due to Statement 1.4.5.1, we can reformulate the definition of the determinant as follows: The determinant of $\hat{T}$ is equal to the factor that multiplies the volume of an *arbitrary* n-dimensional parallelepiped after the transformation $\hat{T}$.

Using this definition, it is easy to derive the fundamental property of determinants: the determinant of a product of two transformations is equal to the product of the two determinants,

$$\det(\hat{A}\hat{B}) = (\det \hat{A})(\det \hat{B}).$$

This property is a simple consequence of the fact that the volume of every parallelepiped transformed by $\hat{A}$ is multiplied by $\det \hat{A}$.

Finally, let us prove an important relationship involving volume in Euclidean space.

Statement 1.4.5.2: The volume of a parallelepiped spanned by vectors $\{\mathbf{v}_1, ..., \mathbf{v}_n\}$ in a Euclidean space $\mathbb{R}^n$ is

$$\text{Vol}\,(\mathbf{v}_1, ..., \mathbf{v}_n) = \sqrt{\det g_{ij}}, \quad g_{ij} \equiv \mathbf{v}_i \cdot \mathbf{v}_j,$$

where $\mathbf{v}_i \cdot \mathbf{v}_j$ is the scalar product of two vectors.

Proof of Statement 1.4.5.2: Let us consider an orthonormal basis $\{\mathbf{e}_j\}$ in this space. The volume of a parallelepiped spanned by $\{\mathbf{e}_j\}$ is, by definition, equal to 1. Since $\{\mathbf{e}_j\}$ is a basis, there exists a linear transformation $\hat{T}$ that brings $\{\mathbf{e}_j\}$ into the given set of vectors $\{\mathbf{v}_j\}$. By Statement 1.4.5.1, the volume of the parallelepiped spanned by $\{\mathbf{v}_j\}$ is equal to $\det \hat{T}$, so it remains to compute that determinant. Let us introduce the components of the transformation $\hat{T}$ in the basis $\{\mathbf{e}_j\}$,

$$\mathbf{v}_j \equiv \hat{T}\mathbf{e}_j = \sum_k T_{jk} \mathbf{e}_k.$$

Since $\mathbf{e}_k \cdot \mathbf{e}_l = \delta_{kl}$, we can express the matrix g_{ij} as

$$g_{ij} = \mathbf{v}_i \cdot \mathbf{v}_j = \left(\sum_k T_{ik} \mathbf{e}_k \right) \cdot \left(\sum_l T_{jl} \mathbf{e}_l \right) = \sum_k T_{ik} T_{jk}.$$

It follows that the matrix g_{ij} is equal to the product of two matrices T_{ij}. Therefore, the determinant of g_{ij} is found as

$$\det g_{ij} = \left(\det T_{ij}\right) \left(\det T_{ji}\right) = \left(\det T_{ij}\right)^2,$$

where we used the fact that the determinant of the transposed matrix is the same as the determinant of the original matrix. Hence,

$$\text{Vol}\,(\mathbf{v}_1, ..., \mathbf{v}_n) = \det T_{ij} = \sqrt{\det g_{ij}},$$

which is the desired formula. ■

1.4.6 Differential forms

Let us now consider **differential forms**, i.e. n-forms defined in the tangent space $T_p\mathcal{M}$ at every point p of a manifold $\mathcal{M}$. An example of a differential form written in local coordinates $\{x, y, z\}$ is

$$\omega = x^2 \mathrm{d}y \wedge \mathrm{d}z - 3y^3 \mathrm{d}x \wedge \mathrm{d}y$$

(in this example, ω is a 2-form).

For scalar functions ("0-forms") f, the 1-form $\mathrm{d}f$ is defined by

$$(\mathrm{d}f) \circ \mathbf{v} \equiv \mathbf{v} \circ f.$$

So the notation "$\mathrm{d}x$" itself can be understood as the differential operator d acting on the function x. It follows that $\mathrm{d}f(x) = f'(x)\mathrm{d}x$, which is a familiar rule of calculus (the chain rule). For this reason, the notation $\mathrm{d}x$ for 1-forms is convenient.

The operator d can be generalized to a linear operator acting on n-forms and producing $(n+1)$-forms; this operator is then called the **exterior differential**. So the exterior differential of an n-form ω is an $(n+1)$-form written as $\mathrm{d}\omega$.

The exterior differential d can be defined either explicitly (see Eq. (1.22) below), or through the properties it satisfies. These properties are the following. For any forms ω_1 and ω_2 and scalar functions f we have the Leibnitz-like rule

$$\mathrm{d}\,(f\omega_1) = \mathrm{d}f \wedge \omega_1 + f\mathrm{d}\omega_1,$$
$$\mathrm{d}(\omega_1 \wedge \omega_2) = (\mathrm{d}\omega_1) \wedge \omega_2 + (-1)^{n_1}\omega_1 \wedge \mathrm{d}\omega_2,$$

where ω_1 is an n_1-form. Heuristically, one gets n sign changes when one "pulls d through an n-form."

The second fundamental property of the exterior differential is $\mathrm{d}\,(\mathrm{d}\omega) = 0$ for any n-form ω. This property can be understood heuristically as follows: "d d" is symmetric with respect to the exchange of d with d; each "d" adds a vector argument into the form, but the result must be a totally antisymmetric form. So we must have $\mathrm{d}\,\mathrm{d} = 0$.

Due to the simplicity of these rules, calculations with differential forms are straightforward.

Example: Consider 1-forms $\omega_1 = x^2\mathrm{d}y$ and $\omega_2 = xy\mathrm{d}x + 2\mathrm{d}y$. The exterior differential of ω_1 is

$$\mathrm{d}\omega_1 = \mathrm{d}\left(x^2\mathrm{d}y\right) = 2x\mathrm{d}x \wedge \mathrm{d}y + x^2\mathrm{d}\,(\mathrm{d}y) = 2x\mathrm{d}x \wedge \mathrm{d}y,$$

because $\mathrm{d}\,(\mathrm{d}y) = 0$. Let us compute $\mathrm{d}\omega_2$:

$$\begin{aligned}\mathrm{d}\omega_2 &= \mathrm{d}(xy) \wedge \mathrm{d}x + xy\mathrm{d}\,(\mathrm{d}x) + 2\mathrm{d}\,(\mathrm{d}y)\\ &= y\mathrm{d}x \wedge \mathrm{d}x + x\mathrm{d}y \wedge \mathrm{d}x\\ &= -x\mathrm{d}x \wedge \mathrm{d}y,\end{aligned}$$

because $\mathrm{d}x \wedge \mathrm{d}x = 0$ and $\mathrm{d}y \wedge \mathrm{d}x = -\mathrm{d}x \wedge \mathrm{d}y$ due to antisymmetry. Consider now the differential of $\mathrm{d}\omega_2$,

$$\mathrm{d}\,(\mathrm{d}\omega_2) = \mathrm{d}\,(-x\mathrm{d}x \wedge \mathrm{d}y) = -\mathrm{d}x \wedge \mathrm{d}x \wedge \mathrm{d}y = 0.$$

Another example calculation is

$$\mathrm{d}\,(xy\mathrm{d}x \wedge \mathrm{d}z) = \mathrm{d}\,(xy) \wedge \mathrm{d}x \wedge \mathrm{d}z = -x\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z.$$

In this way, arbitrary differentials can be computed. ■

Practice problems: **(a)** Evaluate the 3-form $\omega_1 \wedge \mathrm{d}\omega_2$ if $\omega_1 = x\mathrm{d}x + y\mathrm{d}y + z\mathrm{d}z$ and $\omega_2 = x\mathrm{d}y - y\mathrm{d}x$.
(b) Compute the application of the 2-form $\mathrm{d}(xy) \wedge (y\mathrm{d}x - \mathrm{d}y)$ to two vectors $\mathbf{v}_1 = x\partial_x$, $\mathbf{v}_2 = \partial_y + \partial_z$.
Answers: **(a)** $2z\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z$. **(b)** $-xy\,(x+1)$. ■

A rather cumbersome but explicit formula can be derived for the differential $\mathrm{d}\omega$ of an arbitrary n-form ω. The $(n+1)$-form $\mathrm{d}\omega$ acts on vectors $\mathbf{v}_1$, ..., $\mathbf{v}_{n+1}$ as follows,

$$\begin{aligned}(\mathrm{d}\omega) \circ (\mathbf{v}_1, ..., \mathbf{v}_{n+1}) \equiv &\sum_{s=1}^{n+1} (-1)^{s-1} \mathbf{v}_s \circ \omega(\mathbf{v}_1, ..., \hat{\mathbf{v}}_s, ..., \mathbf{v}_{n+1})\\ &+ \sum_{1 \le r < s \le n+1} (-1)^{r+s-1} \omega([\mathbf{v}_r, \mathbf{v}_s], \mathbf{v}_1, ..., \hat{\mathbf{v}}_r, ..., \hat{\mathbf{v}}_s, ..., \mathbf{v}_{n+1}),\end{aligned} \tag{1.22}$$

where the hat over a vector, $\hat{\mathbf{v}}_s$, indicates the *absence* of the vector $\mathbf{v}_s$ among the listed arguments of ω. As an example of using this formula, consider a 1-form ω, then $\mathrm{d}\omega$ is the 2-form defined by

$$(\mathrm{d}\omega) \circ (\mathbf{x}, \mathbf{y}) \equiv \mathbf{x} \circ (\omega(\mathbf{y})) - \mathbf{y} \circ (\omega(\mathbf{x})) - \omega([\mathbf{x}, \mathbf{y}]). \tag{1.23}$$

It is straightforward to check that this formula actually defines a bilinear map $\mathrm{d}\omega$ that does not contain derivatives of $\mathbf{x}$ or $\mathbf{y}$.

Statement 1.4.6.1: Equation (1.23) defines a bilinear form $\mathrm{d}\omega$ despite the apparent presence of derivatives of $\mathbf{x}$ and $\mathbf{y}$.

Proof of Statement 1.4.6.1: Obviously the formula (1.23) defines an antisymmetric function of $(\mathbf{x}, \mathbf{y})$. So it is sufficient to show that

$$(\mathrm{d}\omega) \circ (\lambda\mathbf{x}, \mathbf{y}) = \lambda(\mathrm{d}\omega) \circ (\mathbf{x}, \mathbf{y})$$

when λ is a scalar function of a point. Once this is proved, it will follow that $\mathrm{d}\omega$ does not contain derivatives of $\mathbf{x}$ or $\mathbf{y}$.

The remaining part of the proof is thus a straightforward calculation:

$$\begin{aligned}(\mathrm{d}\omega) \circ (\lambda\mathbf{x}, \mathbf{y}) &= \lambda\mathbf{x} \circ \omega(\mathbf{y}) - \mathbf{y} \circ (\lambda\omega(\mathbf{x})) - \omega([\lambda\mathbf{x}, \mathbf{y}])\\ &= \lambda\left(\mathbf{x} \circ \omega(\mathbf{y}) - \mathbf{y} \circ \omega(\mathbf{x}) - \omega([\mathbf{x}, \mathbf{y}])\right)\\ &\quad - (\mathbf{y} \circ \lambda)\,\omega(\mathbf{x}) + \omega\left((\mathbf{y} \circ \lambda)\,\mathbf{x}\right)\\ &= \lambda(\mathrm{d}\omega) \circ (\mathbf{x}, \mathbf{y}).\end{aligned}$$

Here we used Statement 1.3.1.1 to express $[\lambda\mathbf{x}, \mathbf{y}]$. ■

There is a useful relationship between the Lie derivative and the exterior differential. For convenience of notation, one introduces the **insertion** operation $\iota_{\mathbf{v}}$ (also called the **interior product**) that "inserts" the vector $\mathbf{v}$ into n-forms as the first argument,

$$(\iota_{\mathbf{v}}\omega) \circ (\mathbf{v}_1, ..., \mathbf{v}_{n-1}) \equiv \omega\,(\mathbf{v}, \mathbf{v}_1, ..., \mathbf{v}_{n-1})\,.$$

The insertion operator produces $(n-1)$-forms out of n-forms.

The Lie derivative operation can be expressed through ι and d. To see how, let us first compute $\mathcal{L}_{\mathbf{v}}\omega$ for some 1-form ω and vector field $\mathbf{v}$. The 1-form $\mathcal{L}_{\mathbf{v}}\omega$ acts on an arbitrary vector $\mathbf{u}$ as

$$(\mathcal{L}_{\mathbf{v}}\omega) \circ \mathbf{u} = \mathcal{L}_{\mathbf{v}}\,(\omega \circ \mathbf{u}) - \omega \circ (\mathcal{L}_{\mathbf{v}}\mathbf{u}) = \mathbf{v} \circ (\omega(\mathbf{u})) - \omega([\mathbf{v}, \mathbf{u}]).$$

Let us compare this with the formula (1.23) for $\mathrm{d}\omega$; by inspection, we notice that

$$(\mathcal{L}_{\mathbf{v}}\omega) \circ \mathbf{u} = (\mathrm{d}\omega) \circ (\mathbf{v}, \mathbf{u}) + \mathbf{u} \circ (\omega(\mathbf{v}))\,.$$

Using the insertion operator $\iota_{\mathbf{v}}$, we can rewrite this as

$$(\mathcal{L}_{\mathbf{v}}\omega) \circ \mathbf{u} = (\iota_{\mathbf{v}}\,(\mathrm{d}\omega)) \circ \mathbf{u} + \mathbf{u} \circ (\iota_{\mathbf{v}}\omega)\,.$$

Finally, note that $\iota_{\mathbf{v}}\omega$ is a scalar function, so

$$\mathbf{u} \circ (\iota_{\mathbf{v}}\omega) \equiv (\mathrm{d}\,(\iota_{\mathbf{v}}\omega)) \circ \mathbf{u}.$$

Thus

$$(\mathcal{L}_{\mathbf{v}}\omega) \circ \mathbf{u} = (\iota_{\mathbf{v}}\,(\mathrm{d}\omega) + \mathrm{d}\,(\iota_{\mathbf{v}}\omega)) \circ \mathbf{u}.$$

Omitting the arbitrary vector $\mathbf{u}$, we obtain the **Cartan homotopy formula**

$$\mathcal{L}_{\mathbf{v}}\omega = \mathrm{d}\,(\iota_{\mathbf{v}}\omega) + \iota_{\mathbf{v}}\,(\mathrm{d}\omega)\,,$$

written also more concisely as

$$\mathcal{L}_{\mathbf{v}} = \mathrm{d}\iota_{\mathbf{v}} + \iota_{\mathbf{v}}\mathrm{d}. \tag{1.24}$$

This important formula holds not only for 1-forms, but also for arbitrary n-forms ω. Note that the exterior differential d brings n-forms into $(n+1)$-forms, while $\mathcal{L}_{\mathbf{v}}$ does not change the number of arguments of n-forms.

The Cartan homotopy formula for n-forms can be derived directly from the explicit definition (1.22) of $\mathrm{d}\omega$ and the Leibnitz property of $\mathcal{L}_{\mathbf{v}}$, for instance by computing $\iota_{\mathbf{v}}\mathrm{d}\omega$ and showing that $\iota_{\mathbf{v}}\mathrm{d}\omega = \mathcal{L}_{\mathbf{v}}\omega - \mathrm{d}\iota_{\mathbf{v}}\omega$. See also e.g. the book [33], §4.5. I omit this straightforward verification; unconvinced readers are encouraged to try an explicit calculation for a 2-form.

The following properties also hold for arbitrary n-form ω, n'-form ω', and vector fields $\mathbf{v}, \mathbf{x}$:

$$\begin{aligned}
\iota_{\mathbf{x}}\left(\omega \wedge \omega'\right) &= (\iota_{\mathbf{x}}\omega) \wedge \omega' + (-1)^n\, \omega \wedge \iota_{\mathbf{x}}\omega', \\
\left(\mathcal{L}_{\mathbf{v}}\mathrm{d}\right)\omega &= \left(\mathrm{d}\mathcal{L}_{\mathbf{v}}\right)\omega, \\
\mathcal{L}_{\mathbf{v}}\left(\omega \wedge \omega'\right) &= \left(\mathcal{L}_{\mathbf{v}}\omega\right) \wedge \omega' + \omega \wedge \mathcal{L}_{\mathbf{v}}\omega', \\
\left(\mathcal{L}_{\mathbf{v}}\iota_{\mathbf{x}}\right)\omega &= \iota_{[\mathbf{v},\mathbf{x}]}\omega + \iota_{\mathbf{x}}\mathcal{L}_{\mathbf{v}}\omega.
\end{aligned}$$

These properties can be verified by straightforward computations which I omit.

1.4.7 *Canonical decomposition of 1-forms and 2-forms

Literature: This material is covered in [2], volume 2, chapter 22, Appendix; see also [16], chapter 5.

In this section I review some classical results obtained in the theory of differential forms. The main statements are the following.

With a suitable choice of local coordinates $\{x_1, ..., x_n\}$ in a suitable domain of an n-dimensional manifold:

- A given 1-form ω can be expressed in one of two ways: either as
$$\omega = x_1\mathrm{d}x_2 + ... + x_{2k-1}\mathrm{d}x_{2k},$$
or as
$$\omega = x_1\mathrm{d}x_2 + ... + x_{2k-1}\mathrm{d}x_{2k} + \mathrm{d}x_{2k+1}.$$
- (The Darboux theorem) A given *closed* 2-form Ω can be expressed as
$$\Omega = \mathrm{d}x_1 \wedge \mathrm{d}x_2 + ... + \mathrm{d}x_{2k-1} \wedge \mathrm{d}x_{2k}.$$
- (The Poincaré lemma) Any closed n-form ω is locally exact: if $\mathrm{d}\omega = 0$, there exists an $(n-1)$-form θ such that $\omega = \mathrm{d}\theta$ in some domain (but perhaps not in the entire manifold).

Canonical decomposition of closed 2-forms. We consider an n-dimensional smooth manifold. In a local coordinate system $\{x_1, ..., x_n\}$, an arbitrary 2-form Ω can be expressed as

$$\Omega = \frac{1}{2}\sum_{i,j=1}^{n} \Omega_{ij}(x_1, ..., x_n)\mathrm{d}x_i \wedge \mathrm{d}x_j, \tag{1.25}$$

where $\Omega_{ij} = -\Omega_{ji}$ is an x-dependent matrix. However, the coordinate system may be chosen so that a particular 2-form is written in a simpler way. For example, a closed 2-form

$$\boldsymbol{A} = \mathrm{d}x_1 \wedge \mathrm{d}x_2 + x_3\mathrm{d}x_1 \wedge \mathrm{d}x_3$$

can be rewritten as

$$\boldsymbol{A} = \mathrm{d}x_1 \wedge \mathrm{d}\left(x_2 + \frac{1}{2}x_3^2\right) = \mathrm{d}x_1 \wedge \mathrm{d}y_2, \quad y_2 \equiv x_2 + \frac{1}{2}x_3^2.$$

In this example, a particular choice of the new coordinates $\{x_1, y_2, x_3, ..., x_n\}$ reduces the 2-form $\boldsymbol{A}$ to a simpler expression involving just two of the coordinates.

It is useful to know how many different coordinates are needed to represent a given 2-form Ω in the simplest possible way. Of course, it is also useful to be able to *determine* the new coordinates explicitly if a closed 2-form Ω is specified in a given coordinate system.

The **Darboux theorem** (proved below) says that for any closed 2-form Ω a suitable local coordinate system $\{x_1, ..., x_n\}$ can be found such that Ω is *locally* (i.e. within a certain domain) expressed as

$$\Omega = \mathrm{d}x_1 \wedge \mathrm{d}x_2 + \mathrm{d}x_3 \wedge \mathrm{d}x_4 + ... + \mathrm{d}x_{2k-1} \wedge \mathrm{d}x_{2k}, \tag{1.26}$$

where $k \leq n/2$ is some number.

It is advantageous to find such **canonical** local coordinates because then calculations with Ω are much easier. The required number $2k \leq n$ of different coordinates in the decomposition (1.26) is called the **rank** of Ω.

If a closed 2-form Ω is given in a certain coordinate system, can one determine the rank of Ω without knowing the canonical coordinates (but knowing that they exist)? In the example with the 2-form $\boldsymbol{A}$ above, it is easy to guess the correct canonical coordinates. Consider another example,

$$\boldsymbol{B} = \mathrm{d}y_1 \wedge \mathrm{d}y_2 + y_1\mathrm{d}y_1 \wedge \mathrm{d}y_3.$$

In this case, it is not obvious how to find suitable canonical coordinates $\{x_1, x_2, x_3, ...\}$ such that $\boldsymbol{B}$ is expressed as in Eq. (1.26).

To determine the rank of a given closed 2-form Ω, the following trick can be used. One considers exterior powers of Ω, denoted

$$\Omega^{\wedge k} = \underbrace{\Omega \wedge ... \wedge \Omega}_{k \text{ times}}.$$

Note that for a 2-form Ω the product $\Omega \wedge \Omega$ does not necessarily vanish. A sufficiently high power of Ω will, of course, vanish. So there will be some number $k \leq n/2$ such that

$$\Omega^{\wedge k} \neq 0, \quad \Omega^{\wedge(k+1)} = 0. \tag{1.27}$$

From the Darboux theorem we know that the decomposition (1.26) is possible in some coordinates $\{x_j\}$. In these coordinates, we can compute

$$\Omega^{\wedge k} = k!\, \mathrm{d}x_1 \wedge ... \wedge \mathrm{d}x_{2k} \neq 0,$$

since by assumption all $\{x_j\}$ are independent coordinates in a local coordinate system. However, $\Omega^{\wedge(k+1)} = 0$. Thus the rank of Ω is $2k$, where k is the largest integer such that $\Omega^{\wedge k} \neq 0$ but $\Omega^{\wedge(k+1)} = 0$. If $\Omega = 0$, we say that its rank is zero.

For example, we can immediately see that the 2-form $\boldsymbol{B}$ specified above has rank 2 because $\boldsymbol{B} \wedge \boldsymbol{B} = 0$. Thus, we should expect to find suitable coordinates $\{x_1, x_2, x_3, ...\}$ such that $\boldsymbol{B} = \mathrm{d}x_1 \wedge \mathrm{d}x_2$. Determining these coordinates in practice may be far from easy; a possible method of finding $\{x_j\}$ is given in the proof of the Darboux theorem below.

Remark: The rank of a 2-form may be different at different points of the manifold. Presently, we assume that we will be dealing only with such 2-forms whose rank remains constant *locally*, i.e. throughout some neighborhood of some point. ■

The maximal rank of a 2-form on an n-dimensional manifold is n (for even n) or $n-1$ (for odd n). For a $2k$-dimensional manifold, a 2-form Ω can be **nondegenerate** if the matrix Ω_{ij} is invertible (has nonzero determinant). This condition can be expressed in the following way: for any nonzero vector $\mathbf{x}$ there exists at least one vector $\mathbf{y}$ such that $\Omega(\mathbf{x}, \mathbf{y}) \neq 0$; in other words, the 1-form $\iota_{\mathbf{x}}\Omega$ is a nonvanishing 1-form. Equivalently: if $\iota_{\mathbf{x}}\Omega = 0$ for some vector $\mathbf{x}$ then $\mathbf{x} = 0$.

Statement 1.4.7: For a $2k$-dimensional manifold, a 2-form Ω of maximal rank $2k$ is nondegenerate.

Proof: Suppose a vector $\mathbf{x}$ is such that the 1-form $\iota_{\mathbf{x}}\Omega$ vanishes, $\iota_{\mathbf{x}}\Omega = 0$. We would like to prove that $\mathbf{x} = 0$. It follows from $\iota_{\mathbf{x}}\Omega = 0$ that

$$\iota_{\mathbf{x}} \left(\Omega^{\wedge k} \right) = k \left(\iota_{\mathbf{x}} \Omega \right) \wedge \Omega^{\wedge(k-1)} = 0.$$

The $2k$-form $\Omega^{\wedge k}$ is nonzero since Ω has rank $2k$, therefore $\Omega^{\wedge k}$ is proportional to the volume form $\mathrm{d}x_1 \wedge ... \wedge \mathrm{d}x_{2k}$ with a nonzero coefficient. It follows that

$$\iota_{\mathbf{x}} \left(\mathrm{d}x_1 \wedge ... \wedge \mathrm{d}x_{2k} \right) = 0.$$

Writing $\mathbf{x} = \sum_{j=1}^{n} a_j \partial / \partial x_j$, it is straightforward to see that all the coefficients a_j must vanish. Therefore, $\mathbf{x} = 0$. ■

Assuming that the Darboux theorem is true, it is easy to see that a nondegenerate 2-form always has maximal rank. If Ω does not have maximal rank, the canonical decomposition of a 2-form Ω is free of some of the coordinates, say of x_n. It follows that the vector $\mathbf{v} \equiv \partial / \partial x_n$ satisfies $\iota_{\mathbf{v}}\Omega = 0$. Hence, Ω is degenerate.

An algebraic version of the Darboux theorem is the following.

Statement 1.4.7: An arbitrary antisymmetric bilinear form (2-form) B can be expressed as

$$B = \theta_1 \wedge \theta_2 + ... + \theta_{2k-1} \wedge \theta_{2k}, \tag{1.28}$$

where θ_j are some suitable 1-forms. The set $\{\theta_j\}$ of these 1-forms is linearly independent. The smallest required number ($2k$) of these n-forms is the rank of B, which is equal to the rank of the matrix B_{ij} in any basis.

Proof: The choice of a cnonical basis for an antisymmetric bilinear form B is a standard task of finite-dimensional linear algebra. We consider an n-dimensional vector space. If $B = 0$, the rank is zero and the statement is proved. If $B \neq 0$, there exists a vector $\mathbf{e}_1$ such that the 1-form $\iota_{\mathbf{e}_1} B$ is nonzero. Thus there exists another vector $\mathbf{e}_2$ such that $B(\mathbf{e}_1, \mathbf{e}_2) = 1$. Since B is antisymmetric, the vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ cannot be parallel to each other, thus they can be completed to a basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, ..., \mathbf{e}_n\}$. We can choose the other basis vectors $\mathbf{e}_3$, ..., $\mathbf{e}_n$ such that they are orthogonal to $\mathbf{e}_1$ and $\mathbf{e}_2$ with respect to B, i.e.

$$B(\mathbf{e}_1, \mathbf{e}_j) = B(\mathbf{e}_2, \mathbf{e}_j) = 0, \quad j = 3, ..., n.$$

This choice is always possible because we may add to each $\mathbf{e}_j$, $j = 3, ..., n$ a suitable linear combination of $\mathbf{e}_1$ and $\mathbf{e}_2$,

$$\tilde{\mathbf{e}}_j = \mathbf{e}_j - B(\mathbf{e}_1, \mathbf{e}_j)\mathbf{e}_2 + B(\mathbf{e}_2, \mathbf{e}_j)\mathbf{e}_1,$$

so that now $B(\mathbf{e}_1, \tilde{\mathbf{e}}_j) = B(\mathbf{e}_2, \tilde{\mathbf{e}}_j) = 0$ for every $j = 3, ..., n$. For brevity, let us denote the resulting basis again by $\{\mathbf{e}_j\}$. The result of this construction is that the two subspaces spanned by $\{\mathbf{e}_1, \mathbf{e}_2\}$ and by $\{\mathbf{e}_3, ..., \mathbf{e}_n\}$ are orthogonal to each other with respect to the bilinear form B.

After choosing the basis $\{\mathbf{e}_1, ..., \mathbf{e}_n\}$ in this way, we compute the dual basis 1-forms $\{\theta_1, ..., \theta_n\}$ for the basis $\{\mathbf{e}_1, ..., \mathbf{e}_n\}$ and represent the 2-form B as

$$B = \theta_1 \wedge \theta_2 + B^{(1)},$$

where the new 2-form $B^{(1)}$ is equal to zero on $\mathbf{e}_1, \mathbf{e}_2$. Hence, it is sufficient to consider the 2-form $B^{(1)}$ within the $(n-2)$-dimensional subspace spanned by $\{\mathbf{e}_3, ..., \mathbf{e}_n\}$. We can now apply the same construction to $B^{(1)}$ in a smaller number of dimensions; either $B^{(1)} = 0$, or we can find θ_3 and θ_4 such that $B^{(1)} = \theta_3 \wedge \theta_4 + B^{(2)}$, etc. The construction eventually stops because at every step the dimensionality of the space is reduced by two. At that point, we will obtain a decomposition of the form (1.28), where all 1-forms θ_j are (by construction) a linearly independent set.

In the basis $\{\mathbf{e}_j\}$, the 2-form B is represented by the $n \times n$ matrix

$$\begin{pmatrix} 0 & 1 & & & & & & \\ -1 & 0 & & 0 & & & 0 & \\ & & \ddots & & & & & \\ & 0 & & 0 & 1 & & & \\ & & & -1 & 0 & & & \\ & & & & & 0 & & \\ & 0 & & & & & \ddots & \\ & & & & & & & 0 \end{pmatrix},$$

in which the lower right $(n-2k) \times (n-2k)$ block is zero.

It is straightforward to verify that the rank of the 2-form B is indeed $2k$ according to the definition (1.27) of the rank. The $2k$-form $B^{\wedge k}$ is nonzero due to the linear independence of $\{\theta_j\}$,

$$B^{\wedge k} = k!\, \theta_1 \wedge ... \wedge \theta_{2k},$$

while $B^{\wedge(k+1)} = 0$. ■

The difference between the algebraic and the differential-geometric versions of the Darboux theorem is the following. According to the algebraic theorem just proved, one may reduce any 2-form Ω to a canonical decomposition at any one point p. This decomposition will involve a certain choice of the 1-forms $\{\theta_j|_p\}$ at each p. However, this procedure will generally generate different sets $\{\theta_j|_p\}$ in tangent spaces at different points p, so there will be no local coordinate system $\{x_j\}$ in which the 1-forms θ_j are equal to the coordinate 1-forms $\mathrm{d}x_j$ at every point p. Using the algebraic version of the Darboux theorem, one could obtain the decomposition (1.26) at any one point p but not at neighboring points.

The differential-geometric Darboux theorem says that one can actually choose a local coordinate system in which the 2-form Ω has a canonical decomposition (1.26) at every point within some domain.

The Darboux theorem. A closed 2-form Ω of local rank $2k$ in a n-dimensional manifold can be written as

$$\Omega = \mathrm{d}x_1 \wedge \mathrm{d}x_2 + ... + \mathrm{d}x_{2k-1} \wedge \mathrm{d}x_{2k}$$

in a suitable local coordinate system.

This theorem is proved in three steps. The first step is to show that there exists a local coordinate system $\{x_1, ..., x_n\}$ such that Ω depends *only* on the first $2k$ coordinates $\{x_1, ..., x_{2k}\}$, i.e.

$$\Omega = \frac{1}{2} \sum_{i,j=1}^{2k} \Omega_{ij}(x_1, ..., x_{2k}) \mathrm{d}x_i \wedge \mathrm{d}x_j.$$

Then it is sufficient to consider the 2-form Ω on a $2k$-dimensional submanifold where the remaining coordinates $\{x_{2k+1}, ..., x_n\}$ have fixed values.

The first step can be proved as follows. If a 2-form $\Omega \neq 0$ has rank $2k < n$ then, by Statement 1.4.7, at every point there exists a decomposition of the form (1.28). It follows that in the tangent space at every point p, there will exist a nonempty subspace of vectors $\mathbf{v}$ such that $\iota_{\mathbf{v}}\Omega = 0$. Since the 2-form Ω is smooth, this subspace will also vary smoothly between points. So at least one smooth vector field $\mathbf{v} \neq 0$ can be chosen such that $\iota_{\mathbf{v}}\Omega = 0$ at every point within some domain. A local coordinate system $\{x_1, ..., x_n\}$ can be chosen such that $\mathbf{v} = \partial/\partial x_n$. Then the property $\iota_{\mathbf{v}}\Omega = 0$ means that the expression for Ω may contain $\mathrm{d}x_1$, ..., $\mathrm{d}x_{n-1}$ but no $\mathrm{d}x_n$. Further, we have $\mathrm{d}\Omega = 0$ and thus (by the Cartan homotopy formula)

$$\mathcal{L}_{\mathbf{v}}\Omega = (\mathrm{d}\iota_{\mathbf{v}} + \iota_{\mathbf{v}}\mathrm{d})\,\Omega = 0.$$

The property $\mathcal{L}_{\mathbf{v}}\Omega = 0$ means that the coefficients $\Omega_{ij}(x_1, ..., x_n)$ in the decomposition (1.25) do not actually depend on the coordinate x_n (see Statement 1.3.2). Thus, it is sufficient to study the restriction of the 2-form Ω to an $(n-1)$-dimensional submanifold of constant x_n. The same construction can be applied to the reduced 2-form Ω. If the rank of Ω is smaller than $n-1$, anther coordinate x_{n-1} can be found such that Ω does not indolve $\mathrm{d}x_{n-1}$ and the coefficients Ω_{ij} do not depend on x_{n-1}. Thus we will reduce Ω to an $(n-2)$-dimensional manifold, etc. Each time, the dimension of the manifold is reduced by one, until we obtain a 2-form Ω of rank $2k$ on a $2k$-dimensional manifold, i.e. a 2-form of maximal rank. A 2-form of maximal rank is nondegenerate (Statement 1.4.7). Thus it is sufficient to prove the Darboux theorem only for nondegenerate 2-forms.

The second step is to prove that there exists a local coordinate system $\{x_1, ..., x_n\}$ such that a given closed, nondegenerate 2-form Ω has *constant* coefficients Ω_{ij} in the coordinates $\{x_j\}$,

$$\Omega = \frac{1}{2} \sum_{i,j=1}^{2n} \Omega^{(0)}_{ij} \mathrm{d}x_i \wedge \mathrm{d}x_j, \quad \Omega^{(0)}_{ij} = -\Omega^{(0)}_{ij} = \text{const}.$$

This involves a differential-geometric argument detailed below. The third step is to reduce the nondegenerate, antisymmetric $2n \times 2n$ matrix $\Omega^{(0)}_{ij}$ to the canonical form

$$\begin{pmatrix} 0 & 1 & & & \\ -1 & 0 & & 0 & \\ & & \ddots & & \\ & 0 & & 0 & 1 \\ & & & -1 & 0 \end{pmatrix}.$$

The reduction is performed through a change of basis. According to Statement 1.4.7, a basis $\{\mathbf{e}_1, ..., \mathbf{e}_{2n}\}$ can be chosen such that the matrix $\Omega^{(0)}_{ij}$ assumes the canonical form shown above (there is no zero block since $\Omega^{(0)}_{ij}$ has maximal rank).

It remains to perform the second step. We fix a point p at which

$$\Omega\big|_p = \frac{1}{2} \sum_{i,j=1}^{2n} \Omega^{(0)}_{ij} \mathrm{d}x_i \wedge \mathrm{d}x_j,$$

and define a new 2-form Ω_0 at all other points by the formula

$$\Omega_0 = \frac{1}{2} \sum_{i,j=1}^{2n} \Omega^{(0)}_{ij} \mathrm{d}x_i \wedge \mathrm{d}x_j.$$

By construction, $\Omega = \Omega_0$ at p, while Ω_0 has "constant coefficients" in the given coordinate system. Now we would like to find a change of coordinates in a neighborhood of p such that Ω has constant coefficients in the new coordinates. This is equivalent to finding a diffeomorphism $f : \mathcal{M} \to \mathcal{M}$ such that $f(p) = p$ and $f(\Omega) = \Omega_0$. Instead of finding f directly, we will find a one-parametric *flow* of diffeomorphisms f_t, $t \in [0, 1]$, such that $f_0 = \mathrm{id}$ (the identity map) and $f_1 = f$ is the map we need.

We would like to find a map f_t that transforms Ω into a 2-form Ω_t which is "between" Ω and Ω_0. The first trick is to write such an interpolation explicitly, i.e. we define

$$\Omega_t \equiv \Omega + t\,(\Omega_0 - \Omega)\,.$$

The 2-form $\Omega_0 - \Omega$ is closed and thus, by the Poincaré lemma, there exists a 1-form θ such that

$$\Omega_0 - \Omega = \mathrm{d}\theta$$

(perhaps, the above will hold only in a smaller star-shaped neighborhood of the initial point p). The second trick is to describe the flow f_t through a tangent vector field $\mathbf{v}_t$. The vector field $\mathbf{v}_t$ will be obtained by the duality map from the 1-form θ, using the 2-form Ω_t as the "metric." Converting a 1-form into a vector field is possible only if Ω_t is nondegenerate at every point and for every $t \in [0, 1]$. We postpone the proof of the nondegeneracy of Ω_t, and presently we assume that we have a vector field $\mathbf{v}_t$ satisfying

$$\iota_{\mathbf{v}_t}\Omega_t = \theta, \quad t \in [0, 1]\,.$$

We now define the flow f_t as the flow generated by the (time-dependent!) vector field $\mathbf{v}_t$. By construction, the flow f_t transforms Ω into a t-dependent 2-form $\Omega(t)$ which satisfies the differential equation

$$\frac{d}{dt}\Omega(t) = \mathcal{L}_{\mathbf{v}_t}\Omega(t).$$

It is then straightforward to verify that the interpolation Ω_t satisfies this equation:

$$\frac{d}{dt}\Omega_t = \Omega_0 - \Omega = \mathrm{d}\theta,$$
$$\mathcal{L}_{\mathbf{v}_t}\Omega_t = (\mathrm{d}\iota_{\mathbf{v}_t} + \iota_{\mathbf{v}_t}\mathrm{d})\,\Omega_t = \mathrm{d}\iota_{\mathbf{v}_t}\Omega_t = \mathrm{d}\theta.$$

Thus $\Omega_t = \Omega(t)$ and is indeed the result of the transformation of the initial 2-form Ω by the flow f_t. The diffeomorphism f_1 corresponding to $t = 1$ is the one that transforms Ω into Ω_0.

It remains to demonstrate that the 2-form Ω_t has maximal rank for every $t \in [0,1]$ and at every point within some domain. We note that Ω_t interpolates between Ω and Ω_0, while the 2-forms Ω and Ω_0 coincide at the chosen point p. Since $\Omega_t = \Omega$ at p, it is clear that Ω_t at p remains nondegenerate for all t; the present task is to show that Ω_t has maximal rank everywhere in some open domain around p. The 2-form Ω_t has maximal rank $2n$ iff $\Omega_t^{\wedge n} \neq 0$. The $2n$-form $\Omega_t^{\wedge n}$ is proportional to the volume form $\mathrm{d}x_1 \wedge ... \wedge \mathrm{d}x_{2n}$, namely

$$\Omega_t^{\wedge n} = f_t(x)\, \mathrm{d}x_1 \wedge ... \wedge \mathrm{d}x_{2n},$$

where the coefficient, temporarily denoted by f_t, is a function of the coordinates $\{x_j\}$. The rank of Ω_t is maximal if $f_t(x) \neq 0$. We know that $f_t(p) \neq 0$ for every $t \in [0,1]$. Since the function f_t is smooth, the domain where $f_t \neq 0$ is a t-dependent open set containing the point p. The intersection of all such domains for every $t \in [0,1]$ is again a nonempty open set (possibly smaller than the set where Ω has maximal rank). Thus, we have found an open domain containing the initial point p where Ω_t is nondegenerate at every t.

Canonical decomposition of 1-forms. The concepts of rank and canonical coordinates can be defined also for 1-forms. Heuristically, the rank of a 1-form is the smallest number of independent coordinates required to express this 1-form. This time, we do not limit ourselves to considering only closed 1-forms; but let us briefly examine the case of a closed 1-form. By the Poincaré lemma, a closed 1-form ω is locally represented as $\omega = \mathrm{d}f$ using some function f. The function f will be nonconstant if $\omega \neq 0$. Thus, f can be used as one of the coordinates in a local coordinate system; so ω can be expressed using just one coordinate. We conclude that a closed, nonzero 1-form has rank 1.

Below we will prove that any 1-form ω can be expressed (using suitable local coordinates) in one of two canonical ways: either

$$\omega = x_1 \mathrm{d}x_2 + ... + x_{2k-1}\mathrm{d}x_{2k} \tag{1.29}$$

(then we say that ω has rank $2k$), or

$$\omega = x_1 \mathrm{d}x_2 + ... + x_{2k-1}\mathrm{d}x_{2k} + \mathrm{d}x_{2k+1} \tag{1.30}$$

(then ω has rank $2k+1$). We may call the coordinates used for such decompositions **canonical coordinates** of a given 1-form.

Knowing that this statement is true, how could we determine the rank of a given 1-form ω, for example,

$$\omega = \mathrm{d}x + xy\mathrm{d}z,$$

without knowing the canonical coordinates? We use the following trick. Consider the 2-form $\mathrm{d}\omega$ and compute its exterior powers, $(\mathrm{d}\omega)^{\wedge k}$, $k = 1, 2, ...$ Write the following sequence of the differential forms,

$$\omega,\ \mathrm{d}\omega,\ \omega \wedge \mathrm{d}\omega,\ \mathrm{d}\omega \wedge \mathrm{d}\omega,\ \omega \wedge \mathrm{d}\omega \wedge \mathrm{d}\omega,\ ... \tag{1.31}$$

The k-th element of this sequence ($k = 1, 2, ...$) is a k-form. Eventually some element of this sequence will be zero, and then all the subsequent elements will also be zero. By substituting the decompositions (1.29) and (1.30) into the sequence (1.31), one finds that the rank of ω is equal to the number of initial nonzero elements in the sequence. The rank of ω is zero if $\omega = 0$; the rank of ω is one if $\omega \neq 0$ but $\mathrm{d}\omega = 0$; the rank of ω is two if $\mathrm{d}\omega \neq 0$ but $\omega \wedge \mathrm{d}\omega = 0$; etc. In this way it is straightforward to compute the rank of a 1-form given in some local coordinates. For example, the rank of $\omega = \mathrm{d}x + xy\mathrm{d}z$ is 3 (at points where $xy \neq 0$) because

$$\omega \wedge \mathrm{d}\omega = x\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z \neq 0, \quad \mathrm{d}\omega \wedge \mathrm{d}\omega = 0.$$

Remark: The Carathéodory theorem is a form of Frobenius theorem. It states that the set of null curves of a 1-form ω is surface-forming iff the form ω has rank 1 or 2, so that $\omega \wedge \mathrm{d}\omega = 0$. (**Null curves** of ω are curves γ such that $\omega \circ \dot{\gamma} = 0$.) This is proved by an explicit construction in local coordinates, showing that a 1-form of rank 3 or higher has non-surface-forming null curves. (A **non-surface-forming** set of curves is a set such that curves from the set can reach any point in a neighborhood of an initial point. See [2], chapter 22.) ■

Proof of the canonical decomposition theorem for 1-forms. The canonical decomposition theorem for 1-forms states that for any 1-form ω, one can choose a local coordinate system $\{x_j\}$ such that ω is locally represented in one of the two ways, (1.29) or (1.30), as long as ω has a locally constant rank.

We prove this theorem by reducing it to the Darboux theorem. First, we determine the rank of a given 1-form ω by using the sequence (1.31). If ω has an odd rank $2k+1$, the 2-form $\mathrm{d}\omega$ has rank $2k$ (according to the definition of rank for closed 2-forms) because $(\mathrm{d}\omega)^{\wedge k} \neq 0$ while $(\mathrm{d}\omega)^{\wedge(k+1)} = 0$. According to the Darboux theorem, we can then choose local coordinates $\{x_j\}$ such that

$$\mathrm{d}\omega = \mathrm{d}x_1 \wedge \mathrm{d}x_2 + ... + \mathrm{d}x_{2k-1} \wedge \mathrm{d}x_{2k}.$$

Applying the Poincaré lemma to the closed 1-form

$$\omega - x_1\mathrm{d}x_2 + ... + x_{2k-1}\mathrm{d}x_{2k},$$

we show that there exists a function f such that

$$\omega = \mathrm{d}f + x_1\mathrm{d}x_2 + ... + x_{2k-1}\mathrm{d}x_{2k}.$$

Since the 1-form ω has rank $2k+1$, we obtain

$$0 \neq \omega \wedge (\mathrm{d}\omega)^{\wedge k} = \mathrm{d}f \wedge k!\, \mathrm{d}x_1 \wedge ... \wedge \mathrm{d}x_{2k}.$$

Hence, the function f can be used as a local coordinate that is independent of $\{x_1, ..., x_{2k}\}$. Denote this function f by x_{2k+1}, we obtain the decomposition (1.30).

In the second case, the 1-form ω has an even rank $2k$. Then the 2-form $\mathrm{d}\omega$ again has rank $2k$, and so we could again arrive at Eq. (1.30), but we would like to eliminate the last term in that equation. So we need to work a little harder.

The idea is to find a function f such that the 1-form $f\omega$ has rank $2k-1$. If such f is found, then by the previously proved case it will follow that

$$f\omega = x_1 \wedge \mathrm{d}x_2 + ... + x_{2k-3} \wedge \mathrm{d}x_{2k-2} + \mathrm{d}x_{2k-1},$$

so we will obtain

$$\begin{aligned}\omega &= \frac{1}{f}\left(x_1\mathrm{d}x_2 + ... + x_{2k-3}\mathrm{d}x_{2k-2} + \mathrm{d}x_{2k-1}\right) \\ &= y_1\mathrm{d}y_2 + ... + y_{2k-3}\mathrm{d}y_{2k-2} + y_{2k-1}\mathrm{d}y_{2k},\end{aligned} \tag{1.32}$$

where we simply relabeled the coordinates as

$$y_1 \equiv \frac{1}{f}x_1,\ y_2 \equiv x_2, ..., y_{2k-1} \equiv \frac{1}{f},\ y_{2k} \equiv x_{2k-1}. \tag{1.33}$$

By assumption, the 1-form ω has rank $2k$ and thus

$$(\mathrm{d}\omega)^{\wedge k} \neq 0, \quad \omega \wedge (\mathrm{d}\omega)^{\wedge k} = 0.$$

Since $(\mathrm{d}\omega)^{\wedge k} \neq 0$, it will follow from Eq. (1.32) that all $\{y_1, ..., y_{2k}\}$ are independent local coordinates, and thus a decomposition of the form (1.29) will be found.

It remains to determine f such that the 1-form $f\omega$ has rank $2k-1$. It is sufficient to find some function f such that

$$\begin{aligned} 0 &= (\mathrm{d}(f\omega))^{\wedge k} = (\mathrm{d}f \wedge \omega + f\mathrm{d}\omega)^{\wedge k} \\ &= kf^{k-1}\mathrm{d}f \wedge \omega \wedge (\mathrm{d}\omega)^{\wedge(k-1)} + f^k (\mathrm{d}\omega)^{\wedge k}. \end{aligned}$$

Thus the function f must satisfy the differential equation

$$(\mathrm{d}\ln f) \wedge \omega \wedge (\mathrm{d}\omega)^{\wedge(k-1)} = -\frac{1}{k}(\mathrm{d}\omega)^{\wedge k}. \tag{1.34}$$

To solve this equation for f (or rather, to show that a solution exists), we need to use some additional information about ω. We know that the 2-form $\mathrm{d}\omega$ has rank $2k$. Therefore, by the Darboux theorem we may choose local coordinates $\{x_1, ..., x_n\}$ such that

$$\mathrm{d}\omega = \mathrm{d}x_1 \wedge \mathrm{d}x_2 + ... + \mathrm{d}x_{2k-1} \wedge \mathrm{d}x_{2k}.$$

Applying the Poincaré lemma to the closed 1-form

$$\omega - x_1\mathrm{d}x_2 + ... + x_{2k-1}\mathrm{d}x_{2k},$$

we show that there exists a function h such that

$$\omega = \mathrm{d}h + x_1\mathrm{d}x_2 + ... + x_{2k-1}\mathrm{d}x_{2k}.$$

Since $\mathrm{d}\omega$ has rank $2k$, we have

$$0 = (\mathrm{d}\omega)^{\wedge(k+1)} = (k+1)!\,\mathrm{d}h \wedge \mathrm{d}x_1 \wedge ... \wedge \mathrm{d}x_{2k},$$

so the function h can depend only on the subset of coordinates consisting of the first $2k$ coordinates $\{x_1, ..., x_{2k}\}$. Hence the 1-form ω depends also only on these coordinates. Let us write

$$\omega = \sum_{j=1}^{2k} a_j \mathrm{d}x_j,$$

where a_j are coefficients that depend only on $\{x_1, ..., x_{2k}\}$. Since both parts of Eq. (1.34) depend only on these coordinates, the function f is a function only of $\{x_1, ..., x_{2k}\}$.

We can now return to the task of determining a suitable function f. First we express the $(2k-1)$-form $\omega \wedge (\mathrm{d}\omega)^{\wedge(k-1)}$ needed for Eq. (1.34) as

$$\begin{aligned} &\omega \wedge (\mathrm{d}\omega)^{\wedge(k-1)} \\ &= (k-1)!\,(a_1\mathrm{d}x_1 + a_2\mathrm{d}x_2) \wedge \mathrm{d}x_3 \wedge ... \wedge \mathrm{d}x_{2k} + ... \\ &+ (k-1)!\,(a_{2k-1}\mathrm{d}x_{2k-1} + a_{2k}\mathrm{d}x_{2k}) \wedge \mathrm{d}x_1 \wedge ... \wedge \mathrm{d}x_{2k-2}. \end{aligned}$$

The unknown 1-form $\mathrm{d}\ln f$ can then be written as

$$\mathrm{d}\ln f \equiv \sum_{j=1}^{2k} l_j \mathrm{d}x_j,$$

where l_j are unknown coefficients depending on $\{x_1, ..., x_{2k}\}$. Using this explicit representation, we compute

$$\begin{aligned} (\mathrm{d}\ln f) \wedge \omega \wedge (\mathrm{d}\omega)^{\wedge(k-1)} &= (k-1)!\,(a_1 l_2 - a_2 l_1 + ... \\ &+ a_{2k-1}l_{2k} - a_{2k}l_{2k-1})\mathrm{d}x_1 \wedge ...\mathrm{d}x_{2k} \end{aligned}$$

and hence is proportional to the $2k$-dimensional volume form

$$(\mathrm{d}\omega)^{\wedge k} = k!\,\mathrm{d}x_1 \wedge ...\mathrm{d}x_{2k}.$$

It is convenient to introduce an auxiliary vector field

$$\mathbf{v} \equiv -a_2\frac{\partial}{\partial x_1} + a_1\frac{\partial}{\partial x_2} - ... - a_{2k}\frac{\partial}{\partial x_{2k-1}} + a_{2k-1}\frac{\partial}{\partial x_{2k}}$$

and write

$$(\mathrm{d}\ln f) \wedge \omega \wedge (\mathrm{d}\omega)^{\wedge(k-1)} = (k-1)!\,(\iota_{\mathbf{v}}\mathrm{d}\ln f)\,\mathrm{d}x_1 \wedge ...\mathrm{d}x_{2k}. \tag{1.35}$$

Since both sides of Eq. (1.34) are proportional to the volume form, the equation is simplified to

$$\iota_{\mathbf{v}}\mathrm{d}\ln f = \mathbf{v} \circ \ln f = -1.$$

Note that the field $\mathbf{v}$ can be also defined without coordinates by the requirement

$$\iota_{\mathbf{v}}\left[\omega \wedge (\mathrm{d}\omega)^{\wedge(k-1)}\right] = 0.$$

In other words, $\mathbf{v}$ is a vector field that annihilates the nonzero $(2k-1)$-form $\omega \wedge (\mathrm{d}\omega)^{\wedge(k-1)}$. (In general, a k-form vanishes on an $(n-k)$-dimensional hypersurface, so a $(2k-1)$-form vanishes on a 1-dimensional subspace. The vector $\mathbf{v}$ is a basis vector in that subspace.)

We found that only restriction on the function f is that its logarithmic derivative in the direction of $\mathbf{v}$ must equal -1. Now it is clear that a solution f exists (at least locally). A particular solution f can be determined by integration along the orbits of $\mathbf{v}$ from arbitrary initial conditions. This completes the proof.

1.4.8 The Poincaré lemma

The Poincaré lemma states that a closed n-form ω is locally exact: there exists an $(n-1)$-form θ such that $\omega = \mathrm{d}\theta$ in a star-shaped neighborhood of some point.

A domain $\mathcal{D}$ is a **star-shaped neighborhood** of a point p if there exists a diffeomorphism flow f_t that contracts the domain $\mathcal{D}$ into the point p. In other words, f_t is a continuous set of diffeomorphisms, defined for $t \in [0, 1]$ and such that $f_0 = \mathrm{id}$ and f_1 maps the entire neighborhood $\mathcal{D}$ into the single point p.

Suppose that $\mathcal{D}$ is a star-shaped neighborhood of p and a suitable flow f_t is given. Let $\mathbf{v}_t$ be the tangent vector field of the flow f_t. Suppose that ω is an exact n-form defined in $\mathcal{D}$. The flow f_t transforms ω into a t-dependent n-form ω_t that satisfies $\omega_0 = \omega$, $\omega_1 = 0$, and

$$\frac{d}{dt}\omega_t = \mathcal{L}_{\mathbf{v}_t}\omega_t.$$

Since $\mathrm{d}\omega = 0$, we also have $\mathrm{d}\omega_t = 0$ since d commutes with diffeomorphisms. Using the Cartan homotopy formula, we find

$$\frac{d}{dt}\omega_t = (\mathrm{d}\iota_{\mathbf{v}_t} + \iota_{\mathbf{v}_t}\mathrm{d})\,\omega = \mathrm{d}\iota_{\mathbf{v}_t}\omega.$$

Now we integrate this relationship over t,

$$-\omega = \omega_1 - \omega_0 = \int_0^1 \left(\frac{d}{dt}\omega_t\right) dt = \mathrm{d}\int_0^1 (\iota_{\mathbf{v}_t}\omega)\,dt.$$

Defining

$$\theta \equiv -\int_0^1 (\iota_{\mathbf{v}_t}\omega)\,dt,$$

we find the required relationship, $\omega = \mathrm{d}\theta$.

1.4.9 Integration of forms

A differential n-forms can be **integrated** over a manifold of dimension n: namely, 1-forms can be integrated over curves, 2-forms over surfaces, etc. The integral of a given n-form ω over a given manifold $\mathcal{A}$ of dimension n is denoted by

$$\int_{\mathcal{A}} \omega.$$

Note that the differential "d" is not written after the integral sign because the n-form ω already contains the correct number of d's.

The integral $\int_{\mathcal{A}} \omega$ is defined by the following procedure. One starts by splitting the manifold $\mathcal{A}$ into a large number N of small n-dimensional paralellepipeds. If all the parallelepipeds are sufficiently small, one can approximate the sides of parallelepipeds by tangent vectors (see Sec. 1.2.5). Thus, to each parallelepiped numbered j (where $j = 1, ..., N$) there corresponds a set of n tangent vectors $\{\mathbf{v}_1^{(j)}, ..., \mathbf{v}_n^{(j)}\}$. The parallelepiped numbered j contributes a small amount to the integral; this amount is equal to some number (the value of the function being integrated) times the n-dimensional volume of the parallelepiped. It is natural to regard the n-volume of a parallelepiped spanned by $\{\mathbf{v}_1, ..., \mathbf{v}_n\}$ as an n-form evaluated on the vectors $\{\mathbf{v}_i\}$ (see Sec. 1.4.1). So the contribution of the j-th parallelepiped to the integral can be naturally described as an application of some n-form to the n vectors $\{\mathbf{v}_1^{(j)}, ..., \mathbf{v}_n^{(j)}\}$. In the integral we are evaluating, this n-form is the given n-form ω. Hence, to define the integral $\int_{\mathcal{A}} \omega$ we evaluate the contribution $\omega \circ (\mathbf{v}_1^{(j)}, ..., \mathbf{v}_n^{(j)})$ of each parallelepiped j and then compute the sum of the contributions over all the parallelepipeds. Thus, the integral $\int_{\mathcal{A}} \omega$ is defined as the limit $N \to \infty$ of that sum,

$$\int_{\mathcal{A}} \omega \equiv \lim_{N\to\infty} \sum_{j=1}^{N} \omega \circ (\mathbf{v}_1^{(j)}, ..., \mathbf{v}_n^{(j)}).$$

This limit is quite similar to the limit involved in the ordinary definition of the Riemann integral. So integrals of n-forms are equivalent to n-fold integrals in the usual sense. However, an important caveat is that integrals of n-forms are always performed over *oriented* manifolds. Since n-forms are antisymmetric, an integral will change sign when the orientation of the manifold is reversed.

The fundamental relationship between integration and the exterior differential is

$$\int_{\mathcal{A}} \mathrm{d}\omega = \int_{\partial\mathcal{A}} \omega,$$

where $\mathcal{A}$ is an n-dimensional manifold (or part of a manifold), $\partial\mathcal{A}$ is its $(n-1)$-dimensional boundary, and ω is an arbitrary $(n-1)$-form. This theorem is a generalization of the Stokes and the Gauss laws, as well as the fundamental theorem of calculus $\int_a^b f'(x)dx = f(b) - f(a)$. Since this is a standard result, we refer the reader to textbooks (such as [1]) for more detailed explanations of this theorem.

1.5 Metric

1.5.1 Motivation: metric on surfaces

I will introduce the concept of a metric by starting from the picture of a manifold as a surface embedded in a Euclidean space $\mathbb{R}^n$. Each point p of $\mathbb{R}^n$ can be represented by an n-dimensional vector $\mathbf{p}$. The natural Pythagorean notion of distance in $\mathbb{R}^n$, namely the **Euclidean metric**, defines the distance between two points $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ as

$$\text{Distance}\,(\mathbf{a}, \mathbf{b}) = \sqrt{(b_1 - a_1)^2 + ... + (b_n - a_n)^2}, \qquad (1.36)$$

where a_n and b_n are standard coordinates (components) of the vectors $\mathbf{a}, \mathbf{b}$. Let us now consider a surface $\mathcal{M}$ embedded into $\mathbb{R}^n$. We would like to compute the distance between two points $\mathbf{a}, \mathbf{b} \in \mathcal{M}$, as measured along some path *within the surface*. However, if the surface has a curved shape then one does not have a simple and general formula analogous to Eq. (1.36) for distances measured *within the surface*. Nevertheless, if two points $\mathbf{a}, \mathbf{b}$ on the surface $\mathcal{M}$ are very close to each other, we may consider a short, almost straight curve segment within $\mathcal{M}$ connecting these points. The length of this curve segment can be accurately estimated using the Euclidean formula (1.36). Then the length of any curve within the surface can be determined by splitting the curve into sufficiently small segments. More precisely, one can integrate the *infinitesimal* lengths of infinitesimal segments along the curve.

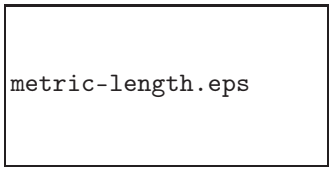

Figure 1.11: A surface is embedded into a Euclidean space $\mathbb{R}^3$. A short curve segment connecting two points $\mathbf{a}$ and $\mathbf{b}$ is approximated by a straight line segment. The length of this straight line segment is computed using the Euclidean metric in $\mathbb{R}^3$.

Let us explore this idea in more detail. To be specific, let us consider a two-dimensional manifold $\mathcal{M}$ embedded in $\mathbb{R}^3$ as a surface $z = F(x, y)$ in standard coordinates $\{x, y, z\}$, as shown in Fig. 1.11. We are interested in computing the length of a curve $\gamma(\tau) \subset \mathcal{M}$. The curve is specified by three functions $\{x(\tau), y(\tau), z(\tau)\}$. Consider a very short curve segment between $\tau = \tau_0$ and $\tau = \tau_0 + \delta\tau$, where $\delta\tau$ is very small. This segment is an almost straight line between the nearby points $\{x_0, y_0, z_0\} \equiv \{x(\tau_0), y(\tau_0), z(\tau_0)\}$ and $\{x_1, y_1, z_1\} \equiv \{x(\tau_0 + \delta\tau), y(\tau_0 + \delta\tau), z(\tau_0 + \delta\tau)\}$ in $\mathbb{R}^3$. So we may compute the length δL of the segment approximately as the Euclidean distance between these points,

$$\delta L \approx \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + (z_1 - z_0)^2}.$$

Further, we may approximate

$$x_1 - x_0 = x(\tau_0 + \delta\tau) - x(\tau_0) \approx \delta\tau \frac{dx}{d\tau}(\tau_0), \quad \text{etc.},$$

and hence

$$\delta L \approx \delta\tau \sqrt{\left(\frac{dx}{d\tau}(\tau_0)\right)^2 + \left(\frac{dy}{d\tau}(\tau_0)\right)^2 + \left(\frac{dz}{d\tau}(\tau_0)\right)^2}.$$

In the limit $\delta\tau \to 0$, this expression becomes precise since the error is of order $\delta\tau^2$. So the length of the curve $\gamma(\tau)$ between $\tau = \tau_0$ and $\tau = \tau_1$ can be found as an integral over the infinitesimal curve segments,

$$L[\gamma] = \int_{\tau_0}^{\tau_1} d\tau \sqrt{\left(\frac{dx}{d\tau}(\tau)\right)^2 + \left(\frac{dy}{d\tau}(\tau)\right)^2 + \left(\frac{dz}{d\tau}(\tau)\right)^2}.$$

So far we have used all three coordinates x, y, z to describe the curve. However, the pair $\{x, y\}$ can be used as local coordinates on $\mathcal{M}$, and so we would like to be able to compute lengths of curves without using the coordinate z. In other words, we would like to use only an intrinsic description of the manifold $\mathcal{M}$. The "metric" is the structure representing the information necessary for computing lengths in the intrinsic description. Let us now see what that information is in our example.

Given the equation $z = F(x, y)$, we may express

$$\frac{dz(\tau)}{d\tau} = \frac{\partial F(x,y)}{\partial x}\frac{dx(\tau)}{d\tau} + \frac{\partial F(x,y)}{\partial y}\frac{dy(\tau)}{d\tau},$$

where it is implied that the derivatives of F are evaluated on the curve $\gamma(\tau)$. Hence, the length δL of an infinitesimal curve segment can be written as

$$\begin{aligned}\delta L &= \delta\tau\sqrt{\left(\frac{dx}{d\tau}\right)^2 + \left(\frac{dy}{d\tau}\right)^2 + \left(\frac{\partial F}{\partial x}\frac{dx}{d\tau} + \frac{\partial F}{\partial y}\frac{dy}{d\tau}\right)^2} \\ &= \delta\tau\sqrt{A\left(\frac{dx}{d\tau}\right)^2 + 2B\frac{dx}{d\tau}\frac{dy}{d\tau} + C\left(\frac{dy}{d\tau}\right)^2}, \qquad (1.37)\end{aligned}$$

where we introduced auxiliary functions A, B, C expressed directly through F as

$$A(x,y) \equiv 1 + \left(\frac{\partial F}{\partial x}\right)^2, \quad B \equiv \frac{\partial F}{\partial x}\frac{\partial F}{\partial y}, \quad C \equiv 1 + \left(\frac{\partial F}{\partial y}\right)^2.$$

Now we note that Eq. (1.37) looks like a quadratic form applied to the two-dimensional vector

$$\dot{\gamma} \equiv \left\{\frac{dx}{d\tau}, \frac{dy}{d\tau}\right\} \in T_{\gamma(\tau)}\mathcal{M}.$$

Recalling the correspondence between tangent vectors from $T_p\mathcal{M}$ and short curve segments (see Sec. 1.2.5), we can verify that the tangent vector $\delta\mathbf{v}$ defined by

$$\delta\mathbf{v} \equiv (\delta\tau)\,\dot{\gamma} = (\delta\tau)\left[\frac{dx}{d\tau}\partial_x + \frac{dy}{d\tau}\partial_y\right] \in T_{\gamma(\tau_0)}\mathcal{M}$$

is the vector representing the short curve segment between $\gamma(\tau_0)$ and $\gamma(\tau_0 + \delta\tau)$. Then δL can be expressed as

$$\delta L = \delta\tau\sqrt{g(\dot{\gamma}, \dot{\gamma})} = \sqrt{g(\delta\mathbf{v}, \delta\mathbf{v})},$$

where g is the following bilinear form,

$$g \equiv A\mathrm{d}x \otimes \mathrm{d}x + B(\mathrm{d}x \otimes \mathrm{d}y + \mathrm{d}y \otimes \mathrm{d}x) + C\mathrm{d}y \otimes \mathrm{d}y.$$

Thus δL (the length of a short curve segment) can be expressed through the bilinear form g alone, without using any other information (such as the function $F(x, y)$ or the embedding of $\mathcal{M}$ in the Euclidean space $\mathbb{R}^3$). The length of the curve between τ_0 and τ_1 can be expressed as

$$L[\gamma] = \int_{\tau_0}^{\tau_1} d\tau\sqrt{g(\dot{\gamma}, \dot{\gamma})}.$$

We conclude that the only information necessary to compute lengths along curves within a surface $\mathcal{M}$ is a symmetric bilinear form g defined in every tangent space $T_p\mathcal{M}$. This bilinear form g is called the **metric** on the manifold $\mathcal{M}$.

In our example, once we computed the metric g (which, in a local coordinate system, is equivalent to determining the three functions A, B, C), we may stop using the embedding of $\mathcal{M}$ in $\mathbb{R}^3$ because it is sufficient to use the local coordinate system $\{x, y\}$. This simplifies the calculations since one needs to work with fewer coordinates, and also allows one to concentrate on the intrinsic properties of the manifold $\mathcal{M}$. Below we will not use the embedding picture to the metric; instead, we will work directly with the metric g, assuming that somehow g is given. It will be unimportant whether or not the metric g comes from an embedding of $\mathcal{M}$ in a larger Euclidean space. An embedding of $\mathcal{M}$ into $\mathbb{R}^n$ can be regarded as merely an auxiliary construction that helps visualize the concept of metric.

In practice, it is sometimes useful to specify the metric on a manifold $\mathcal{M}$ by selecting a particular embedding of $\mathcal{M}$ into a Euclidean space $\mathbb{R}^n$ and declaring that the metric g on $\mathcal{M}$ is the natural metric induced by this embedding, through the construction of infinitesimal curve segments outlined above. The metric g is then called the **induced metric** on $\mathcal{M}$ with respect to the given embedding. In most cases, it is easier to specify a metric g through components in in a local coordinate system. In the example above, this would mean that we specify the local coordinates as $\{x, y\}$ and declare that the coefficients A, B, C of the metric are some given functions of x, y.

1.5.2 Definition

After this motivation, let us now turn to the formal definition of a metric.

A **metric** on a manifold $\mathcal{M}$ is a nondegenerate, symmetric bilinear form $g(\mathbf{u}, \mathbf{v})$ defined in the tangent space $T_p\mathcal{M}$ at each point $p \in \mathcal{M}$. A symmetric bilinear form g is called **nondegenerate** if for any vector $\mathbf{u} \neq 0$ there exists a vector $\mathbf{v}$ such that $g(\mathbf{u}, \mathbf{v}) \neq 0$; in other words, no vector can be orthogonal to *every* other vector.

The nondegeneracy condition prohibits "uninteresting" metrics, such as $g = 0$. It is also important to note that the nondegeneracy condition permits a vector $\mathbf{x}$ to be specified uniquely if its scalar products $g(\mathbf{v}, \mathbf{x})$ with every other vector $\mathbf{v}$ are known. (If there were two vectors $\mathbf{x} \neq \mathbf{x}'$ such that $g(\mathbf{v}, \mathbf{x}) = g(\mathbf{v}, \mathbf{x}')$ for all $\mathbf{v}$, then the vector $\mathbf{x} - \mathbf{x}' \neq 0$ would be orthogonal to every vector, but this is excluded by the assumption of nondegeneracy of g.)

It is known from standard linear algebra that symmetric bilinear forms have **signature**, which is a set of signs independent of the choice of a basis. In GR we shall usually consider four-dimensional metrics with the signature $(+---)$ as appropriate for a locally Lorentzian physical theory, but many results will be the same for arbitrary dimension and signatures of the metric. Manifolds with a metric with signature $(++...+)$ are called **Riemannian**, and **pseudo-Riemannian** if the metric is not sign-definite. The familiar metric with a Lorentzian signature is the **Minkowski metric**, $\eta_{\mu\nu} = \mathrm{diag}(1, -1, -1, -1)$.

Self-test question: The Minkowski metric η admits **null** vectors $\mathbf{n}$ such that $\eta(\mathbf{n}, \mathbf{n}) = 0$; for example, $\mathbf{n} = \{1, 1, 0, 0\}$ is a null vector. Does this mean that the metric η is degenerate? *Answer:* No. ■

In GR, the metric describes physically measured lengths and time intervals. We focus on the mathematical properties of the metric for now.

1.5.3 Examples of metrics

As a first example, consider a flat Minkowski space $\mathcal{M} \equiv \mathbb{R}^4$ with standard coordinates $\{t, x, y, z\}$. The Minkowski metric g can be specified by the following index-free expression,

$$g(\mathbf{u}, \mathbf{v}) = (\mathbf{u} \circ t)\,(\mathbf{v} \circ t) - (\mathbf{u} \circ x)\,(\mathbf{v} \circ x) \\ - (\mathbf{u} \circ y)\,(\mathbf{v} \circ y) - (\mathbf{u} \circ z)\,(\mathbf{v} \circ z)\,.$$

In this expression, the coordinates $\{t, x, y, z\}$ are interpreted as scalar functions on the manifold $\mathcal{M}$, and the vector fields $\mathbf{u}, \mathbf{v}$ are interpreted as derivative operations applied to these scalar functions. We can then directly compute

$$g(\frac{\partial}{\partial t}, \frac{\partial}{\partial t}) = 1, \quad g(\frac{\partial}{\partial t}, \frac{\partial}{\partial x}) = 0, \quad g(\frac{\partial}{\partial t} + \frac{\partial}{\partial x}, \frac{\partial}{\partial t} + \frac{\partial}{\partial x}) = 0,$$

etc. This metric can be also written as

$$g = \mathrm{d}t \otimes \mathrm{d}t - \mathrm{d}x \otimes \mathrm{d}x - \mathrm{d}y \otimes \mathrm{d}y - \mathrm{d}z \otimes \mathrm{d}z,$$

where $\mathrm{d}t, \mathrm{d}x, \mathrm{d}y, \mathrm{d}z$ are the 1-forms defined through the coordinate functions t, x, y, z.

Remark on notation: In the physics literature, one usually finds the Minkowski metric written as follows,

$$ds^2 = dt^2 - dx^2 - dy^2 - dz^2.$$

Strictly speaking, this notation is inconsistent; for instance, dt^2 stands for the bilinear form $\mathrm{d}t \otimes \mathrm{d}t$, and yet "$ds^2$" cannot be interpreted as $\mathrm{d}s \otimes \mathrm{d}s$ with any 1-form $\mathrm{d}s$. Expressions of this kind are understood as being no more than a traditional (or "jargon") notation for the metric, in which "ds^2" is simply a symbol for the bilinear form g. Also, physicists write "$dt\,dx$" when they need the symmetric tensor product $\frac{1}{2}(\mathrm{d}t \otimes \mathrm{d}x + \mathrm{d}x \otimes \mathrm{d}t)$. I will frequently use the "physicists's" metric notation for brevity. To call attention to the traditional character of this notation, I will denote the metric by g instead of the "ds^2" and I will write dx^2 and $dt\,dx$ instead of the rigorous but cumbersome notation $\mathrm{d}x \otimes \mathrm{d}x$, $\frac{1}{2}(\mathrm{d}t \otimes \mathrm{d}x + \mathrm{d}x \otimes \mathrm{d}t)$ that does not yield any computational advantage. In those cases, I write "dx" with an italic "d" because there is no connection to the exterior differential "d." ■

More complicated metrics may involve nontrivial coefficients at the coordinate 1-forms and/or non-diagonal terms, e.g.

$$g = f_1(t, x, y, z)dt^2 + f_2(t, x, y, z)dt\,dx + ...$$

Here are some further examples of metrics describing spacetimes important in physics.

A **Schwarzschild spacetime** in the coordinates $\{t, r, \theta, \phi\}$ is described by the metric

$$g = \left(1 - \frac{2M}{r}\right) dt^2 - \left(1 - \frac{2M}{r}\right)^{-1} dr^2 - r^2\left(d\theta^2 + (\sin^2\theta)d\phi^2\right) \tag{1.38}$$

This spacetime is generated by a nonrotating black hole of mass M centered at $r = 0$.

A **de Sitter spacetime** in spatially flat coordinates $\{t, x, y, z\}$ has the metric

$$g = dt^2 - e^{2Ht}(dx^2 + dy^2 + dz^2). \tag{1.39}$$

This spacetime is used to describe (approximately) the universe that is undergoing an accelerated expansion (**cosmological inflation**).

Remark: In GR, a physical spacetime $\mathcal{M}$ is a four-dimensional manifold with a metric g; the metric should have the Lorentzian signature $(+---)$. A simple-minded view of a curved spacetime $\mathcal{M}$ is that $\mathcal{M} = \mathbb{R}^4$ with coordinates $\{x^\mu\} \equiv \{t, x, y, z\}$, and a metric g is specified through the components $g_{\alpha\beta}(t, x, y, z)$. However, this picture is insufficiently general; for instance, spacetimes containing black holes do not have this simple structure. In some cases, it turns out that the full manifold $\mathcal{M}$ is not covered by the coordinate system $\{x^\mu\}$ in which a metric $g_{\alpha\beta}$ was originally specified. ■

Practice problem: Compute the scalar product $g(\mathbf{a}, \mathbf{b})$ of vectors $\mathbf{a} = t\partial_t + x\partial_x$ and $\mathbf{b} = t\partial_x - x\partial_t$ in the metric $g = dt^2 - dx^2$. Give an example of a vector orthogonal to $\mathbf{a}$ in this metric. ■

1.5.4 Orthonormal frames

This section is a very brief introduction to the notion of orthonormal frames. This subject is more extensively developed in Sec. 6.1.1.

In each tangent space $T_p\mathcal{M}$, one can choose an orthonormal basis $\{\mathbf{e}_a\}$, where a is a label enumerating the basis vectors. We thus obtain a set of vector fields.[7] Such a basis of vector fields is called an **orthonormal frame** (in four dimensions, a **tetrad** or a **vierbein**). Orthonormality means that

$$g(\mathbf{e}_a, \mathbf{e}_b) = \eta_{ab},$$

where η_{ab} is a diagonal matrix having diagonal elements equal to ± 1, depending on the signature of the metric g. For instance, in GR one uses metrics with Lorentzian signature $(+---)$, so η_{ab} is the matrix

$$\eta_{ab} = \mathrm{diag}(1, -1, -1, -1) = \begin{pmatrix} 1 & & & \\ & -1 & & \\ & & -1 & \\ & & & -1 \end{pmatrix}.$$

The dual basis consists of 1-forms $\{\boldsymbol{\theta}^a\}$ that can be expressed as follows,

$$\boldsymbol{\theta}^a \circ \mathbf{u} \equiv \eta_{aa} g(\mathbf{e}_a, \mathbf{u}).$$

The 1-forms $\{\boldsymbol{\theta}^a\}$ are a basis in the cotangent space $T_p^*\mathcal{M}$. The metric g, as a rank (0,2) tensor, can be recovered from the basis $\{\boldsymbol{\theta}^a\}$ as

$$g = \sum_{a,b} \eta_{ab} \boldsymbol{\theta}^a \otimes \boldsymbol{\theta}^b; \quad g^{-1} = \sum_{a,b} \eta_{ab} \mathbf{e}_a \otimes \mathbf{e}_b. \tag{1.40}$$

A derivation of Eq. (1.40) can be found in Sec. 6.1.1.

1.5.5 Correspondence of vectors and covectors

A metric g determines a one-to-one map between vectors and covectors (and thus between vector fields and 1-forms). I denote this map by $\hat{g}$, so $\hat{g}\mathbf{v}$ is the 1-form corresponding to a vector field $\mathbf{v}$. By definition, the 1-form $\hat{g}\mathbf{v}$ acts on vectors $\mathbf{x}$ as follows,

$$(\hat{g}\mathbf{v}) \circ \mathbf{x} \equiv g(\mathbf{v}, \mathbf{x}).$$

Then the dual basis 1-forms $\{\boldsymbol{\theta}^a\}$ are expressed as $\boldsymbol{\theta}^a = \eta_{aa}\hat{g}\mathbf{e}_a$.

[7] If it turns out that an orthonormal frame cannot be chosen smoothly throughout the entire manifold $\mathcal{M}$, we assume that the vector fields $\{\mathbf{e}_a\}$ are smooth within some suitable patch of $\mathcal{M}$.

Since the correspondence between vectors and 1-forms is one-to-one, there exists the inverse map (from 1-forms to vectors), denoted $\hat{g}^{-1}$. This map converts a 1-form ω into the vector $\mathbf{v} = \hat{g}^{-1}\omega$ such that

$$g(\mathbf{v}, \mathbf{u}) \equiv g(\hat{g}^{-1}\omega, \mathbf{u}) \equiv \omega \circ \mathbf{u}$$

for any vector $\mathbf{u}$. Then the scalar product is also defined on 1-forms and is denoted by g^{-1}:

$$g^{-1}(\omega_1, \omega_2) \equiv g(\hat{g}^{-1}\omega_1, \hat{g}^{-1}\omega_2).$$

Note that the dual basis $\{\theta^a\}$ is orthonormal with respect to the scalar product g^{-1}.

Practice problem: **(a)** The metric is $g = dt^2 - t^2 dx^2$. Determine the 1-form $\hat{g}\mathbf{v}$ where $\mathbf{v} = \partial_t - \partial_x$.
(b) The metric is $g = dt\, dx$ (symmetric tensor product). Determine the vector $\hat{g}^{-1}\mathrm{d}t$ and the scalar products $g^{-1}(\mathrm{d}x, \mathrm{d}x)$, $g^{-1}(\mathrm{d}t, \mathrm{d}x)$. ■

Remark: In the index notation, the scalar product form g^{-1} is specified by the matrix $g^{\alpha\beta}$, which is inverse to the matrix $g_{\alpha\beta}$ representing the components of the metric g. ■

For a scalar function f, the 1-form $\mathrm{d}f$ called the **gradient** of f acts on vectors $\mathbf{x}$ as

$$(\mathrm{d}f) \circ \mathbf{x} \equiv \mathbf{x} \circ f,$$

and the corresponding vector field $\hat{g}^{-1}(\mathrm{d}f)$ may be called the **contravariant gradient** of f.

Example: For a vector field $\mathbf{x}$ and a scalar function f, we have

$$\mathbf{x} \circ f = g(\mathbf{x}, \hat{g}^{-1}\mathrm{d}f) = g^{-1}(\hat{g}\mathbf{x}, \mathrm{d}f).$$

The derivative of a scalar function ϕ in the direction of the vector $\hat{g}^{-1}\mathrm{d}f$, i.e. the scalar quantity $(\hat{g}^{-1}\mathrm{d}f) \circ \phi$, can be also expressed as

$$\left(\hat{g}^{-1}\mathrm{d}f\right) \circ \phi = g^{-1}(\mathrm{d}f, \mathrm{d}\phi) = \left(\hat{g}^{-1}\mathrm{d}\phi\right) \circ f.$$

■

Practice problem: Suppose $S(\mathbf{v})$ is a transformation-valued 1-form such that for any vectors $\mathbf{u}, \mathbf{v}, \mathbf{w}$ we have

$$S(\mathbf{v})\mathbf{w} = -S(\mathbf{w})\mathbf{v},$$
$$g(S(\mathbf{v})\mathbf{w}, \mathbf{u}) = g(\mathbf{w}, S(\mathbf{v})\mathbf{u}).$$

Show that these conditions are satisfied only if $S(\mathbf{v})\mathbf{w} \equiv 0$ for all $\mathbf{v}, \mathbf{w}$.

Hint: Define the auxiliary trilinear form $S(\mathbf{v}, \mathbf{w}, \mathbf{u}) \equiv g(S(\mathbf{v})\mathbf{w}, \mathbf{u})$ and investigate its symmetries. ■

1.5.6 The Levi-Civita tensor ε

Ordinary vector algebra in three-dimensional space makes use of the **Levi-Civita symbol** ε_{ijk}, which is defined as the totally antisymmetric array of numbers,

$$\varepsilon_{123} = 1, \quad \varepsilon_{ijk} = -\varepsilon_{ikj} = -\varepsilon_{jik}, \quad i, j, k = 1, 2, 3.$$

In curved space, the role of this symbol is played by a totally antisymmetric tensor field, called the Levi-Civita tensor.

Consider a four-dimensional manifold (to be specific and to avoid unnecessary complications in the notation). In four dimensions, the **Levi-Civita tensor** has rank four and is defined as the 4-form

$$\boldsymbol{\varepsilon} = \theta^0 \wedge \theta^1 \wedge \theta^2 \wedge \theta^3,$$

where $\{\theta^a\}$ is an orthonormal basis of 1-forms (see Sec. 1.5.4). We now review the motivation for this definition and the properties of the tensor $\boldsymbol{\varepsilon}$.

In a three-dimensional Euclidean space, the Levi-Civita symbol ε_{ijk} is closely related to volume. It is known that one can calculate the (oriented) volume of a parallelepiped spanned by three vectors $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$ by using the following explicit formula,

$$\mathrm{Vol}\,(\mathbf{a}, \mathbf{b}, \mathbf{c}) = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = \sum_{i,j,k} \varepsilon_{ijk} a^i b^j c^k,$$

where a^i, b^j, c^k are the Euclidean components of the three vectors (i.e. components in an orthogonal basis). We have seen in Sec. 1.4.4 that the oriented volume of such a parallelepiped is equal to the value of a 3-form evaluated on the vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$. Thus, the array of numbers ε_{ijk} can be interpreted as the array of components of the 3-form $\boldsymbol{\varepsilon}$ in the orthogonal basis. Then we have

$$\mathrm{Vol}\,(\mathbf{a}, \mathbf{b}, \mathbf{c}) = \boldsymbol{\varepsilon} \circ (\mathbf{a}, \mathbf{b}, \mathbf{c}).$$

Let us generalize this construction to a curved, four-dimensional manifold $\mathcal{M}$. In a tangent space at a fixed point $p \in \mathcal{M}$, we look for a totally antisymmetric form $\boldsymbol{\varepsilon}$ such that $\boldsymbol{\varepsilon}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4)$ is equal to the oriented 4-volume of the parallelepiped spanned by the tangent vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4$. If $\{\mathbf{e}_a\}$ is an orthonormal basis with a "positive" orientation then the 4-volume of the parallelepiped spanned by $\{\mathbf{e}_a\}$ is equal to 1. Thus, we are looking for a 4-form $\boldsymbol{\varepsilon}$ such that

$$\boldsymbol{\varepsilon}(\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3) = 1.$$

Since $\theta^a \circ \mathbf{e}_b = \delta^a_b$, it is clear that $\boldsymbol{\varepsilon} = \theta^0 \wedge \theta^1 \wedge \theta^2 \wedge \theta^3$ is one such 4-form.

We defined $\boldsymbol{\varepsilon}$ through a particular basis $\{\theta^a\}$, so we need to investigate whether and to what extent $\boldsymbol{\varepsilon}$ depends on the choice of the basis. Of course, $\boldsymbol{\varepsilon}$ changes sign when the orientation of the basis is reversed. It turns out that $\boldsymbol{\varepsilon}$ does not actually depend on the choice of the basis, as long as the orientation of the basis is fixed (see Statement 1.5.6.1 below). In particular, the orthonormal frame $\{\mathbf{e}_\alpha\}$ may be defined only locally, i.e. different frames $\{\mathbf{e}_\alpha\}$ need to be used in different parts of the manifold $\mathcal{M}$. Nevertheless, the tensor $\boldsymbol{\varepsilon}$ is defined uniquely and *globally*, i.e. throughout the entire manifold, as long as the manifold is globally orientable. A manifold is **globally orientable** if a choice of orthonormal frame can be made in every chart such that the orientation of the orthonormal frames is the same in every overlapping region between two charts. A Möbius strip is an example of a manifold that is not globally orientable. It seems that in General Relativity one never needs to consider nonorientable manifolds.

Statement 1.5.6.1: The Levi-Civita tensor

$$\boldsymbol{\varepsilon} = \theta^0 \wedge \theta^1 \wedge \theta^2 \wedge \theta^3$$

is invariant under changes of orthonormal basis $\{\theta^a\}$, except that it changes sign when the orientation of the basis is reversed.

Proof: Let us consider a linear transformation $\hat{T}$ that brings the basis $\{\mathbf{e}_a\}$ into another basis $\{\tilde{\mathbf{e}}_a\}$,

$$\tilde{\mathbf{e}}_a = \hat{T}\mathbf{e}_a, \quad \tilde{\theta}^a = \hat{T}^{-1}\theta^a.$$

By definition of the determinant (see Sec. 1.4.5), the oriented volume of the parallelepiped spanned by $\{\tilde{\mathbf{e}}_a\}$ is equal to $\det\hat{T}$. Thus, if the new basis $\{\tilde{\mathbf{e}}_a\}$ is orthonormal and has the same orientation as the old orthonormal basis, the oriented volume of the parallelepiped spanned by $\{\tilde{\mathbf{e}}_a\}$ is also equal to 1. Hence, we must have $\det\hat{T} = 1$. On the other hand,

$$\tilde{\varepsilon} \equiv \tilde{\theta}^0 \wedge \tilde{\theta}^1 \wedge \tilde{\theta}^2 \wedge \tilde{\theta}^3 = (\det\hat{T}^{-1})\theta^0 \wedge \theta^1 \wedge \theta^2 \wedge \theta^3 = \varepsilon,$$

by the same logic as

$$\tilde{\mathbf{e}}_1 \wedge \tilde{\mathbf{e}}_2 \wedge \tilde{\mathbf{e}}_3 \wedge \tilde{\mathbf{e}}_4 = (\det\hat{T})\mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \mathbf{e}_3 \wedge \mathbf{e}_4.$$

Therefore the new 4-form $\tilde{\varepsilon}$ defined through the new basis is equal to the old 4-form ε. ■

The Levi-Civita tensor ε is independent of the choice of an orthonormal basis $\{\theta^a\}$, but it *does* depend on the metric g. Let us now determine the explicit formula for the Levi-Civita tensor ε as a function of the metric g. All 4-forms in four-dimensional space are proportional to each other, so the Levi-Civita tensor can be written in a local coordinate system $\{x^\mu\}$ as

$$\varepsilon = f\,\mathrm{d}x^0 \wedge \mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3,$$

where f is some unknown scalar function.

It remains to determine the unknown coefficient f. To do that, we can use a suitable modification of Statement 1.4.5.2. By definition, the matrix elements $g_{\mu\nu}$ are equal to scalar products $g(\partial_\mu, \partial_\nu)$, where $\{\partial_\mu\} \equiv \{\partial/\partial x^\mu\}$ is a coordinate basis. Thus one can show that the 4-volume of the parallelepiped spanned by the tangent vectors $\{\partial_\mu\}$ is equal to $\pm\sqrt{|\det g_{\mu\nu}|}$. The factor "$\pm$" reflects the fact that the volume is oriented and may be negative if the coordinate basis $\{\partial_\mu\}$ has a negative orientation. The absolute value $|\det g_{\mu\nu}|$ is needed to compensate for the signature of the metric g, which might make the determinant negative. Thus

$$f \equiv \varepsilon(\partial_0, \partial_1, \partial_2, \partial_3) = \mathrm{Vol}(\partial_0, \partial_1, \partial_2, \partial_3) = \pm\sqrt{|\det g_{\mu\nu}|}.$$

It is convenient to assume that the coordinate basis $\{\partial_\mu\}$ has positive orientation. Hence the final formula for ε in local coordinates is

$$\varepsilon = \sqrt{|\det g_{\mu\nu}|}\,\mathrm{d}x^0 \wedge \mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3.$$

In this way, the functional dependence of ε on the metric g is made explicit.

It is useful to note that $\mathrm{d}\varepsilon = 0$. This is so because $\mathrm{d}\varepsilon$ is a 5-form, and there are no nonzero 5-forms in four-dimensional space.

The main use of the Levi-Civita tensor in GR is to compute 4-dimensional volume in curved space.[8] Since the tangent vectors are an approximate representation of short curve segments (Sec. 1.2.5), the 4-form ε yields a good approximation to the 4-dimensional volume for very small volumes. Therefore, one can compute the 4-dimensional volume of a spacetime domain by integrating the 4-form ε over that domain, and one can also integrate scalar functions times ε over a domain. The integration is defined purely geometrically, regardless of the choice of the coordinate system, once the metric g is fixed. For this reason (independence of coordinates), physicists know the Levi-Civita tensor ε by the name **covariant volume element**.

Sometimes the tensor ε is simply called the **volume 4-form** since the tensor $\varepsilon|_p$ evaluates the volume in each tangent space $T_p\mathcal{M}$. It is clear that the construction of the Levi-Civita tensor can be straightforwardly generalized to an N-dimensional manifold with a given metric and orientation. Through this generalization one obtains an N-form, which is frequently denoted by the symbol "Vol" as I have done above. Below I will also sometimes denote the Levi-Civita tensor by Vol, particularly when it is used only for integration over a manifold. I will use the notation ε when I need to use its specific properties as a 4-form, e.g. when I need to evaluate it on some tangent vectors or differentiate it.

[8] There are other, more advanced applications of the Levi-Civita tensor, such as the Hodge star operation, which I will consider below.

1.6 Affine connection

1.6.1 Motivation

To formulate the laws of physics, we need to write differential equations for some vector and tensor fields on the spacetime manifold. These differential equations involve derivatives of these vector or tensor fields in various directions. Thus we need to have a directional derivative operation that acts on arbitrary tensors.

One could think at first that the Lie derivative $\mathcal{L}_\mathbf{v}$ is to be used as a directional derivative operation. However, $\mathcal{L}_\mathbf{v}$ is not a true directional derivative because it depends on the behavior of the vector field $\mathbf{v}$ in the neighborhood of a point p and not only on the value $\mathbf{v}|_p$. A true directional derivative of a tensor A in a direction $\mathbf{v}$ at a point p must depend *only* on the value $\mathbf{v}|_p$ at the given point. Defining such a true directional derivative requires some information about how to "connect" neighbor tangent spaces $T_p\mathcal{M}$ and $T_{p'}\mathcal{M}$, where p' and p are "near-by" points in $\mathcal{M}$. If such a "connection" between the neighbor tangent spaces were defined, tensors at different points could be subtracted. For instance, $A|_{p'} - A|_p$ would be defined as a tensor in $T_p\mathcal{M}$, and the derivative of a tensor A along a curve γ at a point $p \equiv \gamma(\tau_0)$ could be defined as the limit

$$\lim_{\delta\tau\to 0} \frac{A(\gamma(\tau_0 + \delta\tau)) - A(\gamma(\tau_0))}{\delta\tau}.$$

Such a true directional derivative operation *can* be defined. It is usually denoted ∇ (pronounced "nabla" or "del") and called a **covariant derivative**, an **affine connection**, or simply a **connection**. Thus, $\nabla_\mathbf{u} A|_p$ is the derivative of a tensor A in the direction of the vector $\mathbf{u}$ at a point p. The quantity $\nabla_\mathbf{u} A|_p$ is a tensor and does not depend on derivatives of $\mathbf{u}$ itself.

It turns out that a connection ∇ is not unique because there are infinitely many ways to "connect" the neighbor tangent spaces, i.e. there are infinitely many possible maps $T_{p'}\mathcal{M} \to T_p\mathcal{M}$. Such a "connection map" can be described by a transformation-valued 1-form

$$C(\mathbf{u}) : T_{p'}\mathcal{M} \to T_p\mathcal{M},$$

where $\mathbf{u}$ is the tangent vector at point p in the direction of point p'. Thus, there are as many connections ∇ as maps C. In

a local coordinate system, the map C is represented by a quantity with three indices, $C^{\lambda}_{\alpha\beta}$, but it is not a tensor in a single tangent space at p: by definition, it is a map connecting *different* tangent spaces. (The map C is related to the non-tensorial Christoffel symbol; see Sec. 1.6.3 below.)

In General Relativity, one formulates the laws of physics using a particular choice of the connection ∇ (called the "Levi-Civita connection," see Sec. 1.6.6 below). This connection has several physically motivated and technically convenient properties with regard to the given metric g on the spacetime manifold. So the Levi-Civita connection ∇ directly involves the metric g and cannot be defined unless a metric is given.

A visual motivation for the Levi-Civita connection, using the idea of embedding a curved manifold in a flat space, is given in Sec. A.4.3 (Appendix A). Presently, I approach the definition of the Levi-Civita connection in a somewhat more abstract way: I first formulate the desirable properties of connections in general and then impose suitable conditions to deduce the Levi-Civita connection.

Since we are studying the general properties of tensors, we will avoid working in a local coordinate system; instead, we will perform calculations in a coordinate-free manner.

1.6.2 General properties of connections

We would like to define a derivative operator $\nabla_{\mathbf{v}}$ which maps scalars to scalars, vectors to vectors, etc., and depends only on the direction of the vector $\mathbf{v}$ at a point p. The condition that $\nabla_{\mathbf{v}} A$ should not contain any derivatives of $\mathbf{v}$ can be written as

$$\nabla_{(\lambda\mathbf{v})} A = \lambda \nabla_{\mathbf{v}} A,$$

where λ is an arbitrary scalar function. If we had such an operation $\nabla_{\mathbf{v}}$, we could define the "gradient" operator ∇ which maps scalars to 1-forms, vectors to (1,1)-tensors, and generally tensors of rank (m, n) to tensors of rank $(m, n+1)$. In the index notation, the symbol ∇ will have its own index; e.g. the "gradient" of a tensor $A_{\alpha\beta}$ is a tensor of rank (0,3) denoted by $\nabla_\mu A_{\alpha\beta} \equiv A_{\alpha\beta;\mu}$.

Remark: Vectors are always boldface in our index-free convention, so there should be no confusion between expressions such as $\nabla_\nu a^\mu$ and $\nabla_{\mathbf{v}}\mathbf{a}$. The former is the index representation of the rank-two tensor $\nabla.\mathbf{a}$, while the latter is the covariant derivative of the vector field $\mathbf{a}$ in the given direction $\mathbf{v}$. The index representation of $\nabla_{\mathbf{v}}\mathbf{a}$ is $v^\alpha a^\mu_{\ ;\alpha}$. The derivative of a scalar function f in the direction of a vector $\mathbf{v}$ may be equivalently rewritten as $\mathbf{v} \circ f = \mathcal{L}_{\mathbf{v}} f = \nabla_{\mathbf{v}} f = (\mathrm{d}f) \circ \mathbf{v}$, according to the convenience of the moment. The index representation of $\mathbf{v} \circ f$ is $v^\alpha f_{,\alpha} = v^\alpha f_{;\alpha}$. ■

Of course, $\nabla_{\mathbf{v}}$ should also act on scalar functions as the usual directional derivative, i.e.

$$\nabla_{\mathbf{v}} f = \mathbf{v} \circ f.$$

In addition, the operator $\nabla_{\mathbf{v}}$ should have the properties analogous to Eqs. (1.4)-(1.6), (1.12)-(1.13), except that the contraction "$\circ$" now means *only* a contraction of tensors (e.g., application of a 1-form to a vector) and not the derivative operation $\mathbf{v} \circ f$.

Remark: Unlike the Lie derivative, the connection ∇ is not expected to satisfy the property

$$\nabla_{\mathbf{u}} (\mathbf{v} \circ f) \neq (\nabla_{\mathbf{u}} \mathbf{v}) \circ f + \mathbf{v} \circ (\nabla_{\mathbf{u}} f).$$

If we assumed that this contraction property holds, we would be forced to have $\nabla_{\mathbf{u}} \equiv \mathcal{L}_{\mathbf{u}}$. ■

1.6.3 The "coordinate derivative" connection

The above properties of the operation ∇ do not yet uniquely specify that operation. There are infinitely many possible affine connections. A simple way to define a connection is to fix a local coordinate system $\{x^\mu\}$ and set

$$\nabla_\lambda \equiv \partial_\lambda \equiv \frac{\partial}{\partial x^\lambda}; \quad (\partial_{\mathbf{v}} A)^{\alpha\beta\gamma\ldots} \equiv v^\lambda \frac{\partial}{\partial x^\lambda} A^{\alpha\beta\gamma\ldots} \equiv v^\lambda A^{\alpha\beta\gamma\ldots}_{\ \ \ \ ,\lambda}$$

for any tensor A. In other words, the component $A^{\alpha\beta\gamma\ldots}_{\ \ \ \ ;\lambda}$ of the covariant derivative of A is calculated as the partial derivative with respect to x^λ of the component $A^{\alpha\beta\gamma\ldots}$ of the tensor A in the fixed coordinate system $\{x^\mu\}$. It is customary to use the semicolon in the expression $A^{\alpha\beta\gamma}_{\ \ \ ;\lambda}$ before the extra index introduced by the covariant derivative.

In the index-free notation, we denote the derivative with respect to the coordinates in a fixed coordinate system by ∂. Thus, the vector $\partial_{\mathbf{u}}\mathbf{v}$ is defined through the components in the fixed coordinate system $\{x^\mu\}$ as

$$(\partial_{\mathbf{u}}\mathbf{v})^\alpha = u^\lambda v^\alpha_{\ ,\lambda}.$$

In the coordinate system $\{x^\mu\}$, the derivative ∂ is just the usual derivative, so it is easy to see that ∂ satisfies all the properties of a connection. Thus, ∂ is a well-defined affine connection which we call the **coordinate derivative** connection **determined by the coordinate system** $\{x^\mu\}$. We stress that the connection ∂ is tied to a *particular* coordinate system $\{x^\mu\}$. In a different coordinate system $\{y^\mu\}$, the components of the vector $\partial_{\mathbf{u}}\mathbf{v}$ do not have the same simple form $u^\lambda \partial v^\alpha / \partial y^\lambda$. Thus, for *each* coordinate system $\{x^\mu\}$ we have a different coordinate connection ∂.

Remark: The notation "∂" does not indicate explicitly the coordinate system $\{x^\mu\}$ through which ∂ is defined. So it is important to identify that coordinate system every time one uses ∂. We will rarely make use of ∂, but when we do, we will make it clear which coordinate system $\{x^\mu\}$ is implied in the definition of ∂. ■

Let us also investigate how an arbitrary connection ∇ is related to the coordinate derivative ∂. It is easy to verify that $\nabla - \partial$ acts on vectors $\mathbf{v}$ linearly (without involving derivatives of $\mathbf{v}$),

$$\nabla_{\mathbf{u}} (\lambda\mathbf{v}) - \partial_{\mathbf{u}} (\lambda\mathbf{v}) = \lambda (\nabla_{\mathbf{u}}\mathbf{v} - \partial_{\mathbf{u}}\mathbf{v}).$$

Hence, for a fixed vector $\mathbf{u}$ the operator $\nabla_{\mathbf{u}} - \partial_{\mathbf{u}}$ is a linear transformation. Equivalently, we can say that $\nabla - \partial$ is a transformation-valued 1-form and write

$$\nabla - \partial \equiv \Gamma; \quad (\nabla_{\mathbf{u}} - \partial_{\mathbf{u}})\,\mathbf{v} = \Gamma(\mathbf{u})\,\mathbf{v}.$$

The tensor Γ is called the **Christoffel tensor** of the connection ∇ with respect to the coordinate system $\{x^\mu\}$ in which ∂ is the coordinate derivative.

For a fixed coordinate derivative ∂, various choices of Γ will give different connections ∇. Thus, the Christoffel tensor Γ parametrizes all the possible covariant derivatives ∇.

Practice problem: Suppose that two different connections ∇ and $\tilde{\nabla}$ are given (not necessarily coordinate derivative connections in any coordinate system). Using the defining properties of a connection, show that the difference $\tilde{\nabla}_{\mathbf{u}}\mathbf{v} - \nabla_{\mathbf{u}}\mathbf{v}$ does not depend on derivatives of $\mathbf{v}$ and can thus be described as a linear transformation of $\mathbf{v}$ (for a fixed $\mathbf{u}$). ■

1.6.4 Compatibility with the metric

To formulate the laws of physics, we need to write differential equations for tensors on the spacetime manifold, and thus we need to use a directional derivative ∇. One possible choice of ∇ is the coordinate derivative ∂ in a fixed coordinate system $\{x^\mu\}$. But the connection $\nabla \equiv \partial$ is inconvenient not only because it depends on a fixed coordinate system (while the laws of physics are expected to be independent of coordinates), but also because it is in general *incompatible with the metric* in the following sense.

Consider a vector $\mathbf{v}$ which is "locally constant" in the direction of some other vector $\mathbf{u}$, i.e. $\nabla_{\mathbf{u}}\mathbf{v} = 0$. Then it is natural to expect that the length of the vector $\mathbf{v}$ is also constant in the direction of $\mathbf{u}$, i.e. that $\nabla_{\mathbf{u}} g(\mathbf{v},\mathbf{v}) = 0$. However, this property does not necessarily hold, since we will generally have

$$\begin{aligned}\nabla_{\mathbf{u}} g(\mathbf{v},\mathbf{v}) &= (\nabla_{\mathbf{u}} g)\circ(\mathbf{v},\mathbf{v}) + g(\nabla_{\mathbf{u}}\mathbf{v},\mathbf{v}) + g(\mathbf{v},\nabla_{\mathbf{u}}\mathbf{v}) \\ &= (\nabla_{\mathbf{u}} g)\circ(\mathbf{v},\mathbf{v}) \neq 0.\end{aligned}$$

This inconvenience will be avoided only if $\nabla_{\mathbf{u}} g = 0$ for any vector $\mathbf{u}$. The property $\nabla g = 0$ is called **compatibility with the metric** (or **metricity** of the connection). An equivalent way to write this property is

$$\mathbf{u}\circ g(\mathbf{v},\mathbf{w}) \equiv \nabla_{\mathbf{u}} g(\mathbf{v},\mathbf{w}) = g(\nabla_{\mathbf{u}}\mathbf{v},\mathbf{w}) + g(\mathbf{v},\nabla_{\mathbf{u}}\mathbf{w}). \quad (1.41)$$

In other words, g behaves as a constant under a covariant derivative $\nabla_{\mathbf{u}}$, and only $\mathbf{v}$ and $\mathbf{w}$ are differentiated when one computes $\nabla_{\mathbf{u}} g(\mathbf{v},\mathbf{w})$.

1.6.5 Torsion and torsion-freeness

Usual partial derivatives of *scalar* functions with respect to coordinates x^μ (in *any* local coordinate system) commute,

$$\partial_\mu \partial_\nu f = \partial_\nu \partial_\mu f.$$

We may ask whether the same property holds for covariant derivatives,

$$\nabla_\mu \nabla_\nu f = \nabla_\nu \nabla_\mu f. \quad (1.42)$$

At the moment this property is written using the index notation. To reformulate the property (1.42) in a geometric way (without introducing coordinates explicitly), we contract Eq. (1.42) with two arbitrary vectors $u^\mu v^\nu$ and rewrite the result using covariant derivatives,

$$\begin{aligned}u^\mu v^\nu \nabla_\mu \nabla_\nu f &= u^\mu \nabla_\mu \left(u^\nu \nabla_\nu f\right) - \left(u^\mu \nabla_\mu u^\nu\right)\left(\nabla_\nu f\right) \\ &= \mathbf{u}\circ(\mathbf{v}\circ f) - (\nabla_{\mathbf{u}}\mathbf{v})\circ f.\end{aligned}$$

Thus, the property (1.42) is equivalent to

$$(\nabla_{\mathbf{u}}\mathbf{v} - \nabla_{\mathbf{v}}\mathbf{u})\circ f = \mathbf{u}\circ(\mathbf{v}\circ f) - \mathbf{v}\circ(\mathbf{u}\circ f) = [\mathbf{u},\mathbf{v}]\circ f.$$

It is not necessarily true that an arbitrary connection ∇ satisfies this property. To describe the extent of the deviation from this property, one defines the **torsion tensor**,

$$T(\mathbf{u},\mathbf{v}) \equiv \nabla_{\mathbf{u}}\mathbf{v} - \nabla_{\mathbf{v}}\mathbf{u} - [\mathbf{u},\mathbf{v}], \quad (1.43)$$

which is a vector-valued 2-form (see Statement 1.6.5.1 below). The property (1.42) is then equivalent to the requirement that $T(\mathbf{u},\mathbf{v}) = 0$ for all $\mathbf{u},\mathbf{v}$; this is called the **torsion-free**ness of the connection ∇. Thus, if a connection ∇ is torsion-free, we have the relation

$$\mathcal{L}_{\mathbf{u}}\mathbf{v} \equiv [\mathbf{u},\mathbf{v}] = \nabla_{\mathbf{u}}\mathbf{v} - \nabla_{\mathbf{v}}\mathbf{u}. \quad (1.44)$$

This relation is similar to Eq. (1.14), except that now covariant derivatives ∇ are used. Note that the coordinate derivative ∂ (defined with respect to a coordinate system $\{x^\mu\}$) is always a torsion-free connection because coordinate derivatives commute.

Statement 1.6.5.1: The function $T(\mathbf{u},\mathbf{v})$ as defined by Eq. (1.43) does not actually involve derivatives of the vectors $\mathbf{u}$ and $\mathbf{v}$. Namely, for an arbitrary scalar function f, we have $T(f\mathbf{u},\mathbf{v}) = fT(\mathbf{u},\mathbf{v})$ and $T(\mathbf{u},f\mathbf{v}) = fT(\mathbf{u},\mathbf{v})$.

Proof of Statement 1.6.5.1: For arbitrary scalars f,ϕ and vectors $\mathbf{u},\mathbf{v}$ we have

$$\begin{aligned}\nabla_{\mathbf{v}}(f\mathbf{u}) &= (\nabla_{\mathbf{v}} f)\,\mathbf{u} + f\nabla_{\mathbf{v}}\mathbf{u} = (\mathbf{v}\circ f)\,\mathbf{u} + f\nabla_{\mathbf{v}}\mathbf{u}, \\ \nabla_{f\mathbf{u}}\mathbf{v} &= f\nabla_{\mathbf{u}}\mathbf{v}, \\ [f\mathbf{u},\mathbf{v}]\circ\phi &= f\mathbf{u}\circ(\mathbf{v}\circ\phi) - \mathbf{v}\circ((f\mathbf{u})\circ\phi) \\ &= f\,[\mathbf{u},\mathbf{v}]\circ\phi - (\mathbf{v}\circ f)\,(\mathbf{u}\circ\phi),\end{aligned}$$

and it follows that

$$\begin{aligned}T(f\mathbf{u},\mathbf{v})\circ\phi &= fT(\mathbf{u},\mathbf{v})\circ\phi - ((\mathbf{v}\circ f)\,\mathbf{u})\circ\phi + (\mathbf{v}\circ f)\,(\mathbf{u}\circ\phi) \\ &= fT(\mathbf{u},\mathbf{v}).\end{aligned}$$

The analogous property for $\mathbf{v}$ will follow from the antisymmetry of T. ■

We have seen that all possible connections ∇ are parameterized as $\nabla = \partial + \Gamma$, where Γ is a rank 3 tensor (the Christoffel tensor) with respect to a fixed coordinate system $\{x^\mu\}$ where ∂ is the coordinate derivative. We can now compute the torsion tensor T through Γ. Since ∂ is torsion-free, it follows immediately from Eq. (1.43) that

$$T(\mathbf{u},\mathbf{v}) = \Gamma(\mathbf{u})\mathbf{v} - \Gamma(\mathbf{v})\mathbf{u}.$$

To visualize this property, we may interpret Γ as a vector-valued bilinear form $\Gamma(\mathbf{u},\mathbf{v}) \equiv \Gamma(\mathbf{u})\mathbf{v}$, and then the property $T(\mathbf{u},\mathbf{v}) = 0$ becomes

$$\Gamma(\mathbf{u},\mathbf{v}) = \Gamma(\mathbf{v},\mathbf{u}).$$

Thus a torsion-free connection ∇ is characterized by a **symmetric** Christoffel tensor.

Practice problem: For a given affine connection ∇, does there always exist a coordinate system $\{x^\mu\}$ where ∇ is equal to the coordinate derivative?

Hint: Consider the torsion of the connection ∇. *Answer:* No. ■

1.6.6 Levi-Civita connection

A covariant derivative ∇ which is torsion-free and compatible with the metric is called the **Levi-Civita connection** or the **metric connection**. In General Relativity, this is the connection used to formulate differential equations for physical quantities. The properties of metricity and torsion-freeness

are imposed for several reasons, including mathematical simplicity. Ultimately, these properties are *physical assumptions*, that is, hypotheses verified by experiments. See Sec. A.4.5 in Appendix A for more discussion and motivation.

We now derive an explicit formula for the Levi-Civita connection, starting from the properties (1.41) and (1.42). It will follow that such a connection is unique. Later we shall see how to describe other connections.

First examples

Before presenting a general derivation, let us make the discussion less abstract by considering examples. First we verify that the Levi-Civita connection in the Euclidean space is just the ordinary, familiar directional derivative. Consider the three-dimensional space $\mathbb{R}^3$with coordinates $\{x, y, z\}$ and the metric

$$g = dx^2 + dy^2 + dz^2,$$

and let us evaluate some covariant derivatives. We will be using only the assumed properties of the Levi-Civita connection and see where this will lead us.

Suppose we need to compute the vector $\nabla_{\partial_x}\partial_x$. The way to compute it is to evaluate scalar products of this vector with other vectors. For instance, using the compatibility of ∇ with the metric, and the properties $g(\partial_x, \partial_x) = 1$, $g(\partial_x, \partial_y) = 0$ etc., we find

$$\begin{aligned} g(\nabla_{\partial_x}\partial_x, \partial_x) &= \frac{1}{2} g(\nabla_{\partial_x}\partial_x, \partial_x) + \frac{1}{2} g(\partial_x, \nabla_{\partial_x}\partial_x) \\ &= \frac{1}{2}\nabla_{\partial_x} g(\partial_x, \partial_x) = \frac{1}{2}\partial_x(1) = 0. \end{aligned}$$

Using the torsion-freeness and the fact that $[\partial_x, \partial_y] = 0$, we also find

$$\begin{aligned} g(\nabla_{\partial_x}\partial_x, \partial_y) &= \nabla_{\partial_x} g(\partial_x, \partial_y) - g(\partial_x, \nabla_{\partial_x}\partial_y) \\ &= 0 - g(\partial_x, [\partial_x, \partial_y] + \nabla_{\partial_y}\partial_x) \\ &= -\frac{1}{2}\nabla_{\partial_y} g(\partial_x, \partial_x) = 0. \end{aligned}$$

In this way, it is easy to see that $\nabla_{\partial_x}\partial_x$ has zero scalar products with every basis vector; thus it is itself zero, $\nabla_{\partial_x}\partial_x = 0$.

Let us now consider $\nabla_{\partial_x}\partial_y$. We have already seen that $g(\nabla_{\partial_x}\partial_y, \partial_x) = 0$, and further we obtain

$$g(\nabla_{\partial_x}\partial_y, \partial_y) = \frac{1}{2}\nabla_{\partial_x} g(\partial_y, \partial_y) = 0.$$

The computation of the scalar product $g(\nabla_{\partial_x}\partial_y, \partial_z)$ is a bit longer. First, we find

$$\begin{aligned} g(\nabla_{\partial_x}\partial_y, \partial_z) &= g([\partial_x, \partial_y] + \nabla_{\partial_y}\partial_x, \partial_z) = -g(\partial_x, \nabla_{\partial_y}\partial_z) \\ &= -g(\nabla_{\partial_y}\partial_z, \partial_x). \end{aligned}$$

The trick is to notice that we now have a similar structure with the coordinates $\{x, y, z\}$ cyclically transposed. Repeating the transposition twice more, we find

$$g(\nabla_{\partial_x}\partial_y, \partial_z) = -g(\nabla_{\partial_x}\partial_y, \partial_z),$$

so it vanishes. Thus, $\nabla_{\partial_x}\partial_y = 0$. Similarly, one can show that $\nabla_{\partial_x}\partial_z = 0$, etc. In other words, in flat space the covariant derivative of every basis vector in any direction equals zero. Now, we consider an arbitrary vector field

$$\mathbf{v} = A\partial_x + B\partial_y + C\partial_z,$$

where A, B, C are functions of x, y, z. We compute

$$\nabla_{\partial_x}\mathbf{v} = (\partial_x A)\,\partial_x + (\partial_x B)\,\partial_y + (\partial_x C)\,\partial_z,$$

since the derivatives ∇_{∂_x} of the vectors ∂_x, ∂_y, ∂_z vanish. In other words, the Levi-Civita connection ∇_{∂_x} simply differentiates the components (A, B, C) of the vector $\mathbf{v}$ in the direction x. This is the behavior familiar from standard vector analysis. Thus, the Levi-Civita connection ∇_{∂_x} in flat space coincides with the familiar directional derivative $\partial/\partial x$.

As a less trivial example, let us consider a two-dimensional space with a local coordinate system (x, y) and the metric

$$g = dx^2 + f(x)dy^2,$$

where $f(x) \neq 0$ is some (known) function that depends only on x. Let us compute $\nabla_{\partial_x}\partial_y$ using the same approach as above. We find

$$\begin{aligned} g(\nabla_{\partial_x}\partial_y, \partial_x) &= g([\partial_x, \partial_y] + \nabla_{\partial_y}\partial_x, \partial_x) = \frac{1}{2}\partial_y g(\partial_x, \partial_x) = 0; \\ g(\nabla_{\partial_x}\partial_y, \partial_y) &= \frac{1}{2}\nabla_{\partial_x} g(\partial_y, \partial_y) = \frac{1}{2}f'(x). \end{aligned}$$

In this way, the vector $\nabla_{\partial_x}\partial_y$ is completely determined through its scalar products with basis vectors ∂_x and ∂_y:

$$\nabla_{\partial_x}\partial_y = \frac{1}{2}\frac{f'(x)}{f(x)}\partial_y.$$

Similarly, one may evaluate other covariant derivatives, such as

$$\begin{aligned} &g(\nabla_{\partial_y}\partial_y, \partial_y) = 0, \\ &g(\nabla_{\partial_y}\partial_y, \partial_x) = -g(\partial_y, \nabla_{\partial_y}\partial_x) = -g(\partial_y, \nabla_{\partial_x}\partial_y) = -\frac{1}{2}f'(x). \end{aligned}$$

Thus

$$\nabla_{\partial_y}\partial_y = -\frac{1}{2}f'(x)\partial_x.$$

Derivation of the Levi-Civita connection

In the examples above, we have computed covariant derivatives of some vectors by using only the defining properties (1.41) and (1.44) of the Levi-Civita connection. We now derive a general formula for $\nabla_{\mathbf{x}}\mathbf{y}$ using the same approach.

If the vector fields $\mathbf{x}, \mathbf{y}$ are given, the vector $\nabla_{\mathbf{x}}\mathbf{y}$ will be unambiguously determined if we compute its scalar product $g(\nabla_{\mathbf{x}}\mathbf{y}, \mathbf{z})$ with an arbitrary vector $\mathbf{z}$. So we start with the scalar product expression $g(\nabla_{\mathbf{x}}\mathbf{y}, \mathbf{z})$ and rewrite it using the assumed properties (1.41) and (1.44),

$$\begin{aligned} g(\nabla_{\mathbf{x}}\mathbf{y}, \mathbf{z}) &= \nabla_{\mathbf{x}} g(\mathbf{y}, \mathbf{z}) - g(\mathbf{y}, \nabla_{\mathbf{x}}\mathbf{z}) \\ &= \mathbf{x} \circ g(\mathbf{y}, \mathbf{z}) - g(\mathbf{y}, [\mathbf{x}, \mathbf{z}]) - g(\mathbf{y}, \nabla_{\mathbf{z}}\mathbf{x}) \\ &= \mathbf{x} \circ g(\mathbf{y}, \mathbf{z}) - g(\mathbf{y}, [\mathbf{x}, \mathbf{z}]) - g(\nabla_{\mathbf{z}}\mathbf{x}, \mathbf{y}). \end{aligned}$$

We have related $g(\nabla_{\mathbf{x}}\mathbf{y}, \mathbf{z})$ to the same structure with $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ cyclically transposed. So now we repeat the same procedure twice more, replacing the term with $\nabla_{\mathbf{z}}\mathbf{x}$ by a term with $\nabla_{\mathbf{y}}\mathbf{z}$, and finally by a term with $\nabla_{\mathbf{x}}\mathbf{y}$. We obtain

$$\begin{aligned} g(\nabla_{\mathbf{x}}\mathbf{y}, \mathbf{z}) &= \mathbf{x} \circ g(\mathbf{y}, \mathbf{z}) - g(\mathbf{y}, [\mathbf{x}, \mathbf{z}]) - \mathbf{z} \circ g(\mathbf{x}, \mathbf{y}) \\ &\quad + g(\mathbf{x}, [\mathbf{z}, \mathbf{y}]) + \mathbf{y} \circ g(\mathbf{x}, \mathbf{z}) - g(\mathbf{z}, [\mathbf{y}, \mathbf{x}]) - g(\nabla_{\mathbf{x}}\mathbf{y}, \mathbf{z}). \end{aligned}$$

Thus the final formula for the Levi-Civita connection is

$$\begin{aligned} g(\nabla_{\mathbf{x}}\mathbf{y}, \mathbf{z}) = \frac{1}{2}\big(&\mathbf{x} \circ g(\mathbf{y}, \mathbf{z}) + \mathbf{y} \circ g(\mathbf{x}, \mathbf{z}) - \mathbf{z} \circ g(\mathbf{x}, \mathbf{y}) \\ &- g(\mathbf{x}, [\mathbf{y}, \mathbf{z}]) - g(\mathbf{y}, [\mathbf{x}, \mathbf{z}]) + g(\mathbf{z}, [\mathbf{x}, \mathbf{y}])\big). \end{aligned} \tag{1.45}$$

This is known as the **Koszul formula**.

Calculation 1.6.6.1: The familiar formula for the covariant derivative in the index notation,

$$\nabla_\mu u^\nu = \partial_\mu u^\nu + \Gamma^\nu_{\alpha\mu} u^\alpha,$$
$$\Gamma^\lambda_{\alpha\beta} \equiv \frac{1}{2} g^{\lambda\mu} \left(\partial_\alpha g_{\mu\beta} + \partial_\beta g_{\mu\alpha} - \partial_\mu g_{\alpha\beta} \right), \tag{1.46}$$

can be derived from Eq. (1.45).

Details: The Christoffel tensor Γ describes the difference between the covariant derivative ∇ and the coordinate derivative ∂ in a particular coordinate system. Let us select a coordinate system and substitute $\mathbf{x} = \partial_\alpha$, $\mathbf{y} = \partial_\beta$, $\mathbf{z} = \partial_\mu$ (where α, β, μ are *fixed* values of the indices) into Eq. (1.45). We note that the commutators of these vectors vanish and so

$$2g(\nabla_{\partial_\alpha} \partial_\beta, \partial_\mu) = \partial_\alpha g_{\beta\mu} + \partial_\beta g_{\mu\alpha} - \partial_\mu g_{\alpha\beta}.$$

Now we can decompose an arbitrary vector $\mathbf{u}$ as $\sum_\lambda u^\lambda \partial_\lambda$ and compute

$$\nabla_\alpha u^\nu = g^{\mu\nu} g(\nabla_{\partial_\alpha} \mathbf{u}, \partial_\nu) = \partial_\alpha u^\nu + \Gamma^\nu_{\alpha\mu} u^\mu.$$

■

Note that the Levi-Civita connection ∇ is defined through the metric g and thus implicitly depends on g. For instance, the vector field $\nabla_\mathbf{u}\mathbf{v}$ involves not only first derivatives of $\mathbf{v}$, but also first derivatives of the metric. This is easy to see by inspection of Eq. (1.45), where the metric g clearly stands under differentiation in the first three terms. It is instructive to verify this statement more directly, without using Eq. (1.45). Let us change the metric tensor g as follows,

$$g \to \tilde{g} = e^{2\lambda} g,$$

where λ is a scalar function. (This transformation of the metric is called a **conformal transformation** because it preserves angles between vectors.) Now we can use Eq. (1.45) to determine the new connection $\tilde{\nabla}$. It is convenient to compute $\tilde{g}(\tilde{\nabla}_\mathbf{x}\mathbf{y}, \mathbf{z})$ using the *new* metric $\tilde{g}$, rather than the old metric. We find

$$\tilde{g}(\tilde{\nabla}_\mathbf{x}\mathbf{y}, \mathbf{z}) = (\mathbf{x} \circ \lambda)\, \tilde{g}(\mathbf{y}, \mathbf{z}) + (\mathbf{y} \circ \lambda)\, \tilde{g}(\mathbf{x}, \mathbf{z}) - (\mathbf{z} \circ \lambda)\, \tilde{g}(\mathbf{x}, \mathbf{y}) + \tilde{g}(\nabla_\mathbf{x}\mathbf{y}, \mathbf{z}). \tag{1.47}$$

This expression contains first derivatives of λ such as $\mathbf{x} \circ \lambda$. Therefore, the Levi-Civita connection indeed involves first derivatives of the metric.

Remark: As we have seen in Sec. 1.6.3, all the possible connections (covariant derivatives) ∇ can be described by choosing a coordinate system $\{x^\mu\}$ (in which the coordinate derivative is ∂) and a transformation-valued 1-form Γ. Then $\nabla_\mathbf{u}\mathbf{v} = \partial_\mathbf{u}\mathbf{v} + \Gamma(\mathbf{u})\mathbf{v}$. Torsion-freeness implies the symmetry $\Gamma(\mathbf{u})\mathbf{v} = \Gamma(\mathbf{v})\mathbf{u}$, but Γ is not otherwise constrained. Of all the possible torsion-free connections, the Levi-Civita connection is described by a particular unique choice of Γ determined by the metricity requirement. ■

Calculation: Recall that the (2,0) tensor g^{-1} is the inverse metric defined on 1-forms, or alternatively the map from vectors to 1-forms to vectors. We now show that $\nabla_\mathbf{u}(g^{-1}) = 0$ for the Levi-Civita connection ∇.

A simple (but less explicit) calculation is as follows. Consider the map $\hat{g}$ from vectors to 1-forms and the map $\hat{g}^{-1}$ from 1-forms to vectors. We have

$$\hat{g}\hat{g}^{-1} = \hat{1},$$

where $\hat{1}$ is the identity map from 1-forms to 1-forms. Then

$$\nabla_\mathbf{u}(\hat{g}\hat{g}^{-1}) = \hat{g}\nabla_\mathbf{u}\hat{g}^{-1} = \nabla_\mathbf{u}\hat{1} = 0,$$

thus $\nabla_\mathbf{u}\hat{g}^{-1} = 0$.

Here is a more detailed calculation along the same lines. For an arbitrary vector $\mathbf{v}$ and 1-form ω, we have

$$\omega \circ \mathbf{v} = g(\hat{g}^{-1}\mathbf{!}, \mathbf{v}).$$

Now we apply $\nabla_\mathbf{u}$ to both sides,

$$\nabla_\mathbf{u}(\omega \circ \mathbf{v}) = \nabla_\mathbf{u} g(\hat{g}^{-1}\mathbf{!}, \mathbf{v}) = g(\nabla_\mathbf{u}(\hat{g}^{-1}\mathbf{!}), \mathbf{v}) + g(\hat{g}^{-1}\mathbf{!}, \nabla_\mathbf{u}\mathbf{v}).$$

Since ∇ satisfies Leibnitz's rule with respect to tensor products and contractions, we find

$$\nabla_\mathbf{u}(\omega \circ \mathbf{v}) = (\nabla_\mathbf{u}\omega) \circ \mathbf{v} + \omega \circ (\nabla_\mathbf{u}\mathbf{v}) = g(\hat{g}^{-1}(\nabla_\mathbf{u}\omega), \mathbf{v}) + g(\hat{g}^{-1}\omega, \nabla_\mathbf{u}\mathbf{v}).$$

The result $\nabla_\mathbf{u}\hat{g}^{-1} = 0$ follows. ■

1.6.7 Killing vectors

The Lie derivative operation $\mathcal{L}_\mathbf{v}$ can be applied to arbitrary tensors, and in particular to the metric tensor g. Let us compute the Lie derivative $\mathcal{L}_\mathbf{v} g$ with respect to a given vector field $\mathbf{v}$. By definition, $\mathcal{L}_\mathbf{v} g$ is a bilinear form acting on two vectors $\mathbf{a}, \mathbf{b}$ by first letting $\mathcal{L}_\mathbf{v}$ act on the entire expression $g(\mathbf{a}, \mathbf{b})$ and then subtracting the Lie derivatives $\mathcal{L}_\mathbf{v}$ of $\mathbf{a}$ and $\mathbf{b}$:

$$(\mathcal{L}_\mathbf{v} g) \circ (\mathbf{a}, \mathbf{b}) \equiv \mathcal{L}_\mathbf{v} \left(g(\mathbf{a}, \mathbf{b}) \right) - g(\mathcal{L}_\mathbf{v}\mathbf{a}, \mathbf{b}) - g(\mathbf{a}, \mathcal{L}_\mathbf{v}\mathbf{b}) = \mathbf{v} \circ g(\mathbf{a}, \mathbf{b}) - g([\mathbf{v}, \mathbf{a}], \mathbf{b}) - g(\mathbf{a}, [\mathbf{v}, \mathbf{b}]).$$

There exists a more convenient formula for the Lie derivative of the metric. This formula uses the Levi-Civita connection ∇.

Statement 1.6.7.1: The Lie derivative of the metric g with respect to a vector $\mathbf{k}$ is a bilinear form that can be computed as

$$(\mathcal{L}_\mathbf{k} g) \circ (\mathbf{a}, \mathbf{b}) = g(\nabla_\mathbf{a}\mathbf{k}, \mathbf{b}) + g(\mathbf{a}, \nabla_\mathbf{b}\mathbf{k}). \tag{1.48}$$

Proof: Let $\mathbf{k}$ be an arbitrary vector field. By definition of $\mathcal{L}_\mathbf{k} g$, we have for arbitrary vector fields $\mathbf{a}, \mathbf{b}$ the expression

$$\mathcal{L}_\mathbf{k} g(\mathbf{a}, \mathbf{b}) = (\mathcal{L}_\mathbf{k} g) \circ (\mathbf{a}, \mathbf{b}) + g(\mathcal{L}_\mathbf{k}\mathbf{a}, \mathbf{b}) + g(\mathbf{a}, \mathcal{L}_\mathbf{k}\mathbf{b}) = (\mathcal{L}_\mathbf{k} g) \circ (\mathbf{a}, \mathbf{b}) + g([\mathbf{k}, \mathbf{a}], \mathbf{b}) + g(\mathbf{a}, [\mathbf{k}, \mathbf{b}]). \tag{1.49}$$

On the other hand, $g(\mathbf{a}, \mathbf{b})$ is a scalar function, and $\mathcal{L}_\mathbf{k} g(\mathbf{a}, \mathbf{b}) = \nabla_\mathbf{k} g(\mathbf{a}, \mathbf{b})$. Then the metricity condition $\nabla_\mathbf{k} g = 0$ yields

$$\mathcal{L}_\mathbf{k} g(\mathbf{a}, \mathbf{b}) = \nabla_\mathbf{k} g(\mathbf{a}, \mathbf{b}) = g(\nabla_\mathbf{k}\mathbf{a}, \mathbf{b}) + g(\mathbf{a}, \nabla_\mathbf{k}\mathbf{b}).$$

Since $[\mathbf{k}, \mathbf{a}] = \nabla_\mathbf{k}\mathbf{a} - \nabla_\mathbf{a}\mathbf{k}$, and similarly for $[\mathbf{k}, \mathbf{b}]$, we obtain Eq. (1.48). ■

A vector field $\mathbf{v}$ such that $\mathcal{L}_\mathbf{v} g = 0$ is called a **Killing vector** for the metric g. The Killing vector describes a special set of directions in space, such that the metric "remains constant" in these directions. It is only due to the presence of a certain symmetry that a metric could have even a single Killing vector field. (By contrast, the Levi-Civita connection gives $\nabla_\mathbf{v} g = 0$ for any vector field $\mathbf{v}$, but this is due to a definition of ∇ rather than a special property of the metric g. Recall that the Lie derivative is defined independently of the metric.)

Example 1.6.7.2: Consider an 3-dimensional space with coordinates $(x^1, ..., x^3) \equiv (x, y, z)$ and the metric

$$g = \sum_{\alpha,\beta=1}^{N} A_{\alpha\beta}(x,z)\mathrm{d}x^\alpha \otimes \mathrm{d}x^\beta,$$

where $A_{\alpha\beta}(x,z)$ are some complicated functions of x and z (but not of y). Then it is intuitively clear that the metric g is "constant in the direction y." The mathematical formulation of this property is that $\mathbf{v} = \partial_y$ is the Killing vector for g.

We can verify this property explicitly. The value of $(\mathcal{L}_\mathbf{v} g) \circ (\mathbf{a}, \mathbf{b})$ depends only on the values of $\mathbf{a}$ and $\mathbf{b}$ at each point p, but not on derivatives of $\mathbf{a}$ and $\mathbf{b}$. To simplify the calculations, let us choose the vector fields $\mathbf{a}$ and $\mathbf{b}$ such that

$$[\mathbf{v}, \mathbf{a}] = 0, \quad [\mathbf{v}, \mathbf{b}] = 0$$

at a point p. In terms of the components a^μ and b^μ, these conditions mean that

$$\left.\frac{\partial a^\mu}{\partial y}\right|_p = 0, \quad \left.\frac{\partial b^\mu}{\partial y}\right|_p = 0$$

at a point p. It is always possible to adjust the derivatives of a^μ and b^μ at any single point p, without changing the values of the vectors at that point. Then we find, at the same point p,

$$\begin{aligned}(\mathcal{L}_\mathbf{v} g) \circ (\mathbf{a}, \mathbf{b}) &= \mathbf{v} \circ g(\mathbf{a}, \mathbf{b}) - g([\mathbf{v}, \mathbf{a}], \mathbf{b}) - g(\mathbf{a}, [\mathbf{v}, \mathbf{b}]) \\ &= \partial_y g(\mathbf{a}, \mathbf{b}) = \partial_y \left[\sum_{\alpha,\beta} A_{\alpha\beta}(x,z) a^\alpha b^\beta\right]_p = 0.\end{aligned}$$

The same calculation goes through at every point p. Thus ∂_y is a Killing vector for g. ■

Example 1.6.7.3: Consider the Schwarzschild metric

$$g = \left(1 - \frac{2M}{r}\right) dt^2 - \left(1 - \frac{2M}{r}\right)^{-1} dr^2 - r^2 \left(d\theta^2 + (\sin^2\theta) d\phi^2\right).$$

This metric has the Killing vectors ∂_t and ∂_ϕ. One describes the corresponding symmetries of the spacetime by saying that the geometry is stationary (independent of time) and azimuthally symmetric (independent of the azimuthal angle ϕ). Note that these vectors commute, $[\partial_t, \partial_\phi] = 0$, meaning that the symmetries are independent of each other. Of course, any linear combination of ∂_t and ∂_ϕ (with constant coefficients) is also a Killing vector for g. ■

Practice problem: Produce an explicit example showing that in general $\mathcal{L}_\mathbf{v} g \neq 0$ for a vector field $\mathbf{v}$ and a metric g. ■

Using Eq. (1.48), the Killing vector property $\mathcal{L}_\mathbf{k} g = 0$ can be rewritten in a more convenient way, as a differential equation for the vector $\mathbf{k}$. This equation is called the **Killing equation**,

$$g(\nabla_\mathbf{a} \mathbf{k}, \mathbf{b}) + g(\mathbf{a}, \nabla_\mathbf{b} \mathbf{k}) = 0, \tag{1.50}$$

and is understood as an identity that should hold for all vectors $\mathbf{a}, \mathbf{b}$.

Remark: In the index notation, the Killing equation (1.50) is

$$\nabla_\mu k_\nu + \nabla_\nu k_\mu = 0 \tag{1.51}$$

and involves the *covariant* components of the Killing vector $\mathbf{k}$. ■

For an arbitrary vector field $\mathbf{x}$, we may consider the bilinear form $B_{(\mathbf{x})}$ defined by

$$B_{(\mathbf{x})}(\mathbf{a}, \mathbf{b}) \equiv g(\nabla_\mathbf{a} \mathbf{x}, \mathbf{b}).$$

We call $B_{(\mathbf{x})}$ the **distortion tensor** corresponding to the vector field $\mathbf{x}$. In the index notation, the distortion tensor is written as

$$\left[B_{(\mathbf{x})}\right]_{\mu\nu} = \nabla_\mu \mathbf{x}_\nu.$$

The Killing equation (1.50) can be easily restated in terms of the distorsion tensor

$$B_{(\mathbf{k})}(\mathbf{a}, \mathbf{b}) \equiv g(\nabla_\mathbf{a} \mathbf{k}, \mathbf{b}) \tag{1.52}$$

of the vector $\mathbf{k}$. Namely, $\mathbf{k}$ is a Killing vector iff its distorsion tensor $B_{(\mathbf{k})}$ is *antisymmetric*, i.e. $B_{(\mathbf{k})}(\mathbf{a}, \mathbf{b}) = -B_{(\mathbf{k})}(\mathbf{b}, \mathbf{a})$. In other words, $B_{(\mathbf{k})}(\mathbf{a}, \mathbf{b})$ is a 2-form if $\mathbf{k}$ is a Killing vector. The following problem derives an explicit representation of this 2-form.

Practice problem: Using Eqs. (1.45) and (1.49), show that

$$B_{(\mathbf{k})}(\mathbf{a}, \mathbf{b}) \equiv g(\nabla_\mathbf{a} \mathbf{k}, \mathbf{b}) = \frac{\mathbf{a} \circ g(\mathbf{k}, \mathbf{b}) - \mathbf{b} \circ g(\mathbf{k}, \mathbf{a}) - g(\mathbf{k}, [\mathbf{a}, \mathbf{b}])}{2}$$

if $\mathbf{k}$ is a Killing vector. Then prove that the 2-form $B_{(\mathbf{k})}$ is equivalently represented as

$$B_{(\mathbf{k})} = \frac{1}{2} \mathrm{d}\left(\hat{g}\mathbf{k}\right),$$

where d is the exterior differential and $\hat{g}\mathbf{k}$ is the 1-form defined by $(\hat{g}\mathbf{k}) \circ \mathbf{v} \equiv g(\mathbf{k}, \mathbf{v})$.

Sketch of a solution: Three of the six terms in Eq. (1.45) cancel because of Eq. (1.49). The rest follows by using the formula (1.23). ■

1.6.8 *Koszul formula and the Lie derivative

In Sec. 1.6.6 we have derived the Koszul formula (1.45) by "brute force," starting from the expression $g(\nabla_\mathbf{x} \mathbf{y}, \mathbf{z})$ and using the defining properties of the Levi-Civita connection ∇. A shorter derivation and a more elegant-looking formula can be found using the following trick.

We will have a formula for ∇ if we can compute the distortion tensor $B_{(\mathbf{x})}$ of an arbitrary vector field $\mathbf{x}$. The trick is to consider the symmetric and the antisymmetric parts of $B_{(\mathbf{x})}$ separately. Let us first consider the symmetric part,

$$\begin{aligned}B_{(\mathbf{x})}(\mathbf{a}, \mathbf{b}) + B_{(\mathbf{x})}(\mathbf{b}, \mathbf{a}) &= g(\nabla_\mathbf{a} \mathbf{x}, \mathbf{b}) + g(\nabla_\mathbf{b} \mathbf{x}, \mathbf{a}) \\ &= g(\nabla_\mathbf{x} \mathbf{a} - \mathcal{L}_\mathbf{x} \mathbf{a}, \mathbf{b}) + g(\mathbf{a}, \nabla_\mathbf{x} \mathbf{b} - \mathcal{L}_\mathbf{x} \mathbf{b}) \\ &= \mathcal{L}_\mathbf{x} g(\mathbf{a}, \mathbf{b}) - g(\mathcal{L}_\mathbf{x} \mathbf{a}, \mathbf{b}) - g(\mathbf{a}, \mathcal{L}_\mathbf{x} \mathbf{b}) \\ &= (\mathcal{L}_\mathbf{x} g) \circ (\mathbf{a}, \mathbf{b}).\end{aligned}$$

Now consider the antisymmetric part,

$$\begin{aligned}B_{(\mathbf{x})}(\mathbf{a}, \mathbf{b}) - B_{(\mathbf{x})}(\mathbf{b}, \mathbf{a}) &= g(\nabla_\mathbf{a} \mathbf{x}, \mathbf{b}) - g(\nabla_\mathbf{b} \mathbf{x}, \mathbf{a}) \\ &= \nabla_\mathbf{a} g(\mathbf{x}, \mathbf{b}) - g(\mathbf{x}, \nabla_\mathbf{a} \mathbf{b}) - \nabla_\mathbf{b} g(\mathbf{x}, \mathbf{a}) + g(\mathbf{x}, \nabla_\mathbf{b} \mathbf{a}) \\ &= \mathbf{a} \circ g(\mathbf{x}, \mathbf{b}) - \mathbf{b} \circ g(\mathbf{x}, \mathbf{a}) - g(\mathbf{x}, [\mathbf{a}, \mathbf{b}]).\end{aligned}$$

Comparing with Eq. (1.23) and denoting by $\hat{g}\mathbf{x}$ the 1-form

$$(\hat{g}\mathbf{x}) \circ \mathbf{a} \equiv g(\mathbf{x}, \mathbf{a}),$$

we find

$$B_{(\mathbf{x})}(\mathbf{a},\mathbf{b}) - B_{(\mathbf{x})}(\mathbf{b},\mathbf{a}) = (\mathrm{d}\hat{g}\mathbf{x}) \circ (\mathbf{a},\mathbf{b}) .$$

Finally, we restore $B_{(\mathbf{x})}$ as a half-sum of its symmetric and antisymmetric parts. Thus, we can rewrite the Koszul formula more concisely as follows,

$$B_{(\mathbf{x})} = \frac{1}{2}\mathrm{d}\left(\hat{g}\mathbf{x}\right) + \frac{1}{2}\mathcal{L}_{\mathbf{x}}g. \tag{1.53}$$

Now it becomes clear why the covariant derivative of a Killing vector $\mathbf{k}$ yields an antisymmetric 2-form $B_{(\mathbf{k})} = \frac{1}{2}\mathrm{d}\hat{g}\mathbf{k}$.

Remark: In the index notation, the formula (1.53) is a nontrivial reinterpretation of the trivial identity

$$\nabla_\mu k_\nu = \frac{1}{2}\left(\nabla_\mu k_\nu - \nabla_\nu k_\mu\right) + \frac{1}{2}\left(\nabla_\mu k_\nu + \nabla_\nu k_\mu\right).$$

The first term is equal to the components of the exterior differential of the 1-form k_μ (the identity $\nabla_\mu k_\nu - \nabla_\nu k_\mu = \partial_\mu k_\nu - \partial_\nu k_\mu$ can be demonstrated explicitly using the Christoffel symbols), while the second term contains the same terms as the Killing equation (1.51). ■

Practice problem: Derive a generalization of the formula (1.53) in case of a connection ∇ that has a given nonzero torsion tensor $T(\mathbf{u},\mathbf{v})$ but is still compatible with the metric g.

Hint: Use Eq. (1.43) and follow the derivation of Eq. (1.53).

Answer:

$$\begin{aligned} 2B_{(\mathbf{x})}(\mathbf{a},\mathbf{b}) &= \left[\mathrm{d}\left(\hat{g}\mathbf{x}\right) + \mathcal{L}_{\mathbf{x}}g\right] \circ (\mathbf{a},\mathbf{b}) \\ &\quad + g(T(\mathbf{x},\mathbf{a}),\mathbf{b}) + g(T(\mathbf{x},\mathbf{b}),\mathbf{a}) - g(T(\mathbf{a},\mathbf{b}),\mathbf{x}). \end{aligned}$$

■

1.6.9 Divergence of a vector field

The divergence of a vector field $\mathbf{u}$ is a number characterizing the change in the volume of space carried by the flow of $\mathbf{u}$.

To be definite, let us consider a *four*-dimensional manifold $\mathcal{M}$. Let $\mathbf{u}$ be a given vector field on $\mathcal{M}$, and consider the orbits of $\mathbf{u}$ as curves parameterized by a parameter τ. We may select a region $\mathcal{V} \subset \mathcal{M}$ at an initial value $\tau = 0$, and let the flow of $\mathbf{u}$ transform this region to a different region $\mathcal{V}(\tau)$. Such a τ-dependent region of space is called **comoving** with the flow of $\mathbf{u}$.

It is more convenient to consider an "infinitesimal" volume, or a **volume element**. A 4-volume element of the comoving volume can be thought of as a parallelepiped spanned by the tangent vector $\mathbf{u}|_p$ and by three vectors connecting the point p to nearby points on the congruence corresponding to the same value of the curve parameter τ. Thus we need to choose three *connecting vectors* $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ for the field $\mathbf{u}$. Then we would like to compute the rate of change of the volume element $V \equiv \varepsilon(\mathbf{u},\mathbf{a},\mathbf{b},\mathbf{c})$, where ε is the Levi-Civita symbol. The result will be the derivative of V along the flow, that is, $\mathcal{L}_{\mathbf{u}}V$.

If the field $\mathbf{u}$ is everywhere timelike and normalized by the condition $g(\mathbf{u},\mathbf{u}) = 1$, we may interpret τ as the "proper time" measured by observers along the orbits. Then the connecting vectors will correspond to "comoving" spatial directions measured at a fixed time τ in the reference frame of an observer. So the derivative $\mathcal{L}_{\mathbf{u}}V$ will be interpreted as the local rate of change of the comoving 3-volume of space. (The 3-volume is completed to a 4-volume by the unit vector $\mathbf{u}$.) However, for the present calculation we do not need to assume that $\mathbf{u}$ is timelike.

Since

$$\mathcal{L}_{\mathbf{u}}\mathbf{a} = \mathcal{L}_{\mathbf{u}}\mathbf{b} = \mathcal{L}_{\mathbf{u}}\mathbf{c} = \mathcal{L}_{\mathbf{u}}\mathbf{u} = 0,$$

the quantity $\mathcal{L}_{\mathbf{u}}V$ can be written as

$$\mathcal{L}_{\mathbf{u}}V = \mathcal{L}_{\mathbf{u}}\varepsilon(\mathbf{u},\mathbf{a},\mathbf{b},\mathbf{c}) = (\mathcal{L}_{\mathbf{u}}\varepsilon) \circ (\mathbf{u},\mathbf{a},\mathbf{b},\mathbf{c}).$$

The tensor $\mathcal{L}_{\mathbf{u}}\varepsilon$ is a totally antisymmetric tensor (4-form), and so it must be proportional to ε itself. Hence we must have $\mathcal{L}_{\mathbf{u}}\varepsilon = f\varepsilon$, where f is some scalar function. This function depends on the field $\mathbf{u}$ and is called the **divergence** or the **expansion** of $\mathbf{u}$, denoted $\mathrm{div}\mathbf{u}$.

Statement 1.6.9.1: The quantity $\mathrm{div}\mathbf{u}$ can be expressed through the Levi-Civita connection ∇ and an orthonormal frame $\{\mathbf{e}_a\}$, such that $g(\mathbf{e}_a,\mathbf{e}_b) = \eta_{ab}$, by the formulas

$$\begin{aligned} \mathrm{div}\mathbf{u} &= \sum_a \eta_{aa} g(\mathbf{e}_a, \nabla_{\mathbf{e}_a}\mathbf{u}) \\ &= \sum_a \eta_{aa} g(\mathbf{e}_a, [\mathbf{e}_a,\mathbf{u}]). \end{aligned}$$

In the index notation, the divergence can be expressed as

$$\mathrm{div}\mathbf{u} = \nabla_\mu u^\mu \equiv u^\mu{}_{;\mu}.$$

Proof of Statement 1.6.9.1: Denote by $\{\theta^a\}$ the orthonormal basis of 1-forms dual to $\{\mathbf{e}_a\}$. We can then compute

$$\begin{aligned} \mathcal{L}_{\mathbf{u}}\varepsilon &= \mathcal{L}_{\mathbf{u}}\left(\theta^0\wedge\theta^1\wedge\theta^2\wedge\theta^3\right) \\ &= \left(\mathcal{L}_{\mathbf{u}}\theta^0\right)\wedge\theta^1\wedge\theta^2\wedge\theta^3 + \theta^0\wedge\left(\mathcal{L}_{\mathbf{u}}\theta^1\right)\wedge\theta^2\wedge\theta^3 + ..., \end{aligned}$$

where we omitted similar terms involving θ^2 and θ^3. We now consider the first term above. The 1-form $\mathcal{L}_{\mathbf{u}}\theta^0$ can be expressed as a linear combination of the basis 1-forms $\{\theta^a\}$,

$$\mathcal{L}_{\mathbf{u}}\theta^0 = \sum_a \theta^a \left(\mathcal{L}_{\mathbf{u}}\theta^0\right) \circ \mathbf{e}_a.$$

Due to antisymmetry of $\wedge$ within $\left(\mathcal{L}_{\mathbf{u}}\theta^0\right)\wedge\theta^1\wedge\theta^2\wedge\theta^3$, only the term proportional to θ^0 survives, so

$$\left(\mathcal{L}_{\mathbf{u}}\theta^0\right)\wedge\theta^1\wedge\theta^2\wedge\theta^3 = \left(\left(\mathcal{L}_{\mathbf{u}}\theta^0\right)\circ\mathbf{e}_0\right)\theta^0\wedge\theta^1\wedge\theta^2\wedge\theta^3.$$

Since $\theta^0 \circ \mathbf{e}_0 = 1$, we can simplify $\left(\mathcal{L}_{\mathbf{u}}\theta^0\right)\circ\mathbf{e}_0$ as follows,

$$\begin{aligned} \left(\mathcal{L}_{\mathbf{u}}\theta^0\right)\circ\mathbf{e}_0 &= \mathcal{L}_{\mathbf{u}}\left(\theta^0\circ\mathbf{e}_0\right) - \theta^0\circ\mathcal{L}_{\mathbf{u}}\mathbf{e}_0 = \theta^0\circ\left(\nabla_{\mathbf{e}_0}\mathbf{u} - \nabla_{\mathbf{u}}\mathbf{e}_0\right) \\ &= \eta_{00}g(\mathbf{e}_0, \nabla_{\mathbf{e}_0}\mathbf{u} - \nabla_{\mathbf{u}}\mathbf{e}_0) = \eta_{00}g(\mathbf{e}_0, \nabla_{\mathbf{e}_0}\mathbf{u}), \end{aligned}$$

where we used the property

$$g(\mathbf{e}_0, [\mathbf{e}_0,\mathbf{u}]) = g(\mathbf{e}_0, \nabla_{\mathbf{e}_0}\mathbf{u}),$$

which is due to

$$g(\mathbf{e}_0,\mathbf{e}_0) = \mathrm{const}, \quad g(\mathbf{e}_0, \nabla_{\mathbf{u}}\mathbf{e}_0) = \frac{1}{2}\nabla_{\mathbf{u}}g(\mathbf{e}_0,\mathbf{e}_0) = 0.$$

Thus

$$\left(\mathcal{L}_{\mathbf{u}}\theta^0\right)\wedge\theta^1\wedge\theta^2\wedge\theta^3 = \eta_{00}g(\mathbf{e}_0, \nabla_{\mathbf{e}_0}\mathbf{u})\varepsilon.$$

Analogous simplifications are performed for other terms, and so we obtain the first required formula. Note that

$$g(\mathbf{e}_a, \nabla_{\mathbf{u}} \mathbf{e}_a) = 0$$

due to the normalization of $\mathbf{e}_a$. Thus

$$g(\mathbf{e}_a, \nabla_{\mathbf{e}_a} \mathbf{u}) = g(\mathbf{e}_a, [\mathbf{e}_a, \mathbf{u}]).$$

The formula for div$\mathbf{u}$ in the index notation can be now derived by expressing the orthonormal basis $\{\mathbf{e}_a\}$ through the coordinate basis $\{\partial_\mu\}$. Suppose that e_a^μ are the components of the orthonormal frame, so that

$$\mathbf{e}_a = e_a^\mu \partial_\mu.$$

Then in the index notation we have

$$g(\mathbf{e}_a, \nabla_{\mathbf{e}_a} \mathbf{u}) = g_{\lambda\mu} e_a^\mu e_a^\nu u^\lambda{}_{;\nu}.$$

Thus

$$\text{div}\mathbf{u} = \sum_a \eta_{aa} g_{\lambda\mu} e_a^\mu e_a^\nu u^\lambda{}_{;\nu}.$$

On the other hand, Eq. (1.40) gives

$$\sum_a \eta_{aa} e_a^\mu e_a^\nu = g^{\mu\nu}.$$

Hence, $\text{div}\mathbf{u} = g_{\lambda\mu} g^{\mu\nu} u^\lambda{}_{;\nu} = u^\mu{}_{;\mu}$. ■

Since $\mathcal{L}_{\mathbf{u}} \boldsymbol{\varepsilon} = (\text{div}\mathbf{u})\boldsymbol{\varepsilon}$, our final result is

$$\mathcal{L}_{\mathbf{u}} V = (\text{div}\mathbf{u})\, \varepsilon(\mathbf{u}, \mathbf{a}, \mathbf{b}, \mathbf{c}) = (\text{div}\mathbf{u})\, V.$$

Thus the quantity div$\mathbf{u}$ is the *relative* rate of change of the volume carried by the flow of $\mathbf{u}$.

The divergence of a vector field can be also expressed in terms of the Hodge star operation.

Statement 1.6.9.2: If η_{ab} is the canonical form of the metric, the divergence of a vector field $\mathbf{u}$ is

$$\text{div}\mathbf{u} = (\det \eta_{ab}) * \mathrm{d} * \hat{g}\mathbf{u}.$$

Remark: This expression for the divergence will be rarely useful in our calculations.

Proof of Statement 1.6.9.2: Since div$\mathbf{u}$ is a scalar, we have $*\text{div}\mathbf{u} = \boldsymbol{\varepsilon}\text{div}\mathbf{u}$, so it is sufficient to show that

$$\boldsymbol{\varepsilon}\text{div}\mathbf{u} = \mathrm{d} * \hat{g}\mathbf{u}.$$

Since $\hat{g}\mathbf{u}$ is a 1-form, we have

$$*\hat{g}\mathbf{u} = \iota_{\mathbf{u}} \boldsymbol{\varepsilon}.$$

Using the property $\mathrm{d}\boldsymbol{\varepsilon} = 0$ and the Cartan homotopy formula (1.24), we compute

$$\mathrm{d} * \hat{g}\mathbf{u} = \mathrm{d}\iota_{\mathbf{u}} \boldsymbol{\varepsilon} = (\mathcal{L}_{\mathbf{u}} - \iota_{\mathbf{u}} \mathrm{d})\, \boldsymbol{\varepsilon} = \mathcal{L}_{\mathbf{u}} \boldsymbol{\varepsilon} \equiv (\text{div}\mathbf{u})\, \boldsymbol{\varepsilon}.$$

■

1.7 Calculations in index-free notation

The index-free notation has its advantages and disadvantages, in comparison with the index notation.

- Index-free notation emphasizes geometric operations, such as the commutator $[\mathbf{a}, \mathbf{b}]$ or scalar product $g(\mathbf{a}, \mathbf{b})$, that would otherwise remain hidden under a mass of indices. For instance, the Bianchi identities are clearly seen to be consequences of the Jacobi identity and the assumptions about the Levi-Civita connection. Also, the geometric relations between tensors are shown explicitly, rather than encoded in the positions of indices. Thus, calculations in index-free notation help develop geometric intuition rather than merely an agility with manipulating indices. In the index notation, the same calculations appear to consist of a certain lucky manipulations of indices (the so-called "juggling of indices"). It remains unclear how to guess the correct sequence of index manipulations that would be needed for a particular calculation.

- Index-free notation is in many cases more concise than the index notation, since one does not need to write indices next to each letter.

- In the index-free approach, every object is defined through geometric operations, so it is impossible to introduce a non-tensorial quantity. For instance, the non-tensorial Christoffel symbol $\Gamma^\lambda_{\alpha\beta}$ in its usual form (1.46) is simply *undefined*; instead one might define a Christoffel *tensor* Γ, which is a transformation-valued 1-form representing the difference between two given affine connections $\tilde{\nabla}$ and ∇,

  $$\Gamma(\mathbf{u})\mathbf{v} \equiv \tilde{\nabla}_{\mathbf{u}} \mathbf{v} - \nabla_{\mathbf{u}} \mathbf{v}.$$

  Within the index notation, one frequently uses non-tensor quantities and calculates their components in a convenient coordinate system, and then one needs to check that the results are tensors. Operations with non-tensor quantities obscure the logic of a derivation and may lead to hard-to-spot errors.

- Index-free notation appears more abstract, whereas expressions in the index notation can be easily interpreted as just "arrays of numbers."

- Index-free notation is unwieldy if complicated contractions or symmetrizations are applied to tensors of high rank. For instance, an expression such as $A^{[\rho}_{\mu\nu;\rho} B^{\beta]\mu\nu}_{\alpha}$ is difficult to manipulate in the index-free notation.

On the whole, it appears that the index-free notation is more suitable for "generic" or "abstract" calculations, such as definitions of new quantities or derivations of general properties of tensors. In calculations involving complicated tensor contractions and symmetrizations, the index notation is better. For specific calculations, e.g. when components of a certain tensor are known in particular coordinates, the index notation is unavoidable.

Perhaps, the index notation should be studied first, because a certain familiarity with the index notation seems to be helpful for learning more abstract material. However, I feel that learning to think in the index-free notation is also helpful, if only for the fact that the entire mathematical literature has been using exclusively index-free notation for several decades.

1.7.1 Abstract index notation

It is somewhat cumbersome to use the index-free notation when dealing with tensors of higher rank. For example, suppose A is a transformation-valued one-form such that $A(\mathbf{v})$ is a linear transformation in the tangent space, i.e. $A(\mathbf{v})\mathbf{w}$ is a vector. Then ∇A is a transformation-valued bilinear form that acts as

$$(\nabla A)(\mathbf{u},\mathbf{v})\mathbf{w} \equiv [\nabla_{\mathbf{u}} A]\,(\mathbf{v})\mathbf{w}.$$

Note that we have some difficulty in making it clear that $(\nabla A)(\mathbf{u},\mathbf{v})$ is a derivative in the direction of $\mathbf{u}$ rather than $\mathbf{v}$, and that only A is to be differentiated but not $\mathbf{v}$ or $\mathbf{w}$. We could write

$$\text{“}(\nabla A)(\mathbf{u},\mathbf{v})\mathbf{w}\text{”} \equiv \nabla_{\mathbf{u}}\left[A(\mathbf{v})\mathbf{w}\right] - A(\nabla_{\mathbf{u}}\mathbf{v})\mathbf{w} - A(\mathbf{v})\nabla_{\mathbf{u}}\mathbf{w},$$

but manipulating such expressions is a rather cumbersome affair. In such cases, it is more convenient to write everything in the index notation, e.g.

$$\text{“}(\nabla A)(\mathbf{u},\mathbf{v})\mathbf{w}\text{”} \equiv u^\mu v^\lambda w^\alpha \nabla_\mu A^\beta_{\alpha\lambda} \equiv u^\mu v^\lambda w^\alpha A^\beta_{\alpha\lambda;\mu}.$$

This notation clearly shows that the vector $\mathbf{u}$ supplies the direction for the derivative, and the vectors $\mathbf{v}$ and $\mathbf{w}$ are contracted with the tensor A, but only A is differentiated. However, when using the index notation we do not actually need to talk about components of the vectors in a basis; in fact, we have not really chosen any explicit basis so far. In most cases, the indices are actually used only as labels indicating that vectors and tensors are contracted with each other in a particular order. This is the idea of the **abstract index notation** introduced by R. Penrose as a re-interpretation of the familiar component notation.[9]

The abstract index notation can be summarized as follows: An index, such as $^\alpha$ in v^α and u_α, does not take values $0, 1, 2, ...$, but is merely a label indicating that v is a vector while u is a 1-form. Repeated indices, e.g. $v^\alpha u_\alpha$, do not mean summation but instead indicate that a certain vector is being contracted with a certain 1-form. No particular basis or coordinate system is chosen or implied, and one is careful to perform only well-defined geometric operations on all quantities. (Geometric operations are the tensor product, the contraction of tensors, the Lie derivative, and the covariant derivative. The ordinary coordinate derivative $\partial/\partial x^\mu$ is disallowed.) The results of such calculations may be interpreted either as coordinate-free expressions or as component expressions “valid in any coordinate system.” Strictly speaking, the abstract index notation disallows non-tensorial quantities such as the Christoffel symbol $\Gamma^\lambda_{\alpha\beta}$, operations with individual components of tensors such as $R_{\alpha 4\beta 4}$, or an implicit choice of a coordinate system (e.g. calculations in locally inertial coordinates).

The approach taken in this book is coordinate-free and the notation is almost everywhere index-free. However, we *do* use the abstract index notation when it is significantly more convenient. For instance, it is straightforward to express complicated tensor contractions such as $A^\rho_{\mu\nu;\rho}B^{\mu\nu}_\alpha$ in the index notation, but cumbersome without indices.

[9] Penrose wrote in [25]: *“What I shall present here is an entirely frame-independent algebra which allows one to calculate with indexed quantities exactly as before (but now with a clear conscience!)...”* The “clear conscience” referred to is the absence of a cordinate system. However, this “conscience” is not really “clear” if one still uses non-tensor quantities such as $\Gamma^\lambda_{\alpha\beta}$ or performs calculations in specially chosen coordinate systems.

Remark: I would like to emphasize that the index notation such as $\nabla_\mu A^\beta_{\alpha\lambda} \equiv A^\beta_{\alpha\lambda;\mu}$ does not imply that the covariant derivative is applied to each component of A separately. Rather, this notation refers to the components of the higher-rank tensor (∇A). For instance, the component $\nabla_1 A^1_{12}$ is not equal to an operator “∇_1” acting on the component A^1_{12} of the tensor A. Rather, the notation $\nabla_1 A^1_{12}$ stands for the component $(\nabla A)^1_{112}$ of the tensor ∇A. If (for example) $A^\beta_{\alpha\lambda} = -A^\beta_{\lambda\alpha}$ then the “obvious” property $A^\beta_{\alpha\lambda;\mu} = -A^\beta_{\lambda\alpha;\mu}$ is not automatically satisfied (as it would be if $A^\beta_{\alpha\lambda;\mu}$ were a component-wise derivative of $A^\beta_{\alpha\lambda}$). This property must be derived as a consequence of the properties of ∇. The symbol ∇_μ is not a collection of fixed tensors $\nabla_0, \nabla_1, \nabla_2, \nabla_3$; rather, ∇_μ represents different operations when applied to different tensors. Therefore, an abstract index expression such as $A^\beta_{\alpha\lambda;\mu}$ must be interpreted and used with care.

1.7.2 Converting expressions into index-free notation

In this section we develop the methods for converting indexed expressions into an index-free form.

The correspondence between the index notation and the index-free notation is established using the basis vectors $\mathbf{e}_\alpha \equiv \partial/\partial x^\alpha$, by assuming *fixed* values of all the indices. For example, consider a covector (1-form) $\mathbf{A}$. It is represented in the index notation by the symbol A_μ, while $\nabla_\alpha A_\beta$ means the (α,β)-component of the rank (0,2) tensor $\nabla\mathbf{A}$. For fixed α and β, the component $\nabla_\alpha A_\beta$ is the number defined by

$$A_{\beta;\alpha} \equiv \nabla_\alpha A_\beta \equiv (\nabla\mathbf{A})_{\alpha\beta} = (\nabla_{\mathbf{e}_\alpha}\mathbf{A}) \circ \mathbf{e}_\beta.$$

An equation written in the index notation, such as

$$A_{\beta;\alpha} + g_{\alpha\beta} - 2A_{\alpha;\beta} = 0, \tag{1.54}$$

is converted to the index-free notation by first rewriting it as

$$\nabla_\alpha A_\beta + g_{\alpha\beta} - 2\nabla_\beta A_\alpha,$$

where we introduced the symbol ∇ explicitly, and then by fixing α, β and inserting basis vectors $\{\mathbf{e}_\alpha\}$, thus obtaining

$$(\nabla_{\mathbf{e}_\alpha}\mathbf{A}) \circ \mathbf{e}_\beta + g\left(\mathbf{e}_\alpha, \mathbf{e}_\beta\right) - 2\left(\nabla_{\mathbf{e}_\beta}\mathbf{A}\right) \circ \mathbf{e}_\alpha = 0.$$

However, we expect that such tensor relations can be expressed *in a geometric way*, i.e. as statements about arbitrary vectors, without using a particular basis $\{\mathbf{e}_\alpha\}$. Indeed, this can be done with a little work.

The basic rule of converting indexed expressions to the index-free notation is to contract each free index with an *arbitrary vector*. The resulting expression will then depend on a certain number of arbitrary vectors but will be independent of the choice of basis. Consider Eq. (1.54) as an example. Since Eq. (1.54) has two free indices (α and β), let us contract it with arbitrary vectors $\mathbf{a}$ and $\mathbf{b}$ having index representations a^α and b^β:

$$a^\alpha b^\beta A_{\beta;\alpha} + a^\alpha b^\beta g_{\alpha\beta} - 2a^\alpha b^\beta A_{\alpha;\beta} = 0.$$

Then we rewrite covariant derivatives explicitly through the symbol ∇:

$$a^\alpha b^\beta \nabla_\alpha A_\beta + a^\alpha b^\beta g_{\alpha\beta} - 2a^\alpha b^\beta \nabla_\beta A_\alpha = 0.$$

Finally, this equation is rewritten in the index-free notation as

$$(\nabla_{\mathbf{a}}\mathbf{A})\circ\mathbf{b}+g(\mathbf{a},\mathbf{b})-2(\nabla_{\mathbf{b}}\mathbf{A})\circ\mathbf{a}=0.$$

Note that $(\nabla_{\mathbf{a}}\mathbf{A})$ is a 1-form acting on vectors. For the purpose of facilitating further calculations with the above expression, it might be useful to rewrite the 1-form $(\nabla_{\mathbf{a}}\mathbf{A})\circ\mathbf{b}$ as

$$\begin{aligned}(\nabla_{\mathbf{a}}\mathbf{A})\circ\mathbf{b}&=\mathbf{a}\circ(\mathbf{A}\circ\mathbf{b})-\mathbf{A}\circ(\nabla_{\mathbf{a}}\mathbf{b})\\&=\mathbf{a}\circ(\mathbf{A}\circ\mathbf{b})-\mathbf{A}\circ([\mathbf{a},\mathbf{b}]+\nabla_{\mathbf{b}}\mathbf{a})\end{aligned}$$

or in some other form. In any case, we now have an equation involving *arbitrary* vectors $\mathbf{a},\mathbf{b}$ and the known 1-form $\mathbf{A}$, and we are not using any basis vectors.

A somewhat more complicated example is an indexed expression involving repeated covariant derivatives, such as

$$v^{\lambda}{}_{;\beta\alpha}-v^{\lambda}{}_{;\alpha\beta}=\nabla_{\alpha}\nabla_{\beta}v^{\lambda}-\nabla_{\beta}\nabla_{\alpha}v^{\lambda}=0. \tag{1.55}$$

Contracting Eq. (1.55) with arbitrary vectors a^{α} and b^{β}, we find

$$a^{\alpha}b^{\beta}\nabla_{\alpha}\nabla_{\beta}v^{\lambda}-a^{\alpha}b^{\beta}\nabla_{\beta}\nabla_{\alpha}v^{\lambda}=0. \tag{1.56}$$

We would like to rewrite this expression in the index-free notation. But now we encounter a difficulty: Nested covariant derivatives, such as $\nabla_{\mathbf{a}}\nabla_{\mathbf{b}}\mathbf{v}$, contain also derivatives of $\mathbf{b}$, which are not present in the expression (1.56). Therefore we first rewrite

$$\begin{aligned}a^{\alpha}b^{\beta}\nabla_{\alpha}\nabla_{\beta}v^{\lambda}&=a^{\alpha}\nabla_{\alpha}\left(b^{\beta}\nabla_{\beta}v^{\lambda}\right)-\left(a^{\alpha}\nabla_{\alpha}b^{\beta}\right)\left(\nabla_{\beta}v^{\lambda}\right)\\&=\nabla_{\mathbf{a}}\nabla_{\mathbf{b}}\mathbf{v}-\nabla_{(\nabla_{\mathbf{a}}\mathbf{b})}\mathbf{v},\end{aligned}$$

and similarly for the other term in Eq. (1.56), and then finally Eq. (1.55) becomes

$$\begin{aligned}&\nabla_{\mathbf{a}}\nabla_{\mathbf{b}}\mathbf{v}-\nabla_{(\nabla_{\mathbf{a}}\mathbf{b})}\mathbf{v}-\nabla_{\mathbf{b}}\nabla_{\mathbf{a}}\mathbf{v}-\nabla_{(\nabla_{\mathbf{b}}\mathbf{a})}\mathbf{v}\\&=[\nabla_{\mathbf{a}},\nabla_{\mathbf{b}}]\,\mathbf{v}-\nabla_{[\mathbf{a},\mathbf{b}]}\mathbf{v}=0.\end{aligned}$$

In the last line, the first commutator is simply a shorthand notation,

$$[\nabla_{\mathbf{a}},\nabla_{\mathbf{b}}]\,\mathbf{v}\equiv\nabla_{\mathbf{a}}\nabla_{\mathbf{b}}\mathbf{v}-\nabla_{\mathbf{b}}\nabla_{\mathbf{a}}\mathbf{v},$$

while the second commutator $[\mathbf{a},\mathbf{b}]\equiv\mathcal{L}_{\mathbf{a}}\mathbf{b}$ is the commutator of vector fields.

Practice problem: Rewrite the following indexed expressions in an index-free manner:

$$\begin{aligned}u^{\alpha}u^{\beta}h_{;\alpha\beta}&=C_{\alpha\beta}v^{\beta}g^{\alpha\lambda}f_{,\lambda};\\u_{\alpha;\beta}-u_{\beta;\alpha}&=v_{\alpha}w_{\beta}-v_{\beta}w_{\alpha};\\v^{\alpha}u_{\beta;\alpha}-u^{\alpha}v_{\beta;\alpha}&=w^{\alpha}w_{\alpha;\beta};\end{aligned}$$

where $u^{\alpha},v^{\alpha},w^{\alpha}$ are indexed representations of vector fields $\mathbf{u},\mathbf{v},\mathbf{w}$; f and h are scalar functions; and $C_{\alpha\beta}$ is a bilinear form $C(\mathbf{x},\mathbf{y})\equiv C_{\alpha\beta}x^{\alpha}y^{\beta}$.

Answers: $\nabla_{\mathbf{u}}\nabla_{\mathbf{u}}h-\nabla_{\nabla_{\mathbf{u}}\mathbf{u}}h=C(\mathbf{v},\hat{g}^{-1}\mathrm{d}f)$, $\mathrm{d}\hat{g}\mathbf{u}=(\hat{g}\mathbf{v})\wedge(\hat{g}\mathbf{w})$, and $g(\mathbf{x},[\mathbf{v},\mathbf{u}])=g(\mathbf{w},\nabla_{\mathbf{x}}\mathbf{w})$, where $\mathbf{x}$ is arbitrary. ■

Note that if we used coordinate basis vectors ∂_{μ} and ∂_{ν} instead of arbitrary vector fields $\mathbf{a}$ and $\mathbf{b}$, we would have $[\partial_{\mu},\partial_{\nu}]=0$ and the term with the commutator would vanish. Thus we would have obtained a simpler formula

$$\nabla_{\partial_{\mu}}\nabla_{\partial_{\nu}}\mathbf{v}-\nabla_{\partial}{}_{.}\nabla_{\partial_{\mu}}\mathbf{v}=0,$$

which looks quite similar to the indexed expression (1.55). The simplification is due to the fact that the commutator term $\nabla_{[\mathbf{a},\mathbf{b}]}\mathbf{v}$ drops out. But the same simplification can be achieved without choosing the coordinate basis vectors. It suffices to assume that the vectors $\mathbf{a}$ and $\mathbf{b}$ commute. Since our expression is actually independent of the *derivatives* of $\mathbf{a}$ and $\mathbf{b}$, we can always choose commuting vector fields $\mathbf{a}$ and $\mathbf{b}$ in a neighborhood of any one point. This choice will be possible in every case when we convert an indexed expression into an index-free form, since any number of simultaneously commuting vector fields can be chosen at a point (see, for example, Statement 1.2.11.2). We will frequently use this method of simplification.

1.7.3 Index-free computations of trace

Traces of tensors *can* be computed in the index-free notation, although in many cases calculations are easier in the index notation. Nevertheless, it is possible to develop a sufficiently powerful formalism, so that the index notation is *in principle* never needed. In this section I explain the index-free approach to calculating the trace. Since the readers will certainly be more familiar with the index notation for traces, I provide a comparison between indexed and index-free approaches.

I begin by reviewing the index-free definition of the trace of a linear transformation. A transformation T is expressed in a tensor form as

$$T=\mathbf{a}_1\otimes\boldsymbol{\omega}_1+\mathbf{a}_2\otimes\boldsymbol{\omega}_2+..., \tag{1.57}$$

where $\mathbf{a}_j$ are some vectors and $\boldsymbol{\omega}_j$ are some covectors; by definition, any (1,1)-tensor can be decomposed in this way. Then the **trace** of T is

$$\mathrm{Tr}\,T=\boldsymbol{\omega}_1\circ\mathbf{a}_1+\boldsymbol{\omega}_2\circ\mathbf{a}_2+... \tag{1.58}$$

However, this definition is not convenient in more complicated cases. For example, a tensor expressed in the index notation as $T^{\lambda}_{\alpha\beta}$ may be interpreted as a transformation-valued 1-form,

$$\mathbf{x}\to T(\mathbf{a})\mathbf{x},\quad (T(\mathbf{a})\mathbf{x})^{\lambda}\equiv T^{\lambda}_{\alpha\beta}a^{\alpha}x^{\beta}.$$

Suppose we would like to compute the trace of T, expressed in the index notation as $T^{\alpha}_{\alpha\beta}$. According to the definition (1.58), we first need to decompose $T^{\lambda}_{\alpha\beta}$ as a linear combination of tensor products, as in Eq. (1.57). This computation is cumbersome because a large number of tensor product terms will be needed in a decomposition of $T^{\lambda}_{\alpha\beta}$. Also, such explicit decompositions usually require a choice of basis. For example, the identity operator can be decomposed as

$$\hat{1}=\sum_{j=1}^{n}\mathbf{e}_j\otimes\boldsymbol{\theta}^j,$$

where $\{\mathbf{e}_j\}$ is a basis and $\{\boldsymbol{\theta}^j\}$ is the corresponding dual basis. (The above relation is called a **decomposition of identity**.) Calculations with such decompositions rapidly become unwieldy for higher-rank tensors, and so we will proceed by another route.

To be specific, let us first show how one can compute the trace $T^{\alpha}_{\alpha\beta}$ in a basis-free manner. We first "lower the index" to obtain $T_{\alpha\beta\lambda}\equiv T^{\mu}_{\alpha\beta}g_{\lambda\mu}$, which is in the index-free notation a trilinear form $T(\mathbf{a},\mathbf{b},\mathbf{c})$,

$$T(\mathbf{a},\mathbf{b},\mathbf{c})\equiv g(T(\mathbf{a})\mathbf{b},\mathbf{c}).$$

Then we need to specify two of three arguments of T which are to be contracted. For instance, the covector $\omega_\lambda \equiv T^\alpha_{\alpha\lambda} \equiv T_{\alpha\beta\lambda}g^{\alpha\beta}$ looks like the contraction of $T(\mathbf{a},\mathbf{b},\mathbf{c})$ with respect to the vectors $\mathbf{a}$ and $\mathbf{b}$. To make this relationship explicit, we use the following notation,

$$\boldsymbol{\omega}(\mathbf{c}) = \mathrm{Tr}_{\mathbf{a}} T(\mathbf{a},\mathbf{a},\mathbf{c}). \tag{1.59}$$

This notation is somewhat verbose but explicit and unambiguous.The vector $\mathbf{a}$ does not enter the final object $\boldsymbol{\omega}(\mathbf{c})$; so $\mathbf{a}$ is merely an auxiliary part of notation. The vector $\mathbf{a}$ in the right-hand side of Eq. (1.59) may be called a **mute vector**, similarly to the mute index α used in summation.

Thus I propose the following index-free notation for the **trace** of a multilinear tensor-valued function $T(\mathbf{a},\mathbf{b},\mathbf{c},...)$ with respect to the arguments $\mathbf{a},\mathbf{b}$. The tensor $T(\mathbf{a},\mathbf{b},\mathbf{c},...)$ is first written as a linear function of $\mathbf{a}\otimes\mathbf{b}$, i.e. $T(\mathbf{a}\otimes\mathbf{b},\mathbf{c},...)$, and then the trace is found by substituting the inverse metric g^{-1}, which is a (2,0)-tensor, instead of the argument $\mathbf{a}\otimes\mathbf{b}$:

$$\mathrm{Tr}_{\mathbf{a}} T(\mathbf{a},\mathbf{a},\mathbf{c},...) \equiv T(\mathbf{a}\otimes\mathbf{b},\mathbf{c},...)\big|_{\mathbf{a}\otimes\mathbf{b}=g^{-1}}.$$

Note that the use of the same vector $\mathbf{a}$ in both arguments of T reflects the fact that g^{-1} is a symmetric (2,0) tensor. This notation often helps simplify terms that vanish due to antisymmetry. An alternative notation, such as $\mathrm{Tr}_{\langle\mathbf{ab}\rangle} T(\mathbf{a},\mathbf{b},\mathbf{c})$, would be more cumbersome to use.

The index notation for the trace above is $g^{\alpha\beta}T_{\alpha\beta\gamma...}$. In this way, the trace operation is understood as a simple substitution $\mathbf{a}\otimes\mathbf{b} = g^{-1}$ into a multilinear function T. An explicit notation for this substitution,

$$\mathrm{Tr}_{\mathbf{a}} T(\mathbf{a},\mathbf{a},\mathbf{c},...) \equiv \mathrm{Tr}_{\mathbf{a}\otimes\mathbf{b}=g^{-1}} T(\mathbf{a},\mathbf{b},\mathbf{c},...) \tag{1.60}$$

could be useful when one needs to specify explicitly that the metric g is being used. For instance, one could then introduce several metrics g, h, ..., and compute traces with respect to these metrics. However, we will be working most of the time with one metric g, and so the verbose and explicit notation (1.60) will not be necessary.

Using the trace notation, we may now rewrite an arbitrary indexed expression in an index-free form. When an indexed expression contains a sum over a pair of "dummy" indices, we can first lower both indices (inserting an extra $g^{\alpha\beta}$) and then replace the $g^{\alpha\beta}$ by a pair of mute vectors.

Example: Let us rewrite the following equation,

$$u^\alpha u^\beta u^\gamma w_{\gamma;\alpha\beta} = C^\lambda{}_{\alpha\lambda\beta} u^\alpha u^\beta,$$

in the index-free notation. Using the techniques from Sec.1.7.2, we rewrite the left-hand side as

$$g(\mathbf{u}, \nabla_{\mathbf{u}}\nabla_{\mathbf{u}}\mathbf{w} - \nabla_{\nabla_{\mathbf{u}}\mathbf{u}}\mathbf{w}).$$

Turning now to the right-hand side, we note that it is only necessary to express the trace $C^\lambda{}_{\alpha\lambda\beta}$, since all other contractions are ordinary. The tensor $C^\lambda{}_{\alpha\mu\beta}$ can be interpreted as a transformation-valued function of two vectors,

$$\mathbf{z} \to C(\mathbf{x},\mathbf{y})\mathbf{z}, \quad [C(\mathbf{x},\mathbf{y})\mathbf{z}]^\lambda \equiv C^\lambda{}_{\alpha\mu\beta} z^\alpha x^\mu y^\beta.$$

(In an actual calculation, such as that in the practice problem on the facing page below, we may be given an explicit, index-free formula for the transformation $C(\mathbf{x},\mathbf{y})\mathbf{z}$.) The trace $C^\lambda{}_{\alpha\lambda\beta}$ is first converted into a contraction with $g^{\lambda\mu}$,

$$C^\lambda{}_{\alpha\lambda\beta} = g^{\lambda\mu} C_{\lambda\alpha\mu\beta}.$$

The tensor $C_{\lambda\alpha\mu\beta}$ is a function of four vectors, expressed through the original tensor $C^\lambda{}_{\alpha\mu\beta}$ as

$$\begin{aligned} C_{\lambda\alpha\mu\beta} a^\lambda z^\alpha x^\mu y^\beta &= g_{\kappa\lambda} a^\kappa C^\lambda{}_{\alpha\mu\beta} z^\alpha x^\mu y^\beta \\ &\equiv g(\mathbf{a}, C(\mathbf{x},\mathbf{y})\mathbf{z}) \equiv C(\mathbf{a},\mathbf{z},\mathbf{x},\mathbf{y}). \end{aligned}$$

Thus we rewrite

$$C^\lambda{}_{\alpha\lambda\beta} u^\alpha u^\beta = g^{\lambda\mu} C_{\lambda\alpha\mu\beta} u^\alpha u^\beta.$$

Finally, we introduce two mute vectors $\mathbf{a},\mathbf{b}$ whose tensor product $a^\lambda b^\mu$ will replace $g^{\lambda\mu}$ in the contraction $g^{\lambda\mu}C_{\lambda\alpha\mu\beta}$. The result of the replacement is

$$\begin{aligned} g^{\lambda\mu} C_{\lambda\alpha\mu\beta} u^\alpha u^\beta \to a^\lambda b^\mu C_{\lambda\alpha\mu\beta} u^\alpha u^\beta &= C(\mathbf{a},\mathbf{u},\mathbf{b},\mathbf{u}) \\ &= g(\mathbf{a}, C(\mathbf{b},\mathbf{u})\mathbf{u}). \end{aligned}$$

The trace of this expression with respect to the mute vectors $\mathbf{a},\mathbf{b}$ is written as

$$\mathrm{Tr}_{\mathbf{a}} g(\mathbf{a}, C(\mathbf{a},\mathbf{u})\mathbf{u}) = C^\lambda{}_{\alpha\lambda\beta} u^\alpha u^\beta.$$

In this way, we are able to rewrite the initial indexed expression in an index-free manner. ■

One sometimes needs a repeated trace over several sets of indices. In these cases I use the notation

$$\mathbf{Tr}_{\mathbf{a},\mathbf{b}} A(\mathbf{a},\mathbf{a},....,\mathbf{b},\mathbf{b},...) \equiv \mathbf{Tr}_{\mathbf{a}}\mathbf{Tr}_{\mathbf{b}} A(\mathbf{a},\mathbf{a},....,\mathbf{b},\mathbf{b},...).$$

The repeated trace can be computed in any order: $\mathrm{Tr}_{\mathbf{a}}\mathrm{Tr}_{\mathbf{b}} = \mathrm{Tr}_{\mathbf{b}}\mathrm{Tr}_{\mathbf{a}}$. Note that each mute vector $\mathbf{a},\mathbf{b},...$ appears exactly twice in the expression under the trace. This is similar to the Einstein summation convention where each repeated index can appear only twice.

To determine the trace such as that in Eq. (1.59) in an actual calculation, we need to have a concrete expression for the tensor T in terms of other known vectors or tensors, and a procedure to implement the trace operation via an index-free calculation. I now describe such a procedure and then give examples.

Suppose we need to compute the trace of a tensor $T(\mathbf{x},\mathbf{y},\mathbf{z},...)$ with respect to $\mathbf{x},\mathbf{y}$. One possibility to perform this calculation in the index-free notation is to use an explicit decomposition of the inverse metric g^{-1} through tensor products of orthonormal frame $\{\mathbf{e}_0,\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3\}$,

$$g^{-1} = \mathbf{e}_0\otimes\mathbf{e}_0 - \mathbf{e}_1\otimes\mathbf{e}_1 - \mathbf{e}_2\otimes\mathbf{e}_2 - \mathbf{e}_3\otimes\mathbf{e}_3, \tag{1.61}$$

which corresponds to the index notation

$$g^{\alpha\beta} = e^\alpha_0 e^\beta_0 - e^\alpha_1 e^\beta_1 - e^\alpha_2 e^\beta_2 - e^\alpha_3 e^\beta_3.$$

Since $T(\mathbf{x},\mathbf{y},\mathbf{z},...)$ is a bilinear form in $\mathbf{x}$ and $\mathbf{y}$, we can compute the trace $\mathrm{Tr}_{\mathbf{x}} T(\mathbf{x},\mathbf{x},\mathbf{z},...)$ by substituting the decomposition (1.61) of the tensor g^{-1} instead of the arguments $\mathbf{x},\mathbf{y}$ in $T(\mathbf{x},\mathbf{y},\mathbf{z})$. Therefore

$$\mathrm{Tr}_{\mathbf{x}} T(\mathbf{x},\mathbf{x},\mathbf{z},...) = T(\mathbf{e}_0,\mathbf{e}_0,\mathbf{z},...) - \sum_{j=1}^{3} T(\mathbf{e}_j,\mathbf{e}_j,\mathbf{z},...). \tag{1.62}$$

Of course, the trace does not depend on the choice of the basis. For instance, we may also use a decomposition of the form

$$g^{-1} = \sum_{j=1}^{n} \mathbf{u}_j \otimes \mathbf{v}_j, \tag{1.63}$$

where $\mathbf{u}_j$ and $\mathbf{v}_j$ are some suitable vectors that are not necessarily linearly independent or orthogonal (and $n \geq 4$). If we use the decomposition (1.63), we will have to write

$$\mathrm{Tr}_{\mathbf{x}} T(\mathbf{x}, \mathbf{x}, \mathbf{z}, ...) = \sum_{j=1}^{n} T(\mathbf{u}_j, \mathbf{v}_j, \mathbf{z}, ...).$$

However, it is somewhat inconvenient to have to choose particular vectors for a decomposition (1.61) or (1.63), especially since the result does not depend on the choice of these vectors. In many cases, a trace $\mathrm{Tr}_{\mathbf{x}} T(\mathbf{x}, \mathbf{x}, \mathbf{z}, ...)$ can be computed without selecting an explicit decomposition of the metric, especially when the tensor T is given by an index-free expression involving fixed vectors, the metric g, and covariant derivatives. The following statement lists the basic building blocks for index-free trace calculations.

Statement 1.7.3.1: The trace operation has the following properties:
Linearity and distributivity:

$$\begin{aligned} \mathrm{Tr}\,(A + B) &= \mathrm{Tr}\, A + \mathrm{Tr}\, B, \\ \mathrm{Tr}_{\mathbf{a}} \mathrm{Tr}_{\mathbf{b}}(...) &= \mathrm{Tr}_{\mathbf{b}} \mathrm{Tr}_{\mathbf{a}}(...), \\ \mathrm{Tr}_{\mathbf{a}} A(...) \otimes B(\mathbf{a}, \mathbf{a}, ...) &= A(...) \otimes \mathrm{Tr}_{\mathbf{a}} B(\mathbf{a}, \mathbf{a}, ...), \\ \mathrm{Tr}_{\mathbf{a}} \nabla_{\mathbf{b}} A(\mathbf{a}, \mathbf{a}, \mathbf{x}, ...) &= \nabla_{\mathbf{b}} \mathrm{Tr}_{\mathbf{a}} A(\mathbf{a}, \mathbf{a}, \mathbf{x}, ...). \end{aligned}$$

Here A, B are arbitrary tensors, g is the metric, $\hat{P}_1$, $\hat{P}_2$ are linear transformations, i.e. (1,1)-tensors, and $\mathbf{a}, \mathbf{b}, ..., \mathbf{x}$ are vectors.
Symmetry properties:

$$\begin{aligned} \mathrm{Tr}_{\mathbf{a}} g(\hat{P}_1 \hat{P}_2 \mathbf{a}, \mathbf{a}) &= \mathrm{Tr}_{\mathbf{a}} g(\hat{P}_2 \hat{P}_1 \mathbf{a}, \mathbf{a}), \\ \mathrm{Tr}_{\mathbf{a}} g(\hat{P}_1 \mathbf{a}, \mathbf{a}) &= \mathrm{Tr}_{\mathbf{a}} g(\hat{P}_1^T \mathbf{a}, \mathbf{a}), \end{aligned}$$

where we denote by $\hat{P}^T$ the **adjoint transformation** to $\hat{P}$ with respect to the metric g: The transformation $\hat{P}^T$ acts on $\mathbf{x}$ by yielding a vector $\hat{P}^T \mathbf{x}$ that satisfies $g(\hat{P}^T \mathbf{x}, \mathbf{y}) \equiv g(\mathbf{x}, \hat{P}\mathbf{y})$ for arbitrary $\mathbf{y}$. (The adjoint transformation is represented in an *orthonormal* basis by the transposed matrix, hence the notation $\hat{P}^T$.)
Relations involving the metric:

$$\begin{aligned} \mathrm{Tr}_{\mathbf{a}} g(\nabla_{\mathbf{a}} \mathbf{x}, \mathbf{a}) &= \mathrm{div}\mathbf{x}; \\ \mathrm{Tr}_{\mathbf{a}} g(\mathbf{a}, \mathbf{a}) &= 4, \end{aligned}$$

where 4 is the dimension of spacetime.
The substitution properties:

$$\mathrm{Tr}_{\mathbf{a}} g(\mathbf{a}, \mathbf{x})\mathbf{a} = \mathbf{x},$$

or more generally,

$$\mathrm{Tr}_{\mathbf{a}} g(\mathbf{a}, \mathbf{x}) A(\mathbf{a}, \mathbf{y}, \mathbf{z}, ...) = A(\mathbf{x}, \mathbf{y}, \mathbf{z}, ...)$$

for an arbitrary tensor-valued function $A(\mathbf{x}, ...)$ linear in $\mathbf{x}$.
Relations involving the Levi-Citiva tensor $\boldsymbol{\varepsilon} \equiv \boldsymbol{\theta}^0 \wedge \boldsymbol{\theta}^1 \wedge \boldsymbol{\theta}^2 \wedge \boldsymbol{\theta}^3$:

$$\begin{aligned} &\mathrm{Tr}_{\mathbf{a}_1, \mathbf{a}_2} \boldsymbol{\varepsilon}(\mathbf{a}_1, \mathbf{a}_2, \mathbf{x}, \mathbf{y}) \boldsymbol{\varepsilon}(\mathbf{a}_1, \mathbf{a}_2, \mathbf{x}', \mathbf{y}') \\ &\quad = -2g(\mathbf{x}, \mathbf{x}')g(\mathbf{y}, \mathbf{y}') + 2g(\mathbf{x}, \mathbf{y}')g(\mathbf{y}, \mathbf{x}'); \\ &\mathrm{Tr}_{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3} \boldsymbol{\varepsilon}(\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{x}) \boldsymbol{\varepsilon}(\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{x}') = -6g(\mathbf{x}, \mathbf{x}'). \end{aligned}$$

Hints: These properties are straightforward to derive using a tensor decomposition of the inverse metric g^{-1} in an orthonormal basis. As a last resort, one could rewrite all expressions in the index notation, but this is not necessary.

Partial proof: Let us first derive the two substitution properties. To show that $\mathrm{Tr}_{\mathbf{a}} g(\mathbf{a}, \mathbf{x})\mathbf{a} = \mathbf{x}$, we decompose $x = \sum_j x_j \mathbf{e}_j$ in an orthonormal basis $\{\mathbf{e}_j\}$ and use Eq. (1.61):

$$\mathrm{Tr}_{\mathbf{a}} g(\mathbf{a}, \mathbf{x})\mathbf{a} = \sum_j g(\mathbf{e}_j, \mathbf{x})\mathbf{e}_j = \sum_j x_j \mathbf{e}_j = \mathbf{x}.$$

For an arbitrary tensor-valued multilinear function $A(...)$, we have $g(\mathbf{a}, \mathbf{x})A(\mathbf{a}, \mathbf{y}, ...) = A(g(\mathbf{a}, \mathbf{x})\mathbf{a}, \mathbf{y}, ...)$. Since we already derived the desired property for $g(\mathbf{a}, \mathbf{x})\mathbf{a}$, and since $A(...)$ is linear in its arguments, it follows that the the same trace property holds also for $A(...)$.

Let us now derive the symmetry properties. Due to the substitution property, we have

$$\mathrm{Tr}_{\mathbf{a}} g(\hat{P}_1 \hat{P}_2 \mathbf{a}, \mathbf{a}) = \mathrm{Tr}_{\mathbf{a}, \mathbf{b}} g(\hat{P}_1 \mathbf{b}, \mathbf{a}) g(\hat{P}_2 \mathbf{a}, \mathbf{b}).$$

The last expression is obviously symmetric with respect to exchange of $\hat{P}_1$ and $\hat{P}_2$. The second symmetry property follows directly from the definition of $\hat{P}^T$. ■

Example: Let us compute the trace of the transformation $\hat{P}\mathbf{x} \equiv \mathbf{x} - 2\mathbf{n}g(\mathbf{n}, \mathbf{x})$, where $\mathbf{n}$ is a given unit vector, $g(\mathbf{n}, \mathbf{n}) = 1$.

First we define the bilinear form $P(\mathbf{x}, \mathbf{y}) \equiv g(\hat{P}\mathbf{x}, \mathbf{y})$; then we compute its trace with respect to $\mathbf{x}, \mathbf{y}$:

$$\begin{aligned} \mathrm{Tr}\, \hat{P} \equiv \mathrm{Tr}_{\mathbf{x}} P(\mathbf{x}, \mathbf{x}) &= \mathrm{Tr}_{\mathbf{x}} \left[g(\mathbf{x}, \mathbf{x}) - 2g(\mathbf{n}, \mathbf{x})g(\mathbf{n}, \mathbf{x}) \right] \\ &= 4 - 2g(\mathbf{n}, \mathbf{n}) = 2. \end{aligned}$$

■

Practice problem: For a fixed vector $\mathbf{u}$ in four-dimensional space, compute the trace

$$\mathrm{Tr}_{\mathbf{a}} g(\mathbf{a}, C(\mathbf{a}, \mathbf{u})\mathbf{u}),$$

where the transformation-valued 2-form C is defined by

$$C(\mathbf{x}, \mathbf{y})\mathbf{z} \equiv g(\mathbf{y}, \mathbf{z})\mathbf{x} - g(\mathbf{x}, \mathbf{z})\mathbf{y}.$$

Answer: $3g(\mathbf{u}, \mathbf{u})$. ■

Derivatives of mute vectors. The tensor under the trace *may* contain derivatives of mute vectors, as long as these derivatives can be moved out of the expression or when the derivatives ultimately cancel out. In other words, the expression under the trace should be a *linear function* of the mute vectors, not really involving their derivatives. Otherwise, the trace expression is *undefined* (meaningless). For instance, the expressions $\mathrm{Tr}_{\mathbf{a}} \nabla_{\mathbf{a}} \mathbf{a}$ and $\mathrm{Tr}_{\mathbf{a}} g(\mathbf{a}, [\mathbf{x}, \mathbf{a}])$ are, strictly speaking, meaningless because the result of a calculation according to Eq. (1.63) depends on the derivatives of the basis vector fields $\{\mathbf{e}_j\}$. However, such meaningless expressions can occur neither by transforming well-defined index-free expressions nor through rewriting well-defined indexed expressions with traces. Only well-defined trace expressions can be obtained in practice. For instance, the following expressions containing derivatives of mute vectors are well-defined,

$$\begin{aligned} \mathrm{Tr}_{\mathbf{a}} \nabla_{\mathbf{x}} g(\mathbf{a}, \nabla_{\mathbf{a}} \mathbf{y}) &= \nabla_{\mathbf{x}} \mathrm{Tr}_{\mathbf{a}} g(\mathbf{a}, \nabla_{\mathbf{a}} \mathbf{y}) = \nabla_{\mathbf{x}} (\mathrm{div}\mathbf{y}); \\ \mathrm{Tr}_{\mathbf{a}} \left(\nabla_{\mathbf{a}} \nabla_{\mathbf{a}} \mathbf{x} - \nabla_{\nabla_{\mathbf{a}} \mathbf{a}} \mathbf{x} \right) &\equiv \Box \mathbf{x} \equiv g^{\mu\nu} \nabla_\mu \nabla_\nu \mathbf{x}. \end{aligned}$$

The last line is the D'Alembert operator acting on a vector field $\mathbf{x}$. The trace is well-defined because the expression under the trace is actually independent of the derivatives of the mute vector $\mathbf{a}$.

Computations with expressions of this kind can be simplified using the following trick. It is always possible to choose basis vector fields so that $\nabla_{\mathbf{e}_j}\mathbf{e}_k = 0$ for all j, k at *one* point in spacetime. Hence, when computing a *well-defined* trace expression containing derivatives, one could simply assume that all the mute vectors are "constant vectors," in the sense that their (covariant) derivatives in any direction are all equal to zero. This assumption is also consistent with the property $\nabla . g = 0$ and a decomposition (1.61) of the metric through a basis that enters the trace via Eq. (1.62).

In this book I will use the simplifying convention that covariant derivatives of mute vectors are always equal to zero. For instance, the D'Alembert operator can then be written more intuitively as follows,

$$\Box \mathbf{x} = \mathrm{Tr}_{\mathbf{a}} \nabla_{\mathbf{a}} \nabla_{\mathbf{a}} \mathbf{x}.$$

1.7.4 Summary of calculation rules

The formalism of coordinate-free, index-free calculations is sufficiently powerful to handle all situations involving vector fields and 1-forms, the metric, and covariant derivatives. (However, the index-free notation becomes cumbersome when one needs to manipulate tensors of higher rank.) In comparison with the index notation, this formalism is more transparent, emphasizes the geometric content of a calculation, and frequently helps to guess the necessary sequence of manipulations that lead to a desired result. I summarize the rules of index-free, coordinate-free calculations for convenience.

Vector fields are denoted by boldface letters $\mathbf{a}, \mathbf{v}, \mathbf{x}$ etc., while scalar functions are denoted by ordinary letters a, b, f, etc. It is sometimes convenient to use Greek letters α, λ, μ for scalar functions. In blackboard calculations, vectors may be underlined or even simply left unmarked if it does not create confusion.

The basic operations of tensor calculus are tensor product, tensor contraction, and the various derivative operations. Explicit tensor products are written as $\mathbf{u} \otimes \mathbf{v}$ and seem to be seldom needed. Tensor contraction is written as an application of a function to a vector, for example $\omega \circ \mathbf{v}$ or $\omega(\mathbf{v})$ if ω is a 1-form, $g(\mathbf{a}, \mathbf{b})$ if g is a metric, or $\iota_{\mathbf{v}}\omega$ if ω is an n-form.

The metric g creates a correspondence between vectors and 1-forms; this correspondence and its inverse are denoted by $\hat{g}$ and $\hat{g}^{-1}$. For instance, $\hat{g}\mathbf{v}$ is a 1-form if $\mathbf{v}$ is a vector.

There are three basic derivative operations: the Lie derivative $\mathcal{L}_{\mathbf{v}}A$ of a tensor A (the application of a vector field to a scalar field is a particular case, $\mathbf{v} \circ \lambda = \mathcal{L}_{\mathbf{v}}\lambda$); the exterior differential $\mathrm{d}\omega$ of an n-form ω (where 0-forms are understood as scalar functions); and the Levi-Civita covariant derivative $\nabla_{\mathbf{v}}A$ of a tensor A in the direction given by a vector $\mathbf{v}$. The quantities $\mathcal{L}_{\mathbf{v}}A$ and $\nabla_{\mathbf{v}}A$ are tensors of the same rank as A. The quantity $\mathrm{d}\omega$ is an $(n+1)$-form if ω is an n-form.

In calculations, it is usually more convenient to use $\nabla_{\mathbf{v}}$ with vectors but d, $\iota_{\mathbf{v}}$, and $\mathcal{L}_{\mathbf{v}}$ with n-forms.

The operations $\iota_{\mathbf{v}}$ and $\nabla_{\mathbf{v}}$ are linear in $\mathbf{v}$ and **local** (depend only on the value of $\mathbf{v}$ at a point but not on its derivatives):

$$\iota_{\lambda\mathbf{v}}(...) = \lambda\iota_{\mathbf{v}}(...), \quad \nabla_{\lambda\mathbf{v}}(...) = \lambda\nabla_{\mathbf{v}}(...).$$

However, $\mathcal{L}_{\mathbf{v}}$ is not local in $\mathbf{v}$,

$$\mathcal{L}_{\lambda\mathbf{v}}(...) \neq \lambda\mathcal{L}_{\mathbf{v}}(...).$$

The precise expression for $\mathcal{L}_{\lambda\mathbf{v}}A$ depends on the rank of the tensor A. The Leibnitz rule with respect to tensor products holds for the derivative operations, with the exception of d which needs an extra sign:

$$\begin{aligned}
\nabla_{\mathbf{a}}(\eta \wedge \theta) &= (\nabla_{\mathbf{a}}\eta) \wedge \theta + \eta \wedge \nabla_{\mathbf{a}}\theta, \\
\mathcal{L}_{\mathbf{a}}(\eta \wedge \theta) &= (\mathcal{L}_{\mathbf{a}}\eta) \wedge \theta + \eta \wedge \mathcal{L}_{\mathbf{a}}\theta, \\
\mathrm{d}(\eta \wedge \theta) &= (\mathrm{d}\eta) \wedge \theta + (-1)^{|\eta|}\eta \wedge \mathrm{d}\theta,
\end{aligned}$$

where $|\eta|$ in the expression $(-1)^{|\eta|}$ means the rank n of an n-form η. Also

$$\iota_{\mathbf{a}}(\eta \wedge \theta) = (\iota_{\mathbf{a}}\eta) \wedge \theta + (-1)^{|\eta|}\eta \wedge \iota_{\mathbf{a}}\theta.$$

Some further rules for calculations are the following (here $\mathbf{a}, \mathbf{b}, \mathbf{c}$ are vectors, λ is a scalar, ω is an n-form, $\boldsymbol{\theta}$ is a 1-form, g is the metric tensor).

$$\begin{aligned}
[\mathbf{a}, \mathbf{b}] \circ \lambda &\equiv \mathbf{a} \circ (\mathbf{b} \circ \lambda) - \mathbf{b} \circ (\mathbf{a} \circ \lambda), \\
\mathcal{L}_{\mathbf{a}}\mathbf{b} &= [\mathbf{a}, \mathbf{b}] = \nabla_{\mathbf{a}}\mathbf{b} - \nabla_{\mathbf{b}}\mathbf{a} = -\mathcal{L}_{\mathbf{b}}\mathbf{a}, \\
\mathcal{L}_{\mathbf{a}}\lambda &= \mathbf{a} \circ \lambda \equiv (\mathrm{d}\lambda) \circ \mathbf{a} \equiv \iota_{\mathbf{a}}\mathrm{d}\lambda = \nabla_{\mathbf{a}}\lambda, \\
\nabla_{\mathbf{a}}\nabla_{\mathbf{b}}\lambda &= \nabla_{\mathbf{b}}\nabla_{\mathbf{a}}\lambda, \\
\nabla_{\mathbf{a}}g &= 0, \\
(\mathrm{d}\boldsymbol{\theta}) \circ (\mathbf{a}, \mathbf{b}) &= \mathbf{a} \circ (\boldsymbol{\theta} \circ \mathbf{b}) - \mathbf{b} \circ (\boldsymbol{\theta} \circ \mathbf{a}) - \boldsymbol{\theta} \circ [\mathbf{a}, \mathbf{b}], \\
(\iota_{\mathbf{a}}\omega) \circ (\mathbf{b}, \mathbf{c}, ...) &\equiv \omega \circ (\mathbf{a}, \mathbf{b}, \mathbf{c}, ...); \\
\mathcal{L}_{\mathbf{a}}\omega &= \mathrm{d}\iota_{\mathbf{a}}\omega + \iota_{\mathbf{a}}\mathrm{d}\omega, \\
\mathrm{d}\mathrm{d}\omega &\equiv 0.
\end{aligned}$$

The symbol $\equiv$ refers to properties that hold directly by definition.

When an expression involves arbitrary vectors and is a local function of these vectors, it is often convenient to choose these vectors so that their (*first*) derivatives vanish. (Second derivatives cannot be constrained to vanish simultaneously with the first derivatives.)

Examples: A consequence of $\nabla_{\mathbf{a}}g = 0$ and the Leibnitz rule is the identity

$$\nabla_{\mathbf{a}}\hat{g}^{-1}\omega = \hat{g}^{-1}\nabla_{\mathbf{a}}\omega$$

for a 1-form ω.

The equivalence of $\nabla_{\mathbf{a}}$ and $\mathcal{L}_{\mathbf{a}}$ on scalars leads to useful simplifications in some calculations with the metric. For instance,

$$\begin{aligned}
\mathbf{a} \circ g(\mathbf{b}, \mathbf{c}) &= \nabla_{\mathbf{a}}g(\mathbf{b}, \mathbf{c}) = \mathcal{L}_{\mathbf{a}}g(\mathbf{b}, \mathbf{c}); \\
\nabla_{\mathbf{a}}g(\mathbf{b}, \mathbf{c}) &= g(\nabla_{\mathbf{a}}\mathbf{b}, \mathbf{c}) + g(\mathbf{b}, \nabla_{\mathbf{a}}\mathbf{c}), \\
\mathcal{L}_{\mathbf{a}}g(\mathbf{b}, \mathbf{c}) &= (\mathcal{L}_{\mathbf{a}}g) \circ (\mathbf{b}, \mathbf{c}) + g([\mathbf{a}, \mathbf{b}], \mathbf{c}) + g(\mathbf{b}, [\mathbf{a}, \mathbf{c}]).
\end{aligned}$$

Depending on the particular calculation, one of these equivalent forms may be used to proceed. ■

The rules for index-free calculations with traces are listed in Statement 1.7.3.1. Additionally, one may use the simplifying convention that all (*first*) covariant derivatives of mute vectors vanish.

Here is another example of an index-free calculation.

Statement 1.7.4.1: Let g and $\tilde{g} = e^{2\lambda}g$ be two metrics related by a conformal transformation. For a vector field $\mathbf{u}$, we can then compute two different divergences $\mathrm{div}\mathbf{u}$ and $\tilde{\mathrm{div}}\mathbf{u}$ with respect to the metrics g and $\tilde{g}$. These divergences are related by the formula

$$\tilde{\mathrm{div}}\mathbf{u} = \mathrm{div}\mathbf{u} + N\mathbf{u} \circ \lambda,$$

where N is the dimension of the manifold.

Proof of Statement 1.7.4.1: The simplest proof is through the definition of div$\mathbf{u}$ as the coefficient in the equation

$$\mathcal{L}_{\mathbf{u}}\boldsymbol{\varepsilon} = (\mathrm{div}\mathbf{u})\,\boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ is the Levi-Civita tensor corresponding to the metric g. The Levi-Civita tensor is defined as the N-form with the property

$$\boldsymbol{\varepsilon}(\mathbf{e}_1, ..., \mathbf{e}_N) = 1,$$

if $\{\mathbf{e}_j\}$ is an orthonormal basis in the metric g. When the metric g is multiplied by $e^{2\lambda}$, each vector in the orthonormal basis is multiplied by $e^{-\lambda}$. It follows that the Levi-Civita tensor is multiplied by $e^{N\lambda}$,

$$\tilde{\boldsymbol{\varepsilon}} = e^{N\lambda}\boldsymbol{\varepsilon}.$$

Then we compute

$$\begin{aligned}\mathcal{L}_{\mathbf{u}}\tilde{\boldsymbol{\varepsilon}} &= \left(\mathcal{L}_{\mathbf{u}}e^{N\lambda}\right)\boldsymbol{\varepsilon} + e^{N\lambda}\mathcal{L}_{\mathbf{u}}\boldsymbol{\varepsilon} = (N\mathbf{u}\circ\lambda)\,\tilde{\boldsymbol{\varepsilon}} + (\mathrm{div}\mathbf{u})\,\tilde{\boldsymbol{\varepsilon}} \\ &= (\tilde{\mathrm{div}}\mathbf{u})\,\tilde{\boldsymbol{\varepsilon}}.\end{aligned}$$

Let us perform another, slightly longer computation of $\tilde{\mathrm{div}}\mathbf{u}$ using the formula involving the trace. We denote by $\tilde{\nabla}$ the Levi-Civita connection for the metric $\tilde{g}$. The inverse metric is $\tilde{g}^{-1} = e^{-2\lambda}g^{-1}$. We will use the formula for the divergence through the trace,

$$\tilde{\mathrm{div}}\mathbf{u} = \mathrm{Tr}_{\mathbf{a}\otimes\mathbf{a}=\tilde{g}^{-1}}\tilde{g}(\tilde{\nabla}_{\mathbf{a}}\mathbf{u}, \mathbf{a}) = e^{-2\lambda}\mathrm{Tr}_{\mathbf{a}}g(\tilde{\nabla}_{\mathbf{a}}\mathbf{u}, \mathbf{a}).$$

Using Eq. (1.47), we compute

$$\begin{aligned}e^{-2\lambda}\mathrm{Tr}_{\mathbf{a}}\tilde{g}(\tilde{\nabla}_{\mathbf{a}}\mathbf{u}, \mathbf{a}) &= \mathrm{div}\mathbf{u} \\ +\mathrm{Tr}_{\mathbf{a}}\left[(\mathbf{a}\circ\lambda)\,g(\mathbf{u}, \mathbf{a}) + (\mathbf{u}\circ\lambda)\,g(\mathbf{a}, \mathbf{a}) - (\mathbf{a}\circ\lambda)\,g(\mathbf{u}, \mathbf{a})\right] \\ = \mathrm{div}\mathbf{u} + N\mathbf{u}\circ\lambda.\end{aligned}$$

■

1.8 Curvature

A motivation for introducing the concept of curvature is to describe in detail the deviation of the geometry of a manifold near each point from the flat geometry. A direct description of this deviation can be given using the concept of parallel transport (see also Sec. 1.9.1). In Appendix A a formula is derived for the parallel transport of a vector along an infinitesimally small closed curve, and thus the connection to the curvature tensor is made explicit (see Sec. A.5.3). In this approach, the curvature tensor is the object describing the transformation of vectors under parallel transport along infinitesimally small closed curves. However, calculations in this approach require using a local coordinate system because otherwise the parallel transport operation cannot be described explicitly. Presently, I introduce the curvature tensor by a different formula that can be directly used in index-free and coordinate-free calculations.

1.8.1 Curvature of a connection

The **curvature of a connection** ∇ is a transformation-valued 2-form $R(\mathbf{u}, \mathbf{v})$ defined by[10]

$$R(\mathbf{u}, \mathbf{v})\mathbf{w} = [\nabla_{\mathbf{u}}, \nabla_{\mathbf{v}}]\,\mathbf{w} - \nabla_{[\mathbf{u},\mathbf{v}]}\mathbf{w}. \tag{1.64}$$

At first glance, it may appear that the right-hand side of the expression (1.64) contains derivatives of the vectors $\mathbf{u}, \mathbf{v}, \mathbf{w}$. However, this is not so.

[10] Equation (1.64) is known as the Ricci identity; it is convenient to regard it as the definition of $R(\mathbf{u}, \mathbf{v})$.

Statement 1.8.1.1: The vector field $R(\mathbf{u}, \mathbf{v})\mathbf{w}$ defined by Eq. (1.64) does not depend on the derivatives of the vector fields $\mathbf{w}$, $\mathbf{u}$, or $\mathbf{v}$. Thus, $R(\mathbf{u}, \mathbf{v})$ is indeed a transformation-valued 2-form.

Proof of Statement 1.8.1.1: Let us verify that no derivatives of $\mathbf{w}$ remain in $R(\mathbf{u}, \mathbf{v})\mathbf{w}$; to that end, we show that $R(\mathbf{u}, \mathbf{v})\lambda\mathbf{w} = \lambda R(\mathbf{u}, \mathbf{v})\mathbf{w}$, where λ is an arbitrary scalar function. This is obtained by a direct computation:

$$\begin{aligned}\nabla_{\mathbf{u}}\nabla_{\mathbf{v}}\lambda\mathbf{w} &= \nabla_{\mathbf{u}}\left(\lambda\nabla_{\mathbf{v}}\mathbf{w} + (\mathbf{v}\circ\lambda)\,\mathbf{w}\right) \\ &= \lambda\nabla_{\mathbf{u}}\nabla_{\mathbf{v}}\mathbf{w} + (\mathbf{u}\circ\lambda)\,\nabla_{\mathbf{v}}\mathbf{w} \\ &\quad + (\mathbf{u}\circ(\mathbf{v}\circ\lambda))\mathbf{w} + (\mathbf{v}\circ\lambda)\nabla_{\mathbf{u}}\mathbf{w}; \\ \nabla_{[\mathbf{u},\mathbf{v}]}\lambda\mathbf{w} &= \lambda\nabla_{[\mathbf{u},\mathbf{v}]}\mathbf{w} + ([\mathbf{u}, \mathbf{v}]\circ\lambda)\mathbf{w};\end{aligned}$$

now it is easy to see that

$$\nabla_{\mathbf{u}}\nabla_{\mathbf{v}}\lambda\mathbf{w} - \nabla_{\mathbf{v}}\nabla_{\mathbf{u}}\lambda\mathbf{w} = \lambda R(\mathbf{u}, \mathbf{v})\mathbf{w} + \nabla_{[\mathbf{u},\mathbf{v}]}\lambda\mathbf{w}.$$

Similarly, we show that $R(\mathbf{u}, \mathbf{v})\mathbf{w}$ is a linear function also of $\mathbf{v}$ (i.e. does not involve derivatives of $\mathbf{v}$) by computing $R(\mathbf{u}, \lambda\mathbf{v})\mathbf{w}$. We use Calculation 1.3.1.1 to express $[\mathbf{u}, \lambda\mathbf{v}]$, and obtain

$$\begin{aligned}\nabla_{\mathbf{u}}\nabla_{\lambda\mathbf{v}}\mathbf{w} &= \nabla_{\mathbf{u}}\left(\lambda\nabla_{\mathbf{v}}\mathbf{w}\right) = \lambda\nabla_{\mathbf{u}}\nabla_{\mathbf{v}}\mathbf{w} + (\mathbf{u}\circ\lambda)\,\nabla_{\mathbf{v}}\mathbf{w}; \\ \nabla_{[\mathbf{u},\lambda\mathbf{v}]}\mathbf{w} &= \nabla_{(\mathbf{u}\circ\lambda)\mathbf{v}+\lambda[\mathbf{u},\mathbf{v}]}\mathbf{w} = (\mathbf{u}\circ\lambda)\,\nabla_{\mathbf{v}}\mathbf{w} + \lambda\nabla_{[\mathbf{u},\mathbf{v}]}\mathbf{w}; \\ R(\mathbf{u}, \lambda\mathbf{v})\mathbf{w} &= \lambda\nabla_{\mathbf{u}}\nabla_{\mathbf{v}}\mathbf{w} - \lambda\nabla_{\mathbf{u}}\nabla_{\mathbf{v}}\mathbf{w} - \lambda\nabla_{[\mathbf{u},\mathbf{v}]}\mathbf{w} = \lambda R(\mathbf{u}, \mathbf{v})\mathbf{w}.\end{aligned}$$

The analogous property for $\mathbf{u}$ follows from the antisymmetry of $R(\mathbf{u}, \mathbf{v})\mathbf{w}$ in $(\mathbf{u}, \mathbf{v})$. ■

The index notation for the curvature tensor, $R_{\alpha\beta\mu}{}^{\nu}$, can be defined by

$$(R(\mathbf{x}, \mathbf{y})\mathbf{z})^{\nu} \equiv x^{\alpha}y^{\beta}z^{\mu}R_{\alpha\beta\mu}{}^{\nu}.$$

We can also define a tensor $R(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d})$ of rank (0,4) by "lowering the index," i.e.

$$R(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) \equiv g(R(\mathbf{a}, \mathbf{b})\mathbf{c}, \mathbf{d}) \equiv R_{\alpha\beta\mu\nu}a^{\alpha}b^{\beta}c^{\mu}d^{\nu}.$$

Either of these equivalent tensors R, when evaluated using the Levi-Civita connection, is called the **Riemann tensor** of the manifold.[11]

1.8.2 Bianchi identities

In this section I review and derive the standard properties of the Riemann tensor.

The antisymmetry property,

$$R(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) = -R(\mathbf{b}, \mathbf{a}, \mathbf{c}, \mathbf{d}), \tag{1.65}$$

follows immediately from the definition (1.64). Additionally, the curvature tensor corresponding to the Levi-Civita connection has the following symmetry properties:

$$R(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) = -R(\mathbf{a}, \mathbf{b}, \mathbf{d}, \mathbf{c}), \tag{1.66}$$

$$R(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) + R(\mathbf{b}, \mathbf{c}, \mathbf{a}, \mathbf{d}) + R(\mathbf{c}, \mathbf{a}, \mathbf{b}, \mathbf{d}) = 0. \tag{1.67}$$

$$R(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) = R(\mathbf{c}, \mathbf{d}, \mathbf{a}, \mathbf{b}). \tag{1.68}$$

The property (1.67) is called the **first Bianchi identity**. These properties hold *only* for the Levi-Civita connection; we always work with that connection here.

[11] We are using the $(-+-)$ sign convention, according to the classification in Misner-Thorne-Wheeler [21].

Remarks: Usually, one simply says "Riemann tensor of the manifold" instead of the more precise but cumbersome phrase "Riemann tensor of the Levi-Civita connection corresponding to the given metric on the manifold." The Riemann tensor is a function of the metric that involves second derivatives of the metric (see Calculation 1.8.4.1 below). It is easier to remember the symmetries (1.65)-(1.68) of the Riemann tensor $R(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d})$ if one thinks of the expression $g(\mathbf{a},\mathbf{c})g(\mathbf{b},\mathbf{d}) - g(\mathbf{a},\mathbf{d})g(\mathbf{b},\mathbf{c})$. This expression (as will be shown below) is actually equal the Riemann tensor of a unit sphere in Euclidean space, if g is the induced metric on the sphere.[12] ■

We will now show in a series of statements that: the property (1.66) is a consequence of torsion-freeness and the compatibility of ∇ with the metric; the property (1.67) is a consequence of torsion-freeness and the Jacobi identity for commutators of vectors; and Eq. (1.68) is a purely algebraic consequence of Eqs. (1.66)-(1.67).

Statement 1.8.2.1: The Riemann tensor has the property $R(\mathbf{a},\mathbf{b},\mathbf{x},\mathbf{x}) = 0$, where $\mathbf{a},\mathbf{b},\mathbf{x}$ are arbitrary vectors. The identity (1.66) then follows due to linearity of $R(\cdot,\cdot,\cdot,\cdot)$.

Idea of proof: Transform the torsion-freeness condition

$$0 = (\nabla_\mathbf{a}\nabla_\mathbf{b} - \nabla_\mathbf{b}\nabla_\mathbf{a} - [\mathbf{a},\mathbf{b}]) \circ g(\mathbf{x},\mathbf{x})$$

using the metricity condition (1.41).

Proof of Statement 1.8.2.1: We have

$$\nabla_\mathbf{a}\nabla_\mathbf{b} g(\mathbf{x},\mathbf{x}) = 2g(\nabla_\mathbf{a}\nabla_\mathbf{b}\mathbf{x},\mathbf{x}) + 2g(\nabla_\mathbf{a}\mathbf{x},\nabla_\mathbf{b}\mathbf{x}),$$

hence

$$[\nabla_\mathbf{a},\nabla_\mathbf{b}]g(\mathbf{x},\mathbf{x}) = 2g([\nabla_\mathbf{a},\nabla_\mathbf{b}]\mathbf{x},\mathbf{x}).$$

Further, we rewrite $[\mathbf{a},\mathbf{b}] \circ g(\mathbf{x},\mathbf{x}) = \nabla_{[\mathbf{a},\mathbf{b}]} g(\mathbf{x},\mathbf{x})$ and find $\nabla_{[\mathbf{a},\mathbf{b}]} g(\mathbf{x},\mathbf{x}) = 2g(\nabla_{[\mathbf{a},\mathbf{b}]}\mathbf{x},\mathbf{x})$. Finally,

$$\begin{aligned} R(\mathbf{a},\mathbf{b},\mathbf{x},\mathbf{x}) &= g(R(\mathbf{a},\mathbf{b})\mathbf{x},\mathbf{x}) \\ &= g([\nabla_\mathbf{a},\nabla_\mathbf{b}]\mathbf{x},\mathbf{x}) - g(\nabla_{[\mathbf{a},\mathbf{b}]}\mathbf{x},\mathbf{x}) \\ &= \frac{1}{2}\left([\nabla_\mathbf{a},\nabla_\mathbf{b}] - \nabla_{[\mathbf{a},\mathbf{b}]}\right) g(\mathbf{x},\mathbf{x}) = 0. \end{aligned}$$

The identity $R(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d}) + R(\mathbf{a},\mathbf{b},\mathbf{d},\mathbf{c}) = 0$ follows by setting $\mathbf{x} = \mathbf{c} + \mathbf{d}$ in $R(\mathbf{a},\mathbf{b},\mathbf{x},\mathbf{x}) = 0$. ■

Simplifying assumption. The calculation in the preceding proof can be made shorter by assuming that the vectors $\mathbf{a},\mathbf{b}$ commute, $[\mathbf{a},\mathbf{b}] = 0$, and that the derivatives of $\mathbf{x}$ in directions $\mathbf{a}$ and $\mathbf{b}$ vanish. This can be assumed without loss of generality because $R(\mathbf{a},\mathbf{b},\mathbf{x},\mathbf{x})$ does not depend on derivatives of $\mathbf{a},\mathbf{b},\mathbf{x}$, but only on values of $\mathbf{a}$ and $\mathbf{b}$ and $\mathbf{x}$ at a point, and thus these derivatives (such as $\nabla_\mathbf{a}\mathbf{b}$ or $\nabla_\mathbf{b}\mathbf{x}$) can be set to zero. (We can always find any finite number of vector fields $\mathbf{a},\mathbf{b},\mathbf{c},...$ in the neighborhood of a point p such that the values $\mathbf{a}(p),\mathbf{b}(p),\mathbf{c}(p),...$ at the point p are prescribed and all first derivatives, such as $\nabla_\mathbf{a}\mathbf{b}$ or $\nabla_\mathbf{b}\mathbf{c}$, vanish at p; but second derivatives, such as $\nabla_\mathbf{a}\nabla_\mathbf{b}\mathbf{c}$, will in general not vanish at p.)

Here is the abbreviated calculation. Assuming that $\nabla_\mathbf{a}\mathbf{x} = \nabla_\mathbf{b}\mathbf{x} = [\mathbf{a},\mathbf{b}] = 0$, we have

$$\begin{aligned} 0 = [\mathbf{a},\mathbf{b}] \circ g(\mathbf{x},\mathbf{x}) &= [\nabla_\mathbf{a},\nabla_\mathbf{b}]\, g(\mathbf{x},\mathbf{x}) \\ &= 2g([\nabla_\mathbf{a},\nabla_\mathbf{b}]\,\mathbf{x},\mathbf{x}) = 2R(\mathbf{a},\mathbf{b},\mathbf{x},\mathbf{x}). \end{aligned}$$

Below we will often use the assumption of vanishing first derivatives to simplify calculations.

Statement 1.8.2.2: The property (1.67) follows by rewriting the Jacobi identity

$$[\mathbf{a},[\mathbf{b},\mathbf{c}]] + [\mathbf{b},[\mathbf{c},\mathbf{a}]] + [\mathbf{c},[\mathbf{a},\mathbf{b}]] = 0$$

using covariant derivatives and the torsion-free condition (1.44).

Proof of Statement 1.8.2.2: We start with

$$\begin{aligned} [\mathbf{a},[\mathbf{b},\mathbf{c}]] &= \nabla_\mathbf{a}[\mathbf{b},\mathbf{c}] - \nabla_{[\mathbf{b},\mathbf{c}]}\mathbf{a} \\ &= \nabla_\mathbf{a}\nabla_\mathbf{b}\mathbf{c} - \nabla_\mathbf{a}\nabla_\mathbf{c}\mathbf{b} - \nabla_{[\mathbf{b},\mathbf{c}]}\mathbf{a}, \end{aligned}$$

and similarly for the other terms. Then we write the Jacobi identity and rearrange the resulting expression:

$$\begin{aligned} &[\mathbf{a},[\mathbf{b},\mathbf{c}]] + [\mathbf{b},[\mathbf{c},\mathbf{a}]] + [\mathbf{c},[\mathbf{a},\mathbf{b}]] = \nabla_\mathbf{a}\nabla_\mathbf{b}\mathbf{c} - \nabla_\mathbf{a}\nabla_\mathbf{c}\mathbf{b} + \nabla_\mathbf{b}\nabla_\mathbf{c}\mathbf{a} \\ &\quad -\nabla_\mathbf{b}\nabla_\mathbf{a}\mathbf{c} + \nabla_\mathbf{c}\nabla_\mathbf{a}\mathbf{b} - \nabla_\mathbf{c}\nabla_\mathbf{b}\mathbf{a} - \nabla_{[\mathbf{b},\mathbf{c}]}\mathbf{a} - \nabla_{[\mathbf{c},\mathbf{a}]}\mathbf{b} - \nabla_{[\mathbf{a},\mathbf{b}]}\mathbf{c} \\ &= R(\mathbf{a},\mathbf{b})\mathbf{c} + R(\mathbf{b},\mathbf{c})\mathbf{a} + R(\mathbf{c},\mathbf{a})\mathbf{b} = 0. \end{aligned}$$

■

Statement 1.8.2.3: Any tensor of rank (0,4) satisfying the transposition properties (1.66) and (1.67) will also satisfy Eq. (1.68).

Proof of Statement 1.8.2.3: For brevity, we will write $\underline{abcd}$ instead of $R(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d})$ within this calculation. First, we use the transposition properties to move the indices c,d to the left:

$$\begin{aligned} 0 &= \underline{abcd} + \underline{bcad} + \underline{cabd} = \\ &= \underline{abcd} + \underline{bcad} + (-\underline{cadb}) \\ &= \underline{abcd} + \underline{bcad} + (\underline{adcb} + \underline{dcab}), \end{aligned}$$

which means that

$$\underline{abcd} - \underline{cdab} = -(\underline{bcad} - \underline{adbc}).$$

This last property can be now used repeatedly on different indices:

$$\begin{aligned} -(\underline{bcad} - \underline{adbc}) &= +(\underline{cabd} - \underline{bdac}), \\ +(\underline{cabd} - \underline{bdac}) &= -(\underline{abcd} - \underline{cdab}). \end{aligned}$$

Therefore $\underline{abcd} - \underline{cdab} = -(\underline{abcd} - \underline{cdab}) = 0$. ■

The Riemann tensor also satisfies the **second Bianchi identity**, written in index notation as

$$\nabla_\kappa R_{\lambda\mu\nu\rho} + \nabla_\lambda R_{\mu\kappa\nu\rho} + \nabla_\mu R_{\kappa\lambda\nu\rho} = 0. \tag{1.69}$$

It is somewhat cumbersome to write this identity using the index-free notation such as $R(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d})$ because only R is to be differentiated and not the vectors $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, $\mathbf{d}$, which is not easy to indicate without using indices. The property (1.69) has important consequences for GR.[13]

Let us now derive the Bianchi identity for the transformation-valued tensor $R(\mathbf{u},\mathbf{v})$ defined by Eq. (1.64). We need to convert the index expression (1.69) into an index-free form. So we introduce arbitrary vector fields $\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{v}$ and consider the expression

$$\begin{aligned} c^\kappa a^\lambda b^\mu v^\nu \nabla_\kappa R_{\lambda\mu\nu\rho} &= \nabla_\mathbf{c}\big(R(\mathbf{a},\mathbf{b})\mathbf{v}\big) - R(\nabla_\mathbf{c}\mathbf{a},\mathbf{b})\mathbf{v} \\ &\quad - R(\mathbf{a},\nabla_\mathbf{c}\mathbf{b})\mathbf{v} - R(\mathbf{a},\mathbf{b})\nabla_\mathbf{c}\mathbf{v}. \end{aligned}$$

(The subtractions are needed to make sure that only R is differentiated.)

[12] Due to the symmetry properties, the Riemann tensor has $\frac{1}{12}n^2(n^2-1)$ independent scalar components for an n-dimensional manifold. This fact, which does not seem to have direct applications in General Relativity, nevertheless enjoys a mention (usually without a clear derivation) in many GR textbooks. Here I follow the venerable tradition.

[13] A generalization of the second Bianchi identity also exists for connections with nonvanishing torsion; see [33], §5.5.

Statement 1.8.2.4: The identity (1.69) holds, due to the vanishing of the sum of the twelve terms obtained via cyclic permutations of $\mathbf{a}, \mathbf{b}, \mathbf{c}$ in the above formula.

Proof of Statement 1.8.2.4: We start the derivation by writing the **Jacobi identity**,

$$[[A, B], C] + [[B, C], A] + [[C, A], B] = 0.$$

This is a purely algebraic relation that holds for arbitrary operations A, B, C in any context (as long as the composition of operations is associative). We now apply the Jacobi identity to the operations $\nabla_{\mathbf{a}}$, $\nabla_{\mathbf{b}}$, $\nabla_{\mathbf{c}}$, and also use the fact that $\nabla_{\mathbf{u}}$ is linear in $\mathbf{u}$, to derive the following identities,

$$[[\nabla_{\mathbf{a}}, \nabla_{\mathbf{b}}], \nabla_{\mathbf{c}}]\,\mathbf{v} + [[\nabla_{\mathbf{b}}, \nabla_{\mathbf{c}}], \nabla_{\mathbf{a}}]\,\mathbf{v} + [[\nabla_{\mathbf{c}}, \nabla_{\mathbf{a}}], \nabla_{\mathbf{b}}]\,\mathbf{v} = 0,$$
$$\nabla_{[[\mathbf{a},\mathbf{b}],\mathbf{c}]}\mathbf{v} + \nabla_{[[\mathbf{b},\mathbf{c}],\mathbf{a}]}\mathbf{v} + \nabla_{[[\mathbf{c},\mathbf{a}],\mathbf{b}]}\mathbf{v} = 0.$$

The Bianchi identity will follow if we subtract the first of these relations from the second and express various terms through $R()$ using Eq. (1.64). For instance,

$$\begin{aligned} -[[\nabla_{\mathbf{a}}, \nabla_{\mathbf{b}}], \nabla_{\mathbf{c}}]\,\mathbf{v} + \nabla_{[[\mathbf{a},\mathbf{b}],\mathbf{c}]}\mathbf{v} &= -R(\mathbf{a},\mathbf{b})\nabla_{\mathbf{c}}\mathbf{v} + \nabla_{\mathbf{c}}R(\mathbf{a},\mathbf{b})\mathbf{v} \\ &\quad -\nabla_{[\mathbf{a},\mathbf{b}]}\nabla_{\mathbf{c}}\mathbf{v} + \nabla_{\mathbf{c}}\nabla_{[\mathbf{a},\mathbf{b}]}\mathbf{v} + \nabla_{[[\mathbf{a},\mathbf{b}],\mathbf{c}]}\mathbf{v} \\ &= \nabla_{\mathbf{c}}\big(R(\mathbf{a},\mathbf{b})\mathbf{v}\big) - R(\mathbf{a},\mathbf{b})\nabla_{\mathbf{c}}\mathbf{v} - R([\mathbf{a},\mathbf{b}],\mathbf{c})\mathbf{v}. \end{aligned}$$

Adding up the cyclic permutations of the last expression in $(\mathbf{a}, \mathbf{b}, \mathbf{c})$ and using Eq. (1.44) and the obvious property $R(\mathbf{a},\mathbf{b})\mathbf{v} = -R(\mathbf{b},\mathbf{a})\mathbf{v}$, we find all the terms required for the Bianchi identity. ■

1.8.3 Ricci tensor and scalar

The **Ricci tensor** $R_{\mu\nu}$ is the bilinear form $\mathrm{Ric}(\mathbf{a}, \mathbf{b})$ defined as the trace of $R(\mathbf{a}, \mathbf{x}, \mathbf{b}, \mathbf{y})$ with respect to $\mathbf{x}$ and $\mathbf{y}$,

$$\mathrm{Ric}(\mathbf{a}, \mathbf{b}) = \mathrm{Tr}_{(\mathbf{x},\mathbf{y})} R(\mathbf{a}, \mathbf{x}, \mathbf{b}, \mathbf{y}).$$

The Ricci tensor is a symmetric bilinear form due to the property (1.68). In the index notation, this contraction of the Riemann tensor is written as

$$R_{\mu\nu} = R_{\mu\alpha\nu\beta} g^{\alpha\beta}.$$

The **Ricci scalar** is the trace of the Ricci tensor,

$$R \equiv \mathrm{Tr}_{(\mathbf{a},\mathbf{b})} \mathrm{Ric}(\mathbf{a}, \mathbf{b}) \equiv g^{\mu\nu} R_{\mu\nu}.$$

It is sometimes more convenient to use the index notation for calculations with the Riemann and Ricci tensors. The following examples illustrate the typical calculations. We will use both the index and the index-free notations.

The **Einstein equation** relates the Ricci tensor $R_{\mu\nu}$ to the energy-momentum tensor of matter, $T_{\mu\nu}$, as follows,

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = -8\pi G T_{\mu\nu}, \tag{1.70}$$

where G is Newton's constant. According to this equation, curvature of the spacetime is related to the energy density, velocity, and pressure of matter.

Statement 1.8.3.1: It follows from the Einstein equation (1.70) and the second Bianchi identity that $T_{\mu\nu}$ is always **covariantly conserved**, $\nabla^\mu T_{\mu\nu} = 0$.

Proof of Statement 1.8.3.1: Contracting Eq. (1.69) with $g^{\kappa\nu} g^{\mu\rho}$, we can easily show that $\nabla^\mu (R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R) = 0$. ■

Example: We consider a four-dimensional spacetime with a known metric g. Consider the following relationship for a symmetric tensor $A_{\mu\nu}$, similar to Eq. (1.70):

$$A_{\mu\nu} + a g_{\mu\nu} A^\alpha_\alpha = S_{\mu\nu},$$

where a is a given constant, $S_{\mu\nu}$ is a given symmetric tensor, and the trace A^α_α is defined by $A^\alpha_\alpha \equiv A_{\mu\nu} g^{\mu\nu}$. We would like to obtain an explicit expression for $A_{\mu\nu}$ in terms of $S_{\mu\nu}$.

In the index-free notation, A is a symmetric bilinear form satisfying

$$A + a g \mathrm{Tr}_{(\mathbf{a},\mathbf{b})} A(\mathbf{a}, \mathbf{b}) = S,$$

or more explicitly

$$A(\mathbf{x}, \mathbf{y}) + a g(\mathbf{x}, \mathbf{y}) \mathrm{Tr}_{(\mathbf{a},\mathbf{b})} A(\mathbf{a}, \mathbf{b}) = S(\mathbf{x}, \mathbf{y}).$$

The unknown tensor A would be readily found from this equation if we knew its trace. So let us compute the trace of both parts with respect to $(\mathbf{x}, \mathbf{y})$; since the trace of g is 4, we find

$$\mathrm{Tr}_{(\mathbf{x},\mathbf{y})} A(\mathbf{x}, \mathbf{y}) + 4a \mathrm{Tr}_{(\mathbf{a},\mathbf{b})} A(\mathbf{a}, \mathbf{b}) = \mathrm{Tr}_{(\mathbf{x},\mathbf{y})} S(\mathbf{x}, \mathbf{y}).$$

It follows that

$$\mathrm{Tr}_{(\mathbf{x},\mathbf{y})} A(\mathbf{x}, \mathbf{y}) = \frac{1}{1+4a} \mathrm{Tr}_{(\mathbf{x},\mathbf{y})} S(\mathbf{x}, \mathbf{y}),$$

and therefore

$$A = S - \frac{a}{1+4a} g.$$

In the index notation, the calculation looks as follows:

$$\begin{aligned} S_{\mu\nu} g^{\mu\nu} &= \left(A_{\mu\nu} + a g_{\mu\nu} A^\alpha_\alpha\right) g^{\mu\nu} = A^\alpha_\alpha + a g_{\mu\nu} g^{\mu\nu} A^\alpha_\alpha \\ &= (1 + 4a)\, A^\alpha_\alpha, \end{aligned}$$

therefore

$$A_{\mu\nu} = S_{\mu\nu} - \frac{a}{1+4a} g_{\mu\nu} S^\lambda_\lambda.$$

Note that there may be *no solutions*, or at any rate no unique expression for A in terms of S, when $a = -\frac{1}{4}$. ■

Practice problem: In the Einstein-Cartan theory (which is a modification of GR), the torsion tensor $T^\lambda_{\mu\nu}$ is related to the spin density tensor $S^\lambda_{\mu\nu}$ of matter by the equation

$$T^\lambda_{\mu\nu} + \delta^\lambda_\mu T^\rho_{\nu\rho} - \delta^\lambda_\nu T^\rho_{\mu\rho} = 8\pi G S^\lambda_{\mu\nu}.$$

Consider a more general relationship,

$$A^\lambda_{\mu\nu} + a\left(\delta^\lambda_\mu A^\rho_{\nu\rho} - \delta^\lambda_\nu A^\rho_{\mu\rho}\right) = S^\lambda_{\mu\nu},$$

where a is a given constant and $S^\lambda_{\mu\nu}$ is a given tensor (both $A^\lambda_{\mu\nu}$ and $S^\lambda_{\mu\nu}$ are antisymmetric in the lower indices). Obtain an explicit expression for $A^\lambda_{\mu\nu}$ in terms of $S^\lambda_{\mu\nu}$.

Hint: Compute a suitable trace of the left-hand side of the given equation.

Answer: If $a \neq -\frac{1}{3}$, the solution is

$$A^\lambda_{\mu\nu} = S^\lambda_{\mu\nu} - \frac{a}{1+3a}\left(\delta^\lambda_\mu S^\rho_{\nu\rho} - \delta^\lambda_\nu S^\rho_{\mu\rho}\right).$$

■

1.8.4 Calculations with the curvature tensor

We will now perform some further calculations involving the curvature tensor. The index-free notation will be used.

Statement 1.8.4.1: A Killing vector $\mathbf{k}$ has the following property (written in index notation),

$$\nabla_\mu \nabla_\nu k_\alpha = R_{\nu\alpha\mu}{}^\beta k_\beta.$$

The index-free version of the given identity is

$$g(\nabla_{\mathbf{a}} \nabla_{\mathbf{b}} \mathbf{k}, \mathbf{c}) - g(\nabla_{\nabla_{\mathbf{a}} \mathbf{b}} \mathbf{k}, \mathbf{c}) = R(\mathbf{b}, \mathbf{c}, \mathbf{a}, \mathbf{k}). \tag{1.71}$$

Outline of proof: We use the first Bianchi identity and the definition of a Killing vector. To simplify the calculations, we assume that first derivatives of the auxiliary vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$ vanish.

Proof of Statement 1.8.4.1: To obtain the index-free expression shown above, we contract the given indexed formula with arbitrary vectors a^μ, b^ν, c^α, and for convenience raise the index on k_α; we find

$$\begin{aligned} a^\mu b^\nu c_\alpha \nabla_\mu \nabla_\nu k^\alpha &= c^\beta g_{\alpha\beta} \left[a^\mu \nabla_\mu (b^\nu \nabla_\nu k^\alpha) - (a^\mu \nabla_\mu b^\nu)(\nabla_\nu k^\alpha) \right] \\ &= g(\mathbf{c}, \nabla_{\mathbf{a}} \nabla_{\mathbf{b}} \mathbf{k}) - g(\mathbf{c}, \nabla_{\nabla_{\mathbf{a}} \mathbf{b}} \mathbf{k}); \\ a^\mu b^\nu c^\alpha R_{\nu\alpha\mu}{}^\beta k_\beta &= R(\mathbf{b}, \mathbf{c}, \mathbf{a}, \mathbf{k}). \end{aligned}$$

We note that both sides of Eq. (1.71) depend only on the values of the auxiliary vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$ but not on their derivatives. So we may simplify calculations by choosing these auxiliary vectors such that all the first derivatives (e.g. $\nabla_{\mathbf{a}} \mathbf{b}$) vanish at a given point. Then we rewrite Eq. (1.71), which we need to prove for vector fields $\mathbf{a}, \mathbf{b}, \mathbf{c}$ with vanishing derivatives, as

$$g(\nabla_{\mathbf{a}} \nabla_{\mathbf{b}} \mathbf{k}, \mathbf{c}) = R(\mathbf{b}, \mathbf{c}, \mathbf{a}, \mathbf{k}). \tag{1.72}$$

By assumption, the vector $\mathbf{k}$ satisfies the Killing equation (for arbitrary $\mathbf{x}, \mathbf{y}$),

$$g(\nabla_{\mathbf{x}} \mathbf{k}, \mathbf{y}) = -g(\nabla_{\mathbf{y}} \mathbf{k}, \mathbf{x}).$$

The plan is to convert $\nabla_{\mathbf{b}} \mathbf{k}$ into $\nabla_{\mathbf{c}} \mathbf{k}$ using the Killing equation, to obtain a term $g(\nabla_{\mathbf{a}} \nabla_{\mathbf{c}} \mathbf{k}, \mathbf{b})$, and to express it through $R(\mathbf{a}, \mathbf{c}, \mathbf{k}, \mathbf{b})$, which essentially differs from what we need on the right side of Eq. (1.72) by a cyclic permutation of $\mathbf{a}, \mathbf{b}, \mathbf{c}$. To implement this plan, we need to move $\nabla_{\mathbf{a}}$ temporarily to the outside of $g(...)$, so that we can use the Killing equation:

$$\begin{aligned} g(\nabla_{\mathbf{a}} \nabla_{\mathbf{b}} \mathbf{k}, \mathbf{c}) &= \nabla_{\mathbf{a}} g(\nabla_{\mathbf{b}} \mathbf{k}, \mathbf{c}) - g(\nabla_{\mathbf{b}} \mathbf{k}, \nabla_{\mathbf{a}} \mathbf{c}); \\ \nabla_{\mathbf{a}} g(\nabla_{\mathbf{b}} \mathbf{k}, \mathbf{c}) &= -\nabla_{\mathbf{a}} g(\nabla_{\mathbf{c}} \mathbf{k}, \mathbf{b}) \\ &= -g(\nabla_{\mathbf{a}} \nabla_{\mathbf{c}} \mathbf{k}, \mathbf{b}) - g(\nabla_{\mathbf{c}} \mathbf{k}, \nabla_{\mathbf{a}} \mathbf{b}). \end{aligned}$$

Since the first derivatives of $\mathbf{a}, \mathbf{b}, \mathbf{c}$ vanish by assumption, we have simply

$$g(\nabla_{\mathbf{a}} \nabla_{\mathbf{b}} \mathbf{k}, \mathbf{c}) = -g(\nabla_{\mathbf{a}} \nabla_{\mathbf{c}} \mathbf{k}, \mathbf{b}).$$

We replace $\nabla_{\mathbf{a}} \nabla_{\mathbf{c}} \mathbf{k}$ with the Riemann tensor,

$$\begin{aligned} g(\nabla_{\mathbf{a}} \nabla_{\mathbf{c}} \mathbf{k}, \mathbf{b}) &= R(\mathbf{a}, \mathbf{c}, \mathbf{k}, \mathbf{b}) + g(\nabla_{\mathbf{c}} \nabla_{\mathbf{a}} \mathbf{k}, \mathbf{b}) + g(\nabla_{[\mathbf{a},\mathbf{c}]} \mathbf{k}, \mathbf{b}) \\ &= R(\mathbf{a}, \mathbf{c}, \mathbf{k}, \mathbf{b}) + g(\nabla_{\mathbf{c}} \nabla_{\mathbf{a}} \mathbf{k}, \mathbf{b}), \end{aligned}$$

since $[\mathbf{a}, \mathbf{c}] = 0$. Thus we have derived the property

$$g(\nabla_{\mathbf{a}} \nabla_{\mathbf{b}} \mathbf{k}, \mathbf{c}) = R(\mathbf{c}, \mathbf{a}, \mathbf{k}, \mathbf{b}) - g(\nabla_{\mathbf{c}} \nabla_{\mathbf{a}} \mathbf{k}, \mathbf{b}),$$

where we cosmetically rearranged the order of the vectors under $R(...)$ to make the bookkeeping easier. Now we note that the structure of the last term on the right-hand side above is exactly the same as that of the left-hand side, except for a cyclic permutation of the vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$. Thus we can apply this relation twice more, until the cyclic permutation $abkc \to cakb \to bcka \to abkc$ is complete. So we find

$$\begin{aligned} g(\nabla_{\mathbf{a}} \nabla_{\mathbf{b}} \mathbf{k}, \mathbf{c}) = R(\mathbf{c}, \mathbf{a}, \mathbf{k}, \mathbf{b}) - R(\mathbf{b}, \mathbf{c}, \mathbf{k}, \mathbf{a}) \\ + R(\mathbf{a}, \mathbf{b}, \mathbf{k}, \mathbf{c}) - g(\nabla_{\mathbf{a}} \nabla_{\mathbf{b}} \mathbf{k}, \mathbf{c}). \end{aligned}$$

Using the first Bianchi identity,

$$R(\mathbf{c}, \mathbf{a}, \mathbf{k}, \mathbf{b}) + R(\mathbf{a}, \mathbf{b}, \mathbf{k}, \mathbf{c}) + R(\mathbf{b}, \mathbf{c}, \mathbf{k}, \mathbf{a}) = 0,$$

we obtain

$$2g(\nabla_{\mathbf{a}} \nabla_{\mathbf{b}} \mathbf{k}, \mathbf{c}) = -2R(\mathbf{b}, \mathbf{c}, \mathbf{k}, \mathbf{a}),$$

which coincides with Eq. (1.72). ■

As another application of the index-free method of calculation, we will derive the formula for the change in the Riemann tensor under a conformal transformation, $g \to \tilde{g} \equiv e^{2\lambda} g$, where $\lambda(x)$ is a fixed scalar function. The nonzero function $e^{2\lambda}$ is called the **conformal factor**. Since the resulting expression contains second derivatives of the conformal factor, it will follow that the Riemann tensor depends on second derivatives of the metric.

Calculation 1.8.4.1: Consider a conformal transformation of the metric, $g \to \tilde{g} \equiv e^{2\lambda} g$. It can be derived from Eq. (1.47) that the Riemann and Ricci tensors of the new metric $\tilde{g}$ are related to the old tensors in the following way. The new Riemann tensor is

$$\begin{aligned} e^{-2\lambda} \tilde{R}(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) &= R(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) \\ &+ H_\lambda(\mathbf{a}, \mathbf{c}) g(\mathbf{b}, \mathbf{d}) - H_\lambda(\mathbf{a}, \mathbf{d}) g(\mathbf{b}, \mathbf{c}) \\ &- H_\lambda(\mathbf{b}, \mathbf{c}) g(\mathbf{a}, \mathbf{d}) + H_\lambda(\mathbf{b}, \mathbf{d}) g(\mathbf{a}, \mathbf{c}) \\ &+ \det \begin{vmatrix} g(\mathbf{a}, \mathbf{c}) & g(\mathbf{b}, \mathbf{c}) & g(\mathbf{l}, \mathbf{c}) \\ g(\mathbf{a}, \mathbf{d}) & g(\mathbf{b}, \mathbf{d}) & g(\mathbf{l}, \mathbf{d}) \\ g(\mathbf{a}, \mathbf{l}) & g(\mathbf{b}, \mathbf{l}) & g(\mathbf{l}, \mathbf{l}) \end{vmatrix}, \end{aligned}$$

where $\mathbf{l} \equiv \hat{g}^{-1} \mathrm{d}\lambda$ and the symmetric bilinear form

$$H_\lambda(\mathbf{a}, \mathbf{b}) = H_\cdot(\mathbf{b}, \mathbf{a}) \equiv g(\nabla_{\mathbf{a}} \mathbf{l}, \mathbf{b}) \equiv \iota_{\mathbf{b}} \nabla_{\mathbf{a}} \mathrm{d}\lambda$$

is called the **Hessian** of the function λ. The Hessian is a tensor representing all the (covariant) second derivatives of λ defined with respect to the old metric g. In the index notation, the Hessian of λ is the tensor $\lambda_{;\alpha\beta}$.

The new Ricci tensor is

$$\begin{aligned} \tilde{\mathrm{Ric}}\,(\mathbf{a}, \mathbf{c}) &= \mathrm{Ric}\,(\mathbf{a}, \mathbf{c}) + g\,(\mathbf{a}, \mathbf{c})\,\Box\lambda \\ &+ (N-2)\left[H_\lambda(\mathbf{a}, \mathbf{c}) + g(\mathbf{l}, \mathbf{l}) g(\mathbf{a}, \mathbf{c}) - g(\mathbf{a}, \mathbf{l}) g(\mathbf{c}, \mathbf{l})\right], \end{aligned}$$

where

$$\Box\lambda \equiv \mathrm{Tr}_{(\mathbf{a},\mathbf{b})} H_\lambda(\mathbf{a}, \mathbf{b})$$

denotes the covariant D'Alembert operator (the **D'Alembertian**) defined with respect to the old metric g. Note that

$$g(\mathbf{l}, \mathbf{l}) \equiv g^{-1}(\mathrm{d}\lambda, \mathrm{d}\lambda).$$

The new Ricci scalar is

$$\begin{aligned} \tilde{R} &= e^{-2\lambda} \mathrm{Tr}_{(\mathbf{a},\mathbf{c})} \tilde{\mathrm{Ric}}\,(\mathbf{a}, \mathbf{c}) \\ &= e^{-2\lambda} \left[R + 2\,(N-1)\,\Box\lambda + (N-1)\,(N-2) g^{-1}(\mathrm{d}\lambda, \mathrm{d}\lambda)\right]. \end{aligned}$$

In the index notation,

$$\Box\lambda = g^{\mu\nu}\lambda_{;\mu\nu}, \quad g^{-1}(\mathrm{d}\lambda, \mathrm{d}\lambda) = g^{\alpha\beta}\lambda_{,\alpha}\lambda_{,\beta}.$$

(Details on page 177.) ■

1.9 Geodesic curves, geodesic vector fields

We have seen before that there is no a priori relation between tangent vectors at different points of the manifold $\mathcal{M}$. In other words, there is no "intrinsic" (i.e. naturally defined) way to carry a vector or a tensor from one point to another. However, we can use a connection ∇ to define the notion of a direction-preserving, or "parallel," transport.

1.9.1 Parallel transport of vectors

A vector field $\mathbf{v}$ is called **parallelly transported** along a curve $\gamma(\tau)$ if the covariant derivative of $\mathbf{v}$ along the curve vanishes,

$$\nabla_{\dot\gamma}\mathbf{v} = 0. \tag{1.73}$$

The interpretation is that we carry a vector $\mathbf{v}$ along the curve γ while keeping the magnitude and direction of $\mathbf{v}$ constant. Heuristically, the only way to control the constancy of $\mathbf{v}$ is through the covariant derivative in the direction $\dot\gamma$, hence the requirement (1.73).

Let us compare this condition with a similar one involving the Lie derivative,

$$\mathcal{L}_{\dot\gamma}\mathbf{v} = 0. \tag{1.74}$$

A vector $\mathbf{v}$ satisfying the condition (1.74) is a **connecting vector** for the congruence of curves to which $\gamma(\tau)$ belongs; it is also called **Lie-propagated** along the curve. The condition (1.73) depends only on the direction of $\dot\gamma$ along the curve, while the Lie derivative in Eq. (1.74) requires us to have a vector field $\dot\gamma$ defined also *around* the curve $\gamma(\tau)$. In other words, we need a congruence of curves and not just one curve. Different choices of the congruence around γ will lead to different connecting vector fields $\mathbf{v}$. The geometric meaning is that $\mathbf{v}$ connects points on neighboring curves corresponding to the *same value* of the parameter τ.

To summarize: Transporting a vector from one point to another along a curve requires to have a relationship between tangent spaces at different points. One needs additional information to produce such a relationship. In the case of Eq. (1.73), this information comes from the chosen connection ∇; in the case of Eq. (1.74), this information comes from the neighboring curves from a congruence containing the given curve γ. Clearly, the construction of a parallel transport closely reflects the intuitive notion of a vector carried along a curve "without change."

Statement 1.9.1.1: The scalar product $g(\mathbf{u}, \mathbf{v})$ is constant along a curve γ if both $\mathbf{u}$ and $\mathbf{v}$ are parallelly transported along the curve. ■

Proof: A straightforward computation gives

$$\nabla_{\dot\gamma} g(\mathbf{u}, \mathbf{v}) = g(\nabla_{\dot\gamma}\mathbf{u}, \mathbf{v}) + g(\mathbf{u}, \nabla_{\dot\gamma}\mathbf{v}) = 0.$$

■

1.9.2 Geodesics

A natural question to ask about a curve $\gamma(\tau)$ is whether the curve parallelly transports its own tangent vector $\dot\gamma(\tau)$. If the tangent vector $\dot\gamma(\tau)$ is indeed parallelly transported along the curve, heuristically one could say that the curve $\gamma(\tau)$ is "straight" in the sense that it is a line whose direction remains unchanged along the line. Within a curved (nonflat) manifold, such lines $\gamma(\tau)$ are the closest analog of straight lines.

By definition, a curve $\gamma(\tau)$ is called a **geodesic** if its tangent vector $\dot\gamma$ is parallelly transported along the curve.

It is worth emphasizing that the parallel transport in GR is defined through the Levi-Civita connection determined by the given metric g. Properties of parallel transport, geodesics, and curvature are very different if a non-Levi-Civita connection is used. In this book we always denote by ∇ the Levi-Civita connection and we use no other connections, so the choice of connection will not be discussed any more.

If we temporarily assume that the tangent vector $\dot\gamma$ is a part of a vector field $\mathbf{v}$, the condition for being geodesic can be written as a differential equation on $\mathbf{v}$, called the **geodesic equation**,

$$\nabla_{\mathbf{v}}\mathbf{v} = 0, \tag{1.75}$$

which however must hold *only* at points $p = \gamma(\tau)$ along the curve. Note that the derivative $\nabla_{\mathbf{v}}\mathbf{v}$ is taken only in the direction of $\mathbf{v}$ itself, so it is sufficient to have the vector field $\mathbf{v}$ defined *only* at points of the curve γ. The derivative $\nabla_{\mathbf{v}}\mathbf{v}$ does not depend on the values of the field $\mathbf{v}$ off the curve γ, while on γ we have $\mathbf{v} = \dot\gamma$. Therefore, the assumption that $\dot\gamma$ is a part of a vector field $\mathbf{v}$ is inessential, and Eq. (1.75) makes sense also for the tangent vector $\dot\gamma$ defined only on the curve γ.

A vector field $\mathbf{v}$ that satisfies the geodesic equation (1.75) at *every* point (not merely along a single orbit) is called a **geodesic vector field**.

It is perhaps not immediately evident that the choice of the parameter τ is important for a geodesic curve, according to this definition. Namely, if $\gamma(\tau)$ is a geodesic and we change the parameter to $\tau = \tau(\sigma)$, then the new curve $\tilde\gamma(\sigma) \equiv \gamma(\tau(\sigma))$ will not necessarily be a geodesic. If the parameter τ is chosen so that $\gamma(\tau)$ is a geodesic then τ is called an **affine parameter** and the vector $\dot\gamma$ is called the **affine tangent vector**.

Statement 1.9.2.1: If $\gamma(\tau)$ is a geodesic and the parameter is changed by $\tau = \tau(\sigma)$, the new curve $\tilde\gamma(\sigma)$ is again a geodesic only if the function $\tau(\sigma)$ is of the form $\tau(\sigma) = a\sigma + b$ with constant a, b and $a \neq 0$. Thus the affine parameter is defined up to an **affine transformation**.

Proof of Statement 1.9.2.1: We have

$$\dot{\tilde\gamma} = \tau'(\sigma)\dot\gamma; \quad \nabla_{\dot{\tilde\gamma}}\dot{\tilde\gamma} = \tau'(\sigma)\left(\dot\gamma \circ \tau'(\sigma)\right)\dot\gamma, \tag{1.76}$$

which vanishes only if $\tau'(\sigma) = \text{const}$ along the curve. ■

The square of the tangent vector, $g(\dot\gamma, \dot\gamma)$, remains constant along a geodesic due to Statement 1.9.1.1. This means that a timelike geodesic remains timelike at all times, a null geodesic remains null, and a spacelike geodesic remains spacelike. It follows from Eq. (1.76) that an affine parameter along a non-null geodesic can always be chosen so that $g(\dot\gamma, \dot\gamma) = \pm 1$: It is sufficient to divide τ by the constant number $\sqrt{g(\dot\gamma, \dot\gamma)}$ to achieve this parameterization. But a null geodesic does not have a naturally fixed affine parameter τ; the freedom to replace $\tau \to a\tau + \beta$ remains for null geodesics.

Physical interpretation of geodesics: According to GR, massive point particles in freefall move along worldlines $\gamma(\tau)$ which are timelike geodesics in spacetime. In that case, it is standard to choose the affine parameter τ as the proper time measured along the particle's worldline, so that $g(\dot\gamma, \dot\gamma) = 1$. On the other hand, electromagnetic waves (in the geometrical optics approximation) can be pictured as rays propagating along null geodesic curves.[14] So I will call null geodesic curves **lightrays**. In a calculation involving lightrays whose frequency is being measured, it is possible to choose the affine parameter for a null geodesic $\gamma(\tau)$ such that the frequency ν measured at a point by an observer with 4-velocity $\mathbf{u}$ is numerically given by

$$\nu = g(\mathbf{u}, \dot\gamma). \tag{1.77}$$

In that case, $h\nu$ (where h is Planck's constant) is the locally measured energy of a single photon propagating along the null geodesic $\gamma(\tau)$. The 4-vector $h\dot\gamma$ can be interpreted as the 4-momentum of that photon.

The statements about the lightrays can be justified by considering a local coordinate system $\{x^\mu\}$ of an observer who has the 4-velocity $\mathbf{u}$ and measures the frequency of an electromagnetic wave near a spacetime point p. In a very small neighborhood of p, the wave looks like a plane wave and the metric g can be brought into the Minkowski form, $g \approx \eta$, by a choice of local coordinates. Then the electromagnetic potential A near the point p can be expressed by the familiar formula

$$A = A_0 \exp\left[i\eta(\mathbf{k}, \mathbf{x})\right],$$

where $\mathbf{k}$ is the wave vector, A_0 is the amplitude of the wave, and $\mathbf{x}$ is a vector representing the local coordinates. (The wave vector $\mathbf{k}$ is a null geodesic vector field.) In a reference frame of the observer, the frequency of the wave is the coefficient at t in the expression $\exp[2\pi i\nu t]$ in the phase of the wave; in other words,

$$\nu = \frac{1}{2\pi} k_0,$$

where k_0 is the time component of the 4-vector $\mathbf{k}$. The 4-velocity $\mathbf{u}$ has components $\{1, 0, 0, 0\}$ in the observer's reference frame. Therefore, the frequency ν can be computed as $\nu = \frac{1}{2\pi}\eta(\mathbf{k}, \mathbf{u})$ in the observer's reference frame. To obtain a formula for ν valid in an arbitrary coordinate system, we need to use the metric g instead of η; hence

$$\nu = \frac{1}{2\pi} g(\mathbf{k}, \mathbf{u}).$$

Since the affine parameter may be multiplied by a constant factor, we can choose the affine parameter τ along the null geodesic $\gamma(\tau)$ such that $\dot\gamma = \frac{1}{2\pi}\mathbf{k}$. Then Eq. (1.77) holds. ■

Remark: Since the geodesic equation $\nabla_{\dot\gamma}\dot\gamma = 0$ is a first-order equation with smooth coefficients,[15] it has a unique solution that depends only on the direction of the vector $\dot\gamma$ at an initial point p. Thus we have a map from vectors $\mathbf{v} \in T_p\mathcal{M}$ at a point p to geodesic lines starting at p. This map is called the **exponentiation** map $\exp : \mathbf{v} \to \gamma(\tau)$. The resulting geodesics are sometimes denoted $\gamma(\tau) = \exp(\mathbf{v}\tau)$, but we will not use this notation since it does not give a computational advantage. ■

A useful property of null geodesics is formulated in the following statement.

Statement 1.9.2.2: A null geodesic $\gamma(\tau)$ remains a geodesic when the metric is rescaled by a conformal transformation, $\tilde g = e^{2\lambda} g$, under a suitable change of the affine parameter $\tau \to f(\tau)$, such that the new affine tangent vector is $e^{-2\lambda}\dot\gamma$. In other words, the shape of a null geodesic is **conformally invariant**.

Idea of proof: Use Eq. (1.47).

Proof of Statement 1.9.2.2: Suppose that the new affine tangent vector is $\tilde{\mathbf{n}} = \phi\mathbf{n}$, where ϕ is an unknown function and $\mathbf{n} = \dot\gamma$. It follows from Eq. (1.47) and $\nabla_{\mathbf{n}}\mathbf{n} = 0$ that, for arbitrary $\mathbf{x}$,

$$\begin{aligned}\tilde g(\tilde\nabla_{\phi\mathbf{n}}(\phi\mathbf{n}), \mathbf{x}) &= 2\phi^2 (\nabla_{\mathbf{n}}\lambda)\, \tilde g(\mathbf{n}, \mathbf{x}) + \tilde g(\nabla_{\phi\mathbf{n}}(\phi\mathbf{n}), \mathbf{x}) \\ &= \tilde g(\mathbf{n}, \mathbf{x})\phi^2 (2\nabla_{\mathbf{n}}\lambda + \nabla_{\mathbf{n}} \ln\phi).\end{aligned}$$

This will vanish if $\phi = e^{-2\lambda}$. The new affine parameter $\tilde\tau$ is found by integrating ϕ along the curve γ. ■

1.9.3 Geodesics extremize proper length

In Riemannian geometry, when the metric is positive-definite, sufficiently short geodesic curves have an important property: they are curves of shortest length among all curves connecting two given points. In GR, the metric is not positive-definite, so geodesic curves extremize the *proper length*. In this section we derive the geodesic equation from the corresponding variational principle.

The **proper length** of a curve segment $\gamma(\tau)$ is

$$L[\gamma] = \int_{\tau_1}^{\tau_2} \sqrt{g(\dot\gamma, \dot\gamma)}\, d\tau. \tag{1.78}$$

The proper length is *extremized* (but not necessarily maximized or minimized) if the functional $L[\gamma]$ does not change in first order in the perturbation when the curve γ is infinitesimally perturbed while the endpoints $\gamma(\tau_1)$, $\gamma(\tau_2)$ remain fixed. We now apply the standard considerations of the variational calculus to this condition and derive the corresponding differential equation for $\dot\gamma$.

A perturbation of the curve γ can be described as the flow of a vector field $\mathbf{t}$ which, we imagine, is directed "transversely" to γ and leaves the endpoints $\gamma(\tau_{1,2})$ in place. Let us denote by σ a parameter along the lines of $\mathbf{t}$, starting from the curve γ. Thus, the flow of $\mathbf{t}$ transforms the curve $\gamma(\tau)$ into a curve $\tilde\gamma(\tau; \sigma)$ when we let the lines of $\mathbf{t}$ carry the points of γ for a parameter distance σ. Denote by $\mathbf{v}$ the vector field obtained by taking tangent vectors to various curves $\tilde\gamma$ transported the initial curve $\gamma(\tau)$ by various distances σ along the flow of $\mathbf{t}$. By construction, $\mathbf{v}$ is a connecting vector for $\mathbf{t}$, and so $\nabla_{\mathbf{t}}\mathbf{v} = \nabla_{\mathbf{v}}\mathbf{t}$. The proper length of a perturbed curve can now be seen as a function of σ,

$$L[\tilde\gamma; \sigma] = \int_{\tilde\gamma} \sqrt{g(\mathbf{v}, \mathbf{v})}\, d\tau,$$

where the integration is performed along the curve $\tilde\gamma(\tau; \sigma)$.

According to the variational principle, the curve γ will extremize the proper length if $dL/d\sigma = 0$ at $\sigma = 0$. The operation $d/d\sigma$ corresponds to applying (under the integral sign) a

[14] This follows from the equations of Maxwell electrodynamics in curved space. In this book I will not derive this property.

[15] This property may not hold at a point where the manifold is not smooth, i.e. at a physical singularity. In this book I do not attempt to consider non-smooth manifolds and do not examine the precise conditions for smoothness.

derivative $\mathcal{L}_\mathbf{t}$ along the flow of $\mathbf{t}$,

$$\frac{d}{d\sigma}L[\tilde{\gamma};\sigma] = \int_{\tilde{\gamma}} \nabla_\mathbf{t}\sqrt{g(\mathbf{v},\mathbf{v})}d\tau = \int_{\tilde{\gamma}} \frac{g(\nabla_\mathbf{v}\mathbf{t},\mathbf{v})}{\sqrt{g(\mathbf{v},\mathbf{v})}}d\tau.$$

Here we have used the property $\nabla_\mathbf{t}\mathbf{v} = \nabla_\mathbf{v}\mathbf{t}$. The last expression above does not contain derivatives of $\mathbf{v}$ and is evaluated on the initial curve γ where $\sigma = 0$, hence

$$\frac{d}{d\sigma}L[\tilde{\gamma};\sigma]\bigg|_{\sigma=0} = \int_\gamma \frac{g(\nabla_\mathbf{v}\mathbf{t},\mathbf{v})}{\sqrt{g(\mathbf{v},\mathbf{v})}}d\tau. \tag{1.79}$$

The curve γ extremizes the proper length if the above integral vanishes for all $\mathbf{t}$.

The expression in Eq. (1.79) contains an undesirable derivative of $\mathbf{t}$. To eliminate that derivative, we now perform the usual trick consisting of integration by parts. Since $\nabla_\mathbf{v}$ acts as a total derivative $\frac{d}{d\tau}$ along γ, while $\mathbf{t} = 0$ at both endpoints of the curve, we have

$$\int_\gamma \nabla_\mathbf{v} g(\mathbf{t}, \frac{\mathbf{v}}{\sqrt{g(\mathbf{v},\mathbf{v})}})d\tau = 0.$$

Subtracting this from Eq. (1.79), we obtain

$$\frac{d}{d\sigma}L[\tilde{\gamma};\sigma]\bigg|_{\sigma=0} = -\int_\gamma g(\mathbf{t}, \nabla_\mathbf{v}\frac{\mathbf{v}}{\sqrt{g(\mathbf{v},\mathbf{v})}})d\tau.$$

This can vanish for all $\mathbf{t}(\tau)$ only if

$$\nabla_\mathbf{v}\frac{\mathbf{v}}{\sqrt{g(\mathbf{v},\mathbf{v})}}\bigg|_{\gamma(\tau)} = \nabla_{\dot\gamma}\frac{\dot\gamma}{\sqrt{g(\dot\gamma,\dot\gamma)}} = 0.$$

This is almost the geodesic equation, except for the normalization of the tangent vector $\dot\gamma$. We now need to assume that $\gamma(\tau)$ is not null. Then a redefinition of the parameter $\tau \to f(\tau)$ can be used to set $g(\dot\gamma,\dot\gamma) = \text{const}$ along γ. Hence we find $\nabla_\mathbf{v}\mathbf{v} = 0$, which is precisely the geodesic equation (1.75). Note that the resulting geodesic equation is *not* invariant under arbitrary changes of parameter $\tau \to f(\tau)$ because we have already assumed a fixed normalization of the tangent vector $\dot\gamma$.

Practice problem: Suppose that a vector field $\mathbf{v}$ satisfies the equation $\nabla_\mathbf{v}\mathbf{v} = \mu\mathbf{v}$ for some given scalar function μ. Show that $\mathbf{v}$ can be made geodesic by a rescaling $\mathbf{v} \to \lambda\mathbf{v}$. Obtain an explicit expression for λ. ■

Null geodesics. In the derivation above, we assumed that $g(\dot\gamma,\dot\gamma) \neq 0$; let us now consider the case of **null curves**, i.e. curves $\gamma(\tau)$ such that $g(\dot\gamma,\dot\gamma) = 0$ along the curve. Such null geodesics *cannot* be obtained by extremizing the functional (1.78), because a variation (even an infinitesimal one) of a null curve $\gamma(\tau)$ will make some portions of the curve spacelike and other portions timelike, and then $\sqrt{g(\dot\gamma,\dot\gamma)}$ will not remain well-defined. To avoid this difficulty, one can simply postulate Eq. (1.75) for null geodesics. Alternatively, a different variational principle can be used to derive the geodesic equation (1.75).

Consider the variational problem with the functional

$$L[\gamma(\tau);N(\tau)] = \int_{\tau_1}^{\tau_2} \frac{g(\dot\gamma,\dot\gamma)}{N(\tau)}d\tau, \tag{1.80}$$

depending on an additional unknown function $N(\tau)$. Since the functional does not contain derivatives of N, this variable plays the role of a Lagrange multiplier. The variation with respect to $N(\tau)$ yields the constraint $g(\dot\gamma,\dot\gamma) = 0$, while the variation with respect to $\gamma(\tau)$ yields the equation

$$N\nabla_\mathbf{v}\mathbf{v} = -(\nabla_\mathbf{v}N)\,\mathbf{v} + \frac{1}{2}g(\mathbf{v},\mathbf{v})\hat{g}^{-1}\mathrm{d}N.$$

Using the constraint $g(\mathbf{v},\mathbf{v}) = 0$ and choosing a function $N(\tau)$ that is constant along the flow of $\mathbf{v}$, one recovers the geodesic equation (1.75). When the Lagrange multiplier $N(\tau)$ is chosen arbitrarily, one obtains the same geodesic curve in a non-affine parameterization.

The variational principle of Eq. (1.80) can be generalized to the case of timelike or spacelike geodesics in the following way. One writes the functional

$$L[\gamma(\tau);N(\tau)] \equiv \int_{\tau_1}^{\tau_2}\left[\frac{g(\dot\gamma,\dot\gamma)}{N} + KN\right]d\tau,$$

where K is a constant and $N(\tau)$ is a Lagrange multiplier. Since N is a function only of τ, we have $\mathcal{L}_\mathbf{t}N = 0$ for a transverse perturbation field $\mathbf{t}$. (orbits of $\mathbf{t}$ connect points at equal τ.) Variation with respect to $N(\tau)$ yields the constraint $g(\dot\gamma,\dot\gamma) = KN^2$, while variation with respect to $\gamma(\tau)$ yields the non-affine geodesic equation

$$\nabla_{\dot\gamma}\frac{\dot\gamma}{N} = 0.$$

The value of K can be chosen freely to control the causal character of the curve. Since K is never in the denominator, while N can be chosen freely, the value $K = 0$ is perfectly admissible. In this way, geodesics of every kind are described through a single variational principle. ■

1.9.4 *Motion under external forces

In classical physics, equations of motion for particles and fields are derived from an **action principle**: The correct trajectories must extremize a certain functional called the **action functional**. As an example, we now derive the equations of motion for a charged massive particle in the presence of an external electromagnetic field in curved spacetime.[16]

The motion of a massive particle corresponds to a timelike worldline $\gamma(\tau)$ in spacetime. The electromagnetic field is described by a 1-form $\boldsymbol{A}$, and the gravitational field by the metric g. We will assume that these fields are fixed, and the question is to determine the trajectory $\gamma(\tau)$ of the particle.

The action functional for the particle is

$$S[\gamma;\boldsymbol{A},g] = -m\int_{\tau_1}^{\tau_2}\sqrt{g(\dot\gamma,\dot\gamma)}d\tau + q\int_{\tau_1}^{\tau_2}(\boldsymbol{A}\circ\dot\gamma)\,d\tau,$$

where m is the rest mass and q is the electric charge of the particle. The first term in the action is m times the proper time along the trajectory, while the second term is simply the integral of the 1-form $\boldsymbol{A}$ along the worldline, which may be written more concisely as $\int_\gamma \boldsymbol{A}$ (see Eq. (1.18) and the preceding explanations).

The equations of motion for the particle are derived by extremizing the functional S with respect to the trajectory $\gamma(\tau)$. Although the actual computation is fairly short, I will go through it slowly and show every step.

[16] I called the electromagnetic field **external** to emphasize that this field is assumed to be given and fixed, rather than determined dynamically.

As in Sec. 1.9.3, we introduce a vector field $\mathbf{t}$ which is transverse to the curve γ and vanishes at $\tau = \tau_{1,2}$. The vector field $\mathbf{t}$ perturbs the curve γ by shifting it to the side. Then we construct the vector field $\mathbf{v}$ by shifting the curve γ along the orbits of $\mathbf{t}$ by different parameter distances σ. The variation of the functional S is found by applying the derivative $\mathcal{L}_\mathbf{t}$ under the integral,

$$\frac{d}{d\sigma} S[\gamma; \boldsymbol{A}, g] = \int_{\tau_1}^{\tau_2} \mathcal{L}_\mathbf{t} \left(-m\sqrt{g(\mathbf{v},\mathbf{v})} + q\boldsymbol{A} \circ \mathbf{v} \right) d\tau. \quad (1.81)$$

The terms under the integral are simplified using $\mathcal{L}_\mathbf{t}\mathbf{v} = 0$ and the Leibnitz property of the Lie derivative. The first term yields

$$\mathcal{L}_\mathbf{t}\sqrt{g(\mathbf{v},\mathbf{v})} = \frac{(\mathcal{L}_\mathbf{t} g)\circ(\mathbf{v},\mathbf{v})}{2\sqrt{g(\mathbf{v},\mathbf{v})}} \overset{1}{=} \frac{g(\nabla_\mathbf{v}\mathbf{t},\mathbf{v})}{\sqrt{g(\mathbf{v},\mathbf{v})}}$$
$$\overset{2}{=} -g(\mathbf{t}, \nabla_\mathbf{v} \frac{\mathbf{v}}{\sqrt{g(\mathbf{v},\mathbf{v})}}) + \nabla_\mathbf{v} \frac{g(\mathbf{t},\mathbf{v})}{\sqrt{g(\mathbf{v},\mathbf{v})}}$$

($\overset{1}{=}$ is via Eq. (1.48), $\overset{2}{=}$ is to separate a total derivative). The integral of the total derivative $\nabla_\mathbf{v}(...)$ vanishes since $\mathbf{t}$ vanishes at the endpoints $\tau_{1,2}$. Thus the first integral term in the right-hand side of Eq. (1.81) is simplified to

$$m\int_{\tau_1}^{\tau_2} g(\mathbf{t}, \nabla_\mathbf{v} \frac{\mathbf{v}}{\sqrt{g(\mathbf{v},\mathbf{v})}}) d\tau.$$

The second term in the right-hand side of Eq. (1.81) becomes

$$\mathcal{L}_\mathbf{t}(\boldsymbol{A}\circ\mathbf{v}) = (\mathcal{L}_\mathbf{t}\boldsymbol{A})\circ\mathbf{v},$$

and then the line integral can be transformed as

$$\int_{\tau_1}^{\tau_2} (\mathcal{L}_\mathbf{t}\boldsymbol{A})\circ\mathbf{v}\, d\tau \overset{1}{=} \int_\gamma \mathcal{L}_\mathbf{t}\boldsymbol{A} \overset{2}{=} \int_\gamma (d\iota_\mathbf{t}\boldsymbol{A} + \iota_\mathbf{t} d\boldsymbol{A})$$
$$\overset{3}{=} \int_\gamma \iota_\mathbf{t} d\boldsymbol{A} \overset{4}{=} \int_{\tau_1}^{\tau_2} (d\boldsymbol{A})\circ(\mathbf{t},\mathbf{v})\, d\tau$$
$$= -\int_{\tau_1}^{\tau_2} \{\iota_\mathbf{t}\iota_\mathbf{v}(d\boldsymbol{A})\}\, d\tau$$

($\overset{1}{=}$ is by definition (1.18) of integrals of 1-forms over curves, $\overset{2}{=}$ is by the Cartan homotopy formula (1.24), $\overset{3}{=}$ is due to $\int_\gamma df = 0$ for a function f that vanishes at the endpoints of the curve, and $\overset{4}{=}$ is by definition of $\iota_\mathbf{t}$). Putting the two terms together and evaluating the derivatives on the initial curve γ, we find

$$\frac{d}{d\sigma} S[\gamma; \boldsymbol{A}, g]\Big|_{\sigma=0} = \int_{\tau_1}^{\tau_2} mg(\mathbf{t}, \nabla_\mathbf{v} \frac{\mathbf{v}}{\sqrt{g(\mathbf{v},\mathbf{v})}}) d\tau$$
$$- \int_{\tau_1}^{\tau_2} q\{\iota_\mathbf{t}\iota_\mathbf{v}(d\boldsymbol{A})\}\, d\tau.$$

The integrand above is linear in the vector field $\mathbf{t}$, and thus can be expressed as an auxiliary 1-form $\boldsymbol{V}$ applied to $\mathbf{t}$,

$$\frac{d}{d\sigma} S[\gamma; \boldsymbol{A}, g]\Big|_{\sigma=0} = \int_{\tau_1}^{\tau_2} (\iota_\mathbf{t}\boldsymbol{V})\, d\tau = \int_{\tau_1}^{\tau_2} (\boldsymbol{V}\circ\mathbf{t})\, d\tau,$$
$$\boldsymbol{V} \equiv m\hat{g}\nabla_\mathbf{v} \frac{\mathbf{v}}{\sqrt{g(\mathbf{v},\mathbf{v})}} - q\iota_\mathbf{v} d\boldsymbol{A}.$$

The action $S[\gamma; \boldsymbol{A}, g]$ is extremized if the derivative $dS/d\sigma = 0$. This can happen for an arbitrary field $\mathbf{t}$ only if the 1-form $\boldsymbol{V}$ vanishes for all τ. Thus we obtain the equation

$$\boldsymbol{V} = m\hat{g}\nabla_\mathbf{v} \frac{\mathbf{v}}{\sqrt{g(\mathbf{v},\mathbf{v})}} - q\iota_\mathbf{v} d\boldsymbol{A} = 0.$$

This is the equation of motion for a charged particle in an external electromagnetic field.

We may rewrite this equation in a more familiar form by choosing the parameter τ such that $g(\mathbf{v},\mathbf{v}) = 1$ and noting that the 2-form $d\boldsymbol{A}$ is the electromagnetic field tensor (2-form), usually denoted $\boldsymbol{F}$:

$$m\nabla_\mathbf{v}\mathbf{v} = q\hat{g}^{-1}\iota_\mathbf{v}\boldsymbol{F}.$$

This can be recognized as a covariant form of Newton's law. The left-hand side is the proper acceleration times mass, and the right-hand side is the Lorentz force. In the index notation, this is written as

$$mv^\alpha v^\mu{}_{;\alpha} = qg^{\mu\nu}F_{\lambda\nu}v^\lambda.$$

Note that the electromagnetic force appears in the right-hand side, while the gravitational force is not explicitly introduced (but is implicitly present because of the covariant derivative in the left-hand side). In GR, the effects of gravitation are described not by forces but by the modification of the equations of motion due to the presence of covariant derivatives.

1.9.5 Deviation of geodesics

Consider a geodesic congruence with an affine tangent vector field $\mathbf{v}$ and an affine parameter τ corresponding to the proper time, so that $g(\mathbf{v},\mathbf{v}) = 1$. Let us investigate how neighboring geodesics deviate from each other with τ. It is natural to measure the deviation between neighbor geodesics using a connecting vector field $\mathbf{c}$, since it would connect points corresponding to the same value of τ. In the reference frame of an observer whose time coordinate is τ, the neighbor geodesic thus has the coordinate $\mathbf{c}\delta$, where δ is "very small." Then the observer moving along a geodesic worldline $\gamma(\tau)$, such that $\dot\gamma = \mathbf{v}|_{p=\gamma(\tau)}$, measures the 4-acceleration of the neighbor observer as

$$\mathbf{a} = \nabla_{\dot\gamma}\nabla_{\dot\gamma}\mathbf{c}.$$

Let us compute this quantity using the properties

$$[\mathbf{v},\mathbf{c}] = \nabla_\mathbf{v}\mathbf{c} - \nabla_\mathbf{c}\mathbf{v} = 0, \quad \nabla_\mathbf{v}\mathbf{v} = 0, \quad \dot\gamma \equiv \mathbf{v}(p).$$

We find, evaluating all the quantities at the point p,

$$\mathbf{a} = \nabla_\mathbf{v}\nabla_\mathbf{v}\mathbf{c} = \nabla_\mathbf{v}\nabla_\mathbf{c}\mathbf{v} = R(\mathbf{v},\mathbf{c})\mathbf{v} + \nabla_\mathbf{c}\nabla_\mathbf{v}\mathbf{v}$$
$$= R(\mathbf{v},\mathbf{c})\mathbf{v}.$$

This is called the equation of **geodesic deviation**. In components, this equation can be written as

$$a^\mu = R_{\lambda\nu\rho}{}^\mu v^\lambda c^\nu v^\rho.$$

The equation of geodesic deviation shows that the Riemann tensor R has a direct physical manifestation: Inertial geodesics will accelerate with respect to each other (exhibiting a gravitational **tidal effect**) iff the Riemann tensor is nonzero. Furthermore, the entire Riemann tensor $R_{\lambda\mu\nu\rho}$ can be recovered if the vector-valued quantity $R(\mathbf{v},\mathbf{c})\mathbf{v}$ is known for arbitrary $\mathbf{v}$ and $\mathbf{c}$; this is derived in the following statement. Thus, all the components of the Riemann tensor at a point can be (in principle) measured using only (a finite number of) geodesic deviation experiments within an infinitesimal neighborhood of that point.

Statement 1.9.5.1: The Riemann tensor at a point can be determined from a finite number of measurements of relative acceleration of *massive* particles in *their* proper reference frame. In other words, $R(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d})$ can be expressed through its special values of the form $R(\mathbf{v},\mathbf{x},\mathbf{v},\mathbf{y})$ for some future-directed, timelike vectors $\mathbf{v}$ and some spacelike vectors $\mathbf{x},\mathbf{y}$ orthogonal to $\mathbf{v}$. (Proof on page 178.) ■

1.10 Example: hypersurface of constant curvature

Let x^μ be the usual rectangular coordinates in a *flat* space $\mathbb{R}^m$ with metric g, and let us consider the hypersurface $x^\mu x_\mu \equiv g(\mathbf{x},\mathbf{x}) = A^2$, where A is a given constant. This hypersurface will be a sphere if g is a Euclidean metric, or a hyperboloid if g is pseudo-Euclidean, but for now we assume that g has a Euclidean signature and $A^2 > 0$. We shall show that this hypersurface is an $(m-1)$-dimensional space of constant curvature; this will elucidate the concept of constant curvature. This result holds regardless of the signature of the metric and the sign of A^2, as long as $A^2 \neq 0$.

1.10.1 Tangent bundle and induced metric

The vector field $n^\mu(x) \equiv \frac{1}{A}x^\mu$ is, by construction, everywhere normal to the hypersurface and normalized to $g(\mathbf{n},\mathbf{n}) = 1$. (The normalization $g(\mathbf{n},\mathbf{n}) = 1$ holds *only* on the hypersurface, but this is sufficient for our calculations.) Vectors $\mathbf{t}$ tangent to the hypersurface satisfy $g(\mathbf{n},\mathbf{t}) = 0$; let us call such vectors simply **tangent**. The equivalent condition for tangent vectors $\mathbf{t}$ is

$$\mathbf{t} \circ (g(\mathbf{x},\mathbf{x}) - A^2) = \mathbf{t} \circ g(\mathbf{x},\mathbf{x}) = 0$$

for $\mathbf{x}$ on the hypersurface.

Statement: The commutator of tangent vector fields is again tangent.

Hint: Tangent vectors to a hypersurface $f(p) = 0$ satisfy $\mathbf{x} \circ f = 0$.

Derivation: This follows directly from the definition of the commutator: The vector field $[\mathbf{x},\mathbf{y}]$ is defined intrinsically through curves lying within the hypersurface and derivatives of functions along these curves. So $[\mathbf{x},\mathbf{y}]$ is a derivative along some curve within the hypersurface, and thus must be a tangent vector field. Here is also an explicit calculation. Suppose that the hypersurface is specified by an equation $f(p) = 0$; then a vector field $\mathbf{x}$ is tangent if $\mathbf{x} \circ f = 0$. By definition of the commutator we have (for tangent $\mathbf{x},\mathbf{y}$)

$$[\mathbf{x},\mathbf{y}] \circ f = \mathbf{x} \circ (\mathbf{y} \circ f) - \mathbf{y} \circ (\mathbf{x} \circ f) = 0.$$

■

Let us construct a self-adjoint **projector** onto the tangent bundle of the hypersurface, that is, a self-adjoint linear transformation P such that $P^2 = P$ and $P\mathbf{u}$ is everywhere tangent for any field $\mathbf{u}$. It is easy to see that the (transformation-valued) tensor field P must be defined as

$$P\mathbf{u} = \mathbf{u} - \mathbf{n}\, g(\mathbf{u},\mathbf{n}) \tag{1.82}$$

and then we have $g(P\mathbf{u},\mathbf{n}) = 0$.

The condition of **self-adjoint**ness,

$$g(P\mathbf{x},\mathbf{y}) = g(\mathbf{x},P\mathbf{y}), \tag{1.83}$$

can be motivated by the consideration that the projection of a vector $\mathbf{x}$ onto the tangent space must preserve the scalar product of $\mathbf{x}$ with vectors that are already tangent,

$$g(\mathbf{x},P\mathbf{y}) = g(P\mathbf{x},P\mathbf{y}) \quad \text{for all } \mathbf{x},\mathbf{y}.$$

This latter requirement expresses the usual geometric picture of projecting "straight down to the surface": a vector $\mathbf{x}$ is modified by subtracting a multiple of the normal vector, which preserves scalar product. By exchanging $\mathbf{x}$ with $\mathbf{y}$, we find

$$g(\mathbf{x},P\mathbf{y}) = g(P\mathbf{x},P\mathbf{y}) = g(P\mathbf{y},P\mathbf{x}) = g(\mathbf{y},P\mathbf{x}),$$

which is the condition (1.83). Conversely, Eq. (1.83) with the property $P^2 = P$ gives $g(\mathbf{x},P\mathbf{y}) = g(P\mathbf{x},P\mathbf{y})$.

The **induced metric** on the hypersurface is simply the same metric as in the larger space $\mathbb{R}^m$ but restricted to tangent vectors. For convenience of notation, we call the induced metric h and define $h(\mathbf{u},\mathbf{v}) = g(\mathbf{u},\mathbf{v})$ for *tangent* vectors $\mathbf{u},\mathbf{v}$. Thus the hypersurface has the structure of a (pseudo-) Riemannian manifold. Let us now compute the curvature tensor of this manifold.

1.10.2 Induced connection

Since the hypersurface has a metric (the induced metric h), we may define the corresponding Levi-Civita connection ∇ (also called the **induced connection**). This connection is defined *only* for vector fields tangent to the hypersurface. We note that there is a natural flat connection ∂ in $\mathbb{R}^m$, which is acts by the partial derivatives with respect to the Cartesian coordinates,

$$(\partial_{\mathbf{u}}\mathbf{v})^\alpha \equiv u^\mu \partial_\mu v^\alpha.$$

We shall now see how to express the connection ∇ on the hypersurface through the existing connection ∂ in the larger space.

The first motivation for deriving the induced connection is the following heuristic consideration. If $\mathbf{t}$ and $\mathbf{u}$ are tangent vector fields, we can compute $\partial_{\mathbf{u}}\mathbf{t}$, but the result is not necessarily a tangent vector field. To make it tangent, we can project $\partial_{\mathbf{u}}\mathbf{t}$ onto the tangent bundle using the projection operator P, i.e. we define

$$\nabla_{\mathbf{u}}\mathbf{t} \equiv P\partial_{\mathbf{u}}\mathbf{t}. \tag{1.84}$$

The result is the "induced" covariant derivative $\nabla_{\mathbf{u}}\mathbf{t}$ that has values in the tangent bundle of the hypersurface.

However, it is not immediately obvious that this "projected" connection ∇ is indeed the Levi-Civita connection that would be defined intrinsically on the hypersurface, without using the already existing connection ∂ in a larger space. Therefore we give a second, more rigorous motivation: The Levi-Civita connection on the hypersurface is defined by Eq. (1.45) if we substitute h, the induced metric, instead of g. Since $h = g$ and the Levi-Civita connection for g is ∂, we obtain for arbitrary tangent vectors $\mathbf{x},\mathbf{y},\mathbf{z}$

$$h(\nabla_{\mathbf{x}}\mathbf{y},\mathbf{z}) = g(\partial_{\mathbf{x}}\mathbf{y},\mathbf{z}). \tag{1.85}$$

However, Eq. (1.85) defines the vector $\nabla_{\mathbf{x}}\mathbf{y}$ only through its scalar product with an arbitrary tangent vector $\mathbf{z}$. The vector $\partial_{\mathbf{x}}\mathbf{y}$ is not necessarily tangent, but Eq. (1.85) shows that $\nabla_{\mathbf{x}}\mathbf{y}$ and $\partial_{\mathbf{x}}\mathbf{y}$ must have equal scalar products with any tangent vector $\mathbf{z}$ (and, of course, $\nabla_{\mathbf{x}}\mathbf{y}$ must be tangent). Therefore, $\nabla_{\mathbf{x}}\mathbf{y}$ must be the projection of $\partial_{\mathbf{x}}\mathbf{y}$ onto the tangent bundle, as shown in Eq. (1.84).

Let us note that $g(\mathbf{n},\mathbf{t}) \equiv 0$ for any tangent vector $\mathbf{t}$, and thus

$$0 = \partial_{\mathbf{u}} g(\mathbf{n},\mathbf{t}) = g(\partial_{\mathbf{u}}\mathbf{n},\mathbf{t}) + g(\mathbf{n},\partial_{\mathbf{u}}\mathbf{t})$$

on the hypersurface. Then we can rewrite the induced connection (1.84) using Eq. (1.82) as

$$\nabla_{\mathbf{u}}\mathbf{t} = \partial_{\mathbf{u}}\mathbf{t} - \mathbf{n}g\left(\mathbf{n},\partial_{\mathbf{u}}\mathbf{t}\right) = \partial_{\mathbf{u}}\mathbf{t} + \mathbf{n}g(\mathbf{t},\partial_{\mathbf{u}}\mathbf{n}) \equiv \partial_{\mathbf{u}}\mathbf{t} + \Gamma(\mathbf{u})\mathbf{t}, \tag{1.86}$$

where the tensor Γ is the following transformation-valued 1-form,

$$\Gamma(\mathbf{u})\mathbf{t} \equiv \mathbf{n}g(\mathbf{t},\partial_{\mathbf{u}}\mathbf{n}) = \Gamma^{\lambda}_{\alpha\beta}u^{\alpha}t^{\beta}; \quad \Gamma^{\lambda}_{\alpha\beta} \equiv n^{\lambda}\partial_{\alpha}n_{\beta}.$$

The tensor Γ is called the **connection 1-form** or the **Christoffel symbol**. Substituting $n^{\mu} = \frac{1}{A}x^{\mu}$, we get

$$\partial_{\alpha}n^{\beta} = \frac{1}{A}\delta^{\beta}_{\alpha}, \quad \partial_{\mathbf{u}}\mathbf{n} = \frac{1}{A}\mathbf{u}, \quad \Gamma(\mathbf{u})\mathbf{t} = \frac{1}{A}\mathbf{n}g(\mathbf{u},\mathbf{t}). \tag{1.87}$$

Statement: The induced connection ∇ is in the Levi-Civita connection on the hypersurface, i.e. for arbitrary tangent vectors $\mathbf{x},\mathbf{y},\mathbf{z}$ we have

$$\nabla_{\mathbf{x}}h(\mathbf{y},\mathbf{z}) - h(\nabla_{\mathbf{x}}\mathbf{y},\mathbf{z}) - h(\mathbf{y},\nabla_{\mathbf{x}}\mathbf{z}) = 0,$$
$$\nabla_{\mathbf{x}}\mathbf{y} - \nabla_{\mathbf{y}}\mathbf{x} - [\mathbf{x},\mathbf{y}] = 0.$$

Derivation:

$$\nabla_{\mathbf{x}}h(\mathbf{y},\mathbf{z}) \equiv \mathbf{x}\circ g(\mathbf{y},\mathbf{z});$$
$$h(\nabla_{\mathbf{x}}\mathbf{y},\mathbf{z}) = g(\partial_{\mathbf{x}}\mathbf{y},\mathbf{z});$$

then we use $\mathbf{x}\circ g(\mathbf{y},\mathbf{z}) - g(\partial_{\mathbf{x}}\mathbf{y},\mathbf{z}) - g(\mathbf{y},\partial_{\mathbf{x}}\mathbf{z}) = 0$ (the flat connection ∂ is compatible with the flat metric g). Torsion-freeness follows by projecting onto the tangent bundle:

$$\nabla_{\mathbf{x}}\mathbf{y} - \nabla_{\mathbf{y}}\mathbf{x} = P(\partial_{\mathbf{x}}\mathbf{y} - \partial_{\mathbf{y}}\mathbf{x}) = P\left[\mathbf{x},\mathbf{y}\right] = \left[\mathbf{x},\mathbf{y}\right],$$

the last equality holds since $[\mathbf{x},\mathbf{y}]$ is tangent. ■

Self-test question: In the above calculation, the connection 1-form Γ is obviously a tensor since it is defined in an invariant, geometric way by Eq. (1.87). However, Γ plays the role of the Christoffel symbol because Eq. (1.86) is written in the component notation as $\nabla_{\alpha}v^{\lambda} = \partial_{\alpha}v^{\lambda} + \Gamma^{\lambda}_{\alpha\beta}v^{\beta}$. It is explained in most GR textbooks that the Christoffel symbol $\Gamma^{\lambda}_{\alpha\beta}$ is not a tensor. Is there a contradiction?

Hint: In the index-free approach, *every* quantity is a tensor.

Answer: "Our" tensor Γ is defined using the "coordinate-based" connection ∂ which depends on a particular, *fixed* coordinate system in $\mathbb{R}^m$, which is in the present case the natural Cartesian coordinate system $\{x^{\mu}\}$. If we pass to a different coordinate system $\{y^{\mu}\}$, say to the spherical coordinate system, the components of the tensor Γ will have to be transformed according to the usual tensor law. The formula for Γ will not be the same in the coordinate system $\{y^{\mu}\}$. One could say that the connection $\partial/\partial x^{\mu}$ itself is no longer $\partial/\partial y^{\mu}$ in a different coordinate system $\{y^{\mu}\}$ and its components also need to be transformed. The standard GR textbooks, on the other hand, define the components $\Gamma^{\lambda}_{\alpha\beta}$ by a *fixed*, non-covariant formula in an *arbitrary* coordinate system. This procedure does not define a tensor. So then it is no surprise that the components $\Gamma^{\lambda}_{\alpha\beta}$ do not transform as components of a tensor. ■

Practice problem: A transformation-valued tensor field S is defined in the tangent bundle of the hypersurface by its action on tangent vector fields $\mathbf{t}$ in the following way,

$$S\mathbf{t} = h(\mathbf{v},\mathbf{t})\mathbf{v} - \frac{1}{17}\mathbf{t},$$

where h is the induced metric, $\mathbf{v}$ is a given tangent vector field which is normalized, $h(\mathbf{v},\mathbf{v}) = 1$. Compute ∇S, which should be a transformation-valued 1-form.

Solution: For arbitrary tangent vector fields $\mathbf{u}$ and $\mathbf{t}$, we have

$$\begin{aligned}(\nabla_{\mathbf{u}}S)\mathbf{t} &= \nabla_{\mathbf{u}}(S\mathbf{t}) - S\nabla_{\mathbf{u}}\mathbf{t}\\ &= \nabla_{\mathbf{u}}\left(h(\mathbf{v},\mathbf{t})\mathbf{v} - \frac{1}{17}\mathbf{t}\right) - h(\mathbf{v},\nabla_{\mathbf{u}}\mathbf{t})\mathbf{v} + \frac{1}{17}\nabla_{\mathbf{u}}\mathbf{t}\\ &= h(\nabla_{\mathbf{u}}\mathbf{v},\mathbf{t})\mathbf{v} + h(\mathbf{v},\mathbf{t})\nabla_{\mathbf{u}}\mathbf{v}\\ &= g(\partial_{\mathbf{u}}\mathbf{v},\mathbf{t})\mathbf{v} + g(\mathbf{v},\mathbf{t})\left(\partial_{\mathbf{u}}\mathbf{v} - \frac{1}{A}\mathbf{n}g(\mathbf{u},\mathbf{v})\right).\end{aligned}$$

■

1.10.3 Riemann tensor within the hypersurface

Finally, let us compute the Riemann tensor of the hypersurface. By definition, we have

$$R(\mathbf{x},\mathbf{y})\mathbf{z} = \left[\nabla_{\mathbf{x}},\nabla_{\mathbf{y}}\right]\mathbf{z} - \nabla_{[\mathbf{x},\mathbf{y}]}\mathbf{z},$$

and we write out the terms using Eq. (1.86),

$$\begin{aligned}\nabla_{\mathbf{x}}\nabla_{\mathbf{y}}\mathbf{z} &= \nabla_{\mathbf{x}}\left(\partial_{\mathbf{y}}\mathbf{z} + \Gamma(\mathbf{y})\mathbf{z}\right)\\ &= \partial_{\mathbf{x}}\partial_{\mathbf{y}}\mathbf{z} + \Gamma(\mathbf{x})\partial_{\mathbf{y}}\mathbf{z} + \partial_{\mathbf{x}}\Gamma(\mathbf{y})\mathbf{z} + \Gamma(\mathbf{x})\Gamma(\mathbf{y})\mathbf{z};\\ \nabla_{[\mathbf{x},\mathbf{y}]}\mathbf{z} &= \partial_{[\mathbf{x},\mathbf{y}]}\mathbf{z} + \Gamma([\mathbf{x},\mathbf{y}])\mathbf{z};\\ \left[\nabla_{\mathbf{x}},\nabla_{\mathbf{y}}\right]\mathbf{z} - \nabla_{[\mathbf{x},\mathbf{y}]}\mathbf{z} &= \underline{\left[\partial_{\mathbf{x}},\partial_{\mathbf{y}}\right]\mathbf{z} - \partial_{[\mathbf{x},\mathbf{y}]}\mathbf{z}} + \left[\Gamma(\mathbf{x}),\Gamma(\mathbf{y})\right]\mathbf{z}\\ &\quad + \underline{\Gamma(\partial_{\mathbf{x}}\mathbf{y} - \partial_{\mathbf{y}}\mathbf{x} - [\mathbf{x},\mathbf{y}])\mathbf{z}} + (\partial_{\mathbf{x}}\Gamma)(\mathbf{y})\mathbf{z} - (\partial_{\mathbf{y}}\Gamma)(\mathbf{x})\mathbf{z}\\ &= \left[\Gamma(\mathbf{x}),\Gamma(\mathbf{y})\right]\mathbf{z} + (\partial_{\mathbf{x}}\Gamma)(\mathbf{y})\mathbf{z} - (\partial_{\mathbf{y}}\Gamma)(\mathbf{x})\mathbf{z}.\end{aligned} \tag{1.88}$$

In the last line we used the fact that ∂ is a flat and torsion-free connection, in order to cancel the underlined terms. (Note the similarity of the last formula to the standard expression for the Riemann tensor in the component notation, $R^{\cdot}_{\cdot\cdot\cdot\cdot} = \partial\Gamma^{\cdot}_{\cdot\cdot} - \partial\Gamma^{\cdot}_{\cdot\cdot} + \Gamma^{\cdot}_{\cdot\cdot}\Gamma^{\cdot}_{\cdot\cdot} - \Gamma^{\cdot}_{\cdot\cdot}\Gamma^{\cdot}_{\cdot\cdot}$; see Eq. (A.11) in Appendix A.)

Now we substitute Eq. (1.87) into Eq. (1.88) and find (for tangent vectors $\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{t}$)

$$\begin{aligned}\Gamma(\mathbf{x})\Gamma(\mathbf{y})\mathbf{z} &= \frac{1}{A}\mathbf{n}g(\mathbf{x},\Gamma(\mathbf{y})\mathbf{z}) = \frac{1}{A}\mathbf{n}g(\mathbf{x},\mathbf{n}(...)) = 0,\\ (\partial_{\mathbf{x}}\Gamma)(\mathbf{y})\mathbf{z} &= \frac{1}{A}(\partial_{\mathbf{x}}\mathbf{n})g(\mathbf{y},\mathbf{z}) = \frac{1}{A^2}\mathbf{x}g(\mathbf{y},\mathbf{z}),\\ R(\mathbf{x},\mathbf{y})\mathbf{z} &= \frac{1}{A^2}\left(\mathbf{x}g(\mathbf{y},\mathbf{z}) - \mathbf{y}g(\mathbf{x},\mathbf{z})\right),\\ R(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{t}) &= \frac{1}{A^2}\left(g(\mathbf{x},\mathbf{t})g(\mathbf{y},\mathbf{z}) - g(\mathbf{x},\mathbf{z})g(\mathbf{y},\mathbf{t})\right)\\ &= \frac{1}{A^2}\left(h(\mathbf{x},\mathbf{t})h(\mathbf{y},\mathbf{z}) - h(\mathbf{x},\mathbf{z})h(\mathbf{y},\mathbf{t})\right),\end{aligned} \tag{1.89}$$

since $g(\mathbf{u},\mathbf{v}) = h(\mathbf{u},\mathbf{v})$ for tangent vectors $\mathbf{u},\mathbf{v}$. In the index notation, the above Riemann tensor is

$$R_{\alpha\beta\gamma\delta} = \frac{1}{A^2}\left(h_{\alpha\delta}h_{\beta\gamma} - h_{\alpha\gamma}h_{\beta\delta}\right).$$

A manifold with a Riemann tensor of form (1.89) is called a space of **constant curvature**. Note also that the second

Bianchi identity forces $A = \text{const}$ for a Riemann tensor of the form (1.89).

Our result is that the hypersurface $x_\mu x^\mu = A^2$ has constant curvature. The actual magnitude of the curvature is A^{-2} if the metric g is Euclidean and $-A^{-2}$ if it is pseudo-Euclidean with signature $(+--...)$, because in the latter case the vector $\mathbf{n}$ is "timelike" and the metric h is negative-definite. (The cases $x_\mu x^\mu = -A^2$ or $g(\mathbf{n}, \mathbf{n}) = -1$ can be treated in a very similar way and the result is essentially the same, with appropriate sign changes.)

Remark: The following two arguments help to visualize the concept of constant curvature.

The first argument appeals to the intuitive understanding that a sphere is a perfectly symmetrically curved surface that has a constant curvature at all points. The above calculation of the Riemann tensor obviously holds for a sphere $|\mathbf{x}|^2 = R^2$ in the usual Euclidean space. It makes intuitive sense to say that a sphere of radius R has a constant curvature equal to $1/R$. Therefore, any space with a Riemann tensor (1.89) is a space of constant curvature, and the value of the curvature is equal to $1/A$.

The second argument is somewhat more abstract. The tensor $R_{\alpha\beta\lambda\mu} = R_{\lambda\mu\alpha\beta}$ can be viewed as a symmetric bilinear form in the space of 2-forms, $R(\omega^{(1)}, \omega^{(2)}) = R^{\alpha\beta\gamma\delta}\omega^{(1)}_{\alpha\beta}\omega^{(2)}_{\gamma\delta}$, or as a linear transformation in the space of 2-forms, $\omega_{\alpha\beta} \to R_{\alpha\beta}{}^{\lambda\mu}\omega_{\lambda\mu}$, where $\omega_{\alpha\beta} = -\omega_{\beta\alpha}$ is an arbitrary 2-form. The space of 2-forms is six-dimensional, and the bilinear form $R(\omega^{(1)}, \omega^{(2)})$ can be represented as a symmetric 6×6 matrix. Due to this symmetry, the corresponding transformation is diagonalizable. Consider the six eigenvalues $\lambda_1(p)$, ..., $\lambda_6(p)$ of that transformation as functions of the point p on the manifold. These six scalar functions characterize (in some way) the magnitude of the curvature of the manifold at various points. Without a detailed analysis of the significance of the eigenvalues λ_j and the corresponding eigenvectors, we may heuristically expect that, for a manifold of *constant* curvature, all the six eigenvalues λ_j should be equal to each other and remain constant throughout the manifold. Thus the transformation $R_{\alpha\beta}{}^{\lambda\mu}$ must be proportional to the identity map in the space of 2-forms,

$$R_{\alpha\beta}{}^{\lambda\mu} = \left(\delta^\mu_\alpha \delta^\lambda_\beta - \delta^\lambda_\alpha \delta^\mu_\beta\right) K, \quad K = \text{const}.$$

(The coefficient K must be constant also by the second Bianchi identity, once we assume that $R_{\alpha\beta}{}^{\lambda\mu}$ has the form given above.) Then the comparison with the sphere shows that the numeric value of the curvature is K. ■

Practice problem: Consider a surface embedded in $\mathbb{R}^m$, specified by an equation $f(p) = 0$, where f is a given scalar function and $p \in \mathbb{R}^m$. Obtain the normal vector field $\mathbf{n}$ in terms of f and the flat metric g in $\mathbb{R}^m$. Compute the induced metric h and the Riemann tensor $R(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{t})$ in terms of $\mathbf{n}$ and g.

Solution: The vector $\mathbf{n}$ normal to the hypersurface is

$$\mathbf{n} = \frac{1}{N}\hat{g}^{-1}(\mathrm{d}f), \quad N \equiv \sqrt{g^{-1}(\mathrm{d}f, \mathrm{d}f)}.$$

In other words, $g(\mathbf{n}, \mathbf{x}) = N^{-1}\partial_{\mathbf{x}} f$ for any vector $\mathbf{x}$. In components:

$$n^\alpha = \frac{g^{\alpha\beta}\partial_\beta f}{N}, \quad N \equiv \sqrt{g^{\alpha\beta}\left(\partial_\alpha f\right)\left(\partial_\beta f\right)}.$$

Note that $g(\mathbf{n}, \mathbf{n}) = 1$ and so $g(\partial_{\mathbf{x}}\mathbf{n}, \mathbf{n}) = 0$ for any vector $\mathbf{x}$. The projection onto the tangent bundle is

$$P\mathbf{x} = \mathbf{x} - \mathbf{n}g(\mathbf{n}, \mathbf{x}).$$

The induced connection ∇ is equal to ∂ projected onto the tangent bundle,

$$\nabla_{\mathbf{x}}\mathbf{y} = P\partial_{\mathbf{x}}\mathbf{y} = \partial_{\mathbf{x}}\mathbf{y} - \mathbf{n}g(\mathbf{n}, \partial_{\mathbf{x}}\mathbf{y}) = \partial_{\mathbf{x}}\mathbf{y} + \mathbf{n}g(\mathbf{y}, \partial_{\mathbf{x}}\mathbf{n}),$$

which can be also written as

$$\nabla_{\mathbf{x}}\mathbf{y} = \partial_{\mathbf{x}}\mathbf{y} + \Gamma(\mathbf{x})\mathbf{y}, \quad \Gamma(\mathbf{x})\mathbf{y} \equiv \mathbf{n}g(\partial_{\mathbf{x}}\mathbf{n}, \mathbf{y}).$$

The connection ∇ is the Levi-Civita connection on the hypersurface. The curvature of ∇ is

$$\begin{aligned} R(\mathbf{x}, \mathbf{y})\mathbf{z} &= [\Gamma(\mathbf{x}), \Gamma(\mathbf{y})]\,\mathbf{z} + (\partial_{\mathbf{x}}\Gamma)(\mathbf{y})\mathbf{z} - (\partial_{\mathbf{y}}\Gamma)(\mathbf{x})\mathbf{z} \\ &= g(\partial_{\mathbf{y}}\mathbf{n}, \mathbf{z})\partial_{\mathbf{x}}\mathbf{n} - g(\partial_{\mathbf{x}}\mathbf{n}, \mathbf{z})\partial_{\mathbf{y}}\mathbf{n}; \\ R(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{t}) &= g(\partial_{\mathbf{y}}\mathbf{n}, \mathbf{z})g(\partial_{\mathbf{x}}\mathbf{n}, \mathbf{t}) - g(\partial_{\mathbf{x}}\mathbf{n}, \mathbf{z})g(\partial_{\mathbf{y}}\mathbf{n}, \mathbf{t}). \end{aligned}$$

In the index notation, we can write

$$\begin{aligned} R_{\kappa\lambda\mu\nu} &= \left((\partial_\lambda n^\alpha)(\partial_\kappa n^\beta) - (\partial_\kappa n^\alpha)(\partial_\lambda n^\beta)\right) g_{\alpha\mu}g_{\beta\nu} \\ &= \left(\partial_\lambda n_\mu\right)\left(\partial_\kappa n_\nu\right) - \left(\partial_\kappa n_\mu\right)\left(\partial_\lambda n_\nu\right), \end{aligned}$$

since the metric $g_{\alpha\beta}$ in the Euclidean space $\mathbb{R}^m$ is flat. It is clear that, in general, the hypersurface $f(p) = 0$ is not a space of constant curvature. ■

2 Geometry of null surfaces

Throughout this chapter, we consider a four-dimensional spacetime with a metric g having the signature $(+---)$, and the Levi-Civita connection ∇. Most results can be straightforwardly generalized to higher dimensions, but it is important that the metric has indefinite signature of the kind $(+--...)$ and, as a consequence, that there exist null directions.

2.1 Null vectors

Recall that a vector $\mathbf{v}$ is called **timelike** if $g(\mathbf{v},\mathbf{v}) > 0$, **null** if $g(\mathbf{v},\mathbf{v}) = 0$, and **spacelike** if $g(\mathbf{v},\mathbf{v}) < 0$. A geodesic curve $\gamma(\tau)$ is called a **lightray** if $g(\dot\gamma,\dot\gamma) = 0$ for all τ. (Lightrays are worldlines of light or other radiation carried by massless particles.)

Null vectors have somewhat peculiar properties which we now review.

Statement 2.1.0.1: **a)** If $\mathbf{n}$ is null and $\mathbf{v}$ is orthogonal to $\mathbf{n}$, i.e. $g(\mathbf{v},\mathbf{n}) = 0$, then $\mathbf{v}$ is either parallel to $\mathbf{n}$, or spacelike. In other words: the orthogonal complement to a null vector $\mathbf{n}$ is spanned by $\mathbf{n}$ and two spacelike vectors. (The **orthogonal complement** $\mathbf{n}^\perp$ of a vector $\mathbf{n}$ is the subspace consisting of vectors $\mathbf{v}$ such that $g(\mathbf{n},\mathbf{v}) = 0$.)
b) If $\mathbf{n}$ is null and $\mathbf{v}$ is some vector such that $g(\mathbf{v},\mathbf{n}) \neq 0$ then there exists a linear combination $\alpha\mathbf{n} + \beta\mathbf{v}$ that is timelike, and another that is spacelike.

Hint: These statements concern vectors in a single tangent space $T_p\mathcal{M}$, so it suffices to consider the Minkowski space with a standard coordinate basis $\{t,x,y,z\}$.

2.1.1 Orthogonal complement spaces

The preceding statement 2.1.0.1 shows that the orthogonal complement $\mathbf{n}^\perp$ to a null vector $\mathbf{n}$ is a three-dimensional subspace spanned by $\mathbf{n}$ and two other spacelike vectors. For calculations, it is convenient to have a projection operator onto this subspace. We shall now investigate the properties of such projections.

We begin with an easier case: the projection onto $\mathbf{n}^\perp$, where $\mathbf{n}$ is *not* null. Recall that a self-adjoint projector is an operator P such that $g(P\mathbf{x},\mathbf{n}) = 0$ for all $\mathbf{x}$ (any vector $\mathbf{x}$ is projected onto $\mathbf{n}^\perp$), $P^2 = P$ ("the projection remains projected"), and the bilinear form corresponding to P is symmetric, $g(P\mathbf{x},\mathbf{y}) = g(\mathbf{x},P\mathbf{y})$ for all vectors $\mathbf{x},\mathbf{y}$ (self-adjointness; see Sec. 1.10.1 for a motivation). For a unit vector $\mathbf{n}$, there is a unique self-adjoint projector onto $\mathbf{n}^\perp$, given by the standard formula

$$P\mathbf{x} = \mathbf{x} - \mathbf{n}g(\mathbf{n},\mathbf{x}). \tag{2.1}$$

(See also Eqs. (1.82) and (1.83) in Sec. 1.10.1.) It is easy to compute the trace of P (see Sec. 1.7.3 for details),

$$g(P\mathbf{x},\mathbf{y}) = g(\mathbf{x},\mathbf{y}) - g(\mathbf{n},\mathbf{x})g(\mathbf{n},\mathbf{y});$$
$$\operatorname{Tr} P = \operatorname{Tr}_{(\mathbf{x},\mathbf{y})} g(P\mathbf{x},\mathbf{y}) = 4 - g(\mathbf{n},\mathbf{n}) = 4 - 1 = 3.$$

For calculations in a subspace, it is convenient to define a **partial metric** h which coincides with the metric g for vectors in the subspace but is zero on vectors orthogonal to the subspace. Thus, h is defined in the entire space but "can only see the subspace." The advantage of introducing the partial metric is that one can work with vectors from the full space, substituting vectors from the subspace only at the end of calculations.

Given a self-adjoint projector P onto the subspace, the partial metric h satisfying the above requirements can be expressed as

$$h(\mathbf{a},\mathbf{b}) = g(P\mathbf{a},P\mathbf{b}) = g(P\mathbf{a},\mathbf{b}).$$

(The last equality is due to $P^2 = P$ and self-adjointness.) For a standard projector (2.1) to the space $\mathbf{n}^\perp$ where $g(\mathbf{n},\mathbf{n}) = 1$, the partial metric is

$$h(\mathbf{a},\mathbf{b}) = g(\mathbf{a},\mathbf{b}) - g(\mathbf{n},\mathbf{a})g(\mathbf{n},\mathbf{b}).$$

Now we turn to the case when the vector $\mathbf{n}$ is null. Then the formula (2.1) fails since $g(\mathbf{n},\mathbf{n}) = 0$, so $g(P\mathbf{u},\mathbf{n}) \neq 0$. Clearly, the transformation $P\mathbf{x}$ needs to subtract from $\mathbf{x}$ some vector whose scalar product with $\mathbf{n}$ is nonzero. So let us try $P\mathbf{x} = \mathbf{x} - \mathbf{l}g(\mathbf{n},\mathbf{x})$, where $\mathbf{l}$ is a vector such as $g(\mathbf{l},\mathbf{n}) = 1$. This gives $g(P\mathbf{x},\mathbf{n}) = 0$ for any $\mathbf{x}$; however, this operator P is not self-adjoint, $P^T\mathbf{x} = \mathbf{x} - \mathbf{n}g(\mathbf{l},\mathbf{x}) \neq P\mathbf{x}$, and we need to symmetrize it. So finally we define the map

$$P\mathbf{x} = \mathbf{x} - \mathbf{l}g(\mathbf{n},\mathbf{x}) - \mathbf{n}g(\mathbf{l},\mathbf{x}), \tag{2.2}$$

where again $g(\mathbf{l},\mathbf{n}) = 1$. It is easy to check that P is now self-adjoint and that $g(P\mathbf{x},\mathbf{n}) = 0$ for any $\mathbf{x}$. The requirement $P^2 = P$ yields the condition

$$0 = g(\mathbf{l},P\mathbf{x}) = g(\mathbf{l},\mathbf{x} - \mathbf{l}g(\mathbf{n},\mathbf{x}) - \mathbf{n}g(\mathbf{l},\mathbf{x})) = g(\mathbf{l},\mathbf{l})g(\mathbf{n},\mathbf{x}),$$

which can be satisfied for all $\mathbf{x}$ only if $\mathbf{l}$ itself is null, $g(\mathbf{l},\mathbf{l}) = 0$. Thus the self-adjoint projector (2.2) is, *by necessity*, also a projector onto the orthogonal complement of another null vector $\mathbf{l}$. The image of P is therefore the set $(\mathbf{l},\mathbf{n})^\perp$ of vectors orthogonal to both $\mathbf{n}$ and $\mathbf{l}$, which is not the entire $\mathbf{n}^\perp$ but a *two-dimensional* subspace within $\mathbf{n}^\perp$. Accordingly, the trace of P is equal to 2:

$$\operatorname{Tr} P = 4 - g(\mathbf{n},\mathbf{l}) - g(\mathbf{n},\mathbf{l}) = 4 - 1 - 1 = 2.$$

Statement: For null vectors $\mathbf{n}$, there exists no self-adjoint projector *onto* the full subspace $\mathbf{n}^\perp$.

Derivation: If P were such a projector, we would have $P\mathbf{n} = \mathbf{n}$. Consider a vector $\mathbf{v}$ such that $g(\mathbf{n},\mathbf{v}) \neq 0$. By assumption the vector $P\mathbf{v}$ must be orthogonal to $\mathbf{n}$, and then the condition of self-adjointness yields a contradiction:

$$0 \neq g(\mathbf{n},\mathbf{v}) = g(P\mathbf{n},\mathbf{v}) = g(\mathbf{n},P\mathbf{v}) = 0.$$

In other words, we cannot project $\mathbf{v}$ while keeping the scalar product of $\mathbf{v}$ and $\mathbf{n}$ constant.

The orthogonal complement subspace $\mathbf{n}^\perp$ has a naturally induced metric, which is just the restriction of the full metric g onto $\mathbf{n}^\perp$. However, the induced metric is *degenerate* since $\mathbf{n}$ is orthogonal to every vector in the subspace. It is awkward

to work with a degenerate metric; for instance, tensors specified only through their scalar products with vectors from $\mathbf{n}^\perp$ become ambiguous (defined up to multiples of $\mathbf{n}$). To avoid this inconvenience, we may further restrict the subspace $\mathbf{n}^\perp$ to a two-dimensional subspace $(\mathbf{l},\mathbf{n})^\perp$ which is the image of a projector P. The induced metric is nondegenerate on the subspace $(\mathbf{l},\mathbf{n})^\perp$. However, this subspace is not uniquely selected since it depends on the choice of a second null vector, $\mathbf{l}$. Despite these difficulties, we will obtain unambiguous results that are independent of the freedom in the choice of $\mathbf{l}$.

Let us now compare the projectors for the spaces $\mathbf{v}^\perp$, $\mathbf{n}^\perp$, and $\mathbf{s}^\perp$, where the vectors $\mathbf{v},\mathbf{n},\mathbf{s}$ are respectively timelike, null, and spacelike:

$$P_\mathbf{v}\mathbf{x} = \mathbf{x} - g(\mathbf{v},\mathbf{x})\mathbf{v}, \quad g(\mathbf{v},\mathbf{v}) = 1, \quad \mathrm{Tr}\, P_\mathbf{v} = 3;$$
$$P_\mathbf{n}\mathbf{x} = \mathbf{x} - \mathbf{l}g(\mathbf{n},\mathbf{x}) - \mathbf{n}g(\mathbf{l},\mathbf{x}), \quad g(\mathbf{n},\mathbf{l}) = 1, \quad \mathrm{Tr}\, P_\mathbf{n} = 2;$$
$$P_\mathbf{s}\mathbf{x} = x + g(\mathbf{s},\mathbf{x})\mathbf{s}, \quad g(\mathbf{s},\mathbf{s}) = -1, \quad \mathrm{Tr}\, P_\mathbf{s} = 3.$$

We find that the timelike and the spacelike cases are quite similar but differ from the null case. The partial metrics are

$$h_\mathbf{v}(\mathbf{a},\mathbf{b}) = g(\mathbf{a},\mathbf{b}) - g(\mathbf{v},\mathbf{a})g(\mathbf{v},\mathbf{b}); \tag{2.3}$$
$$h_\mathbf{n}(\mathbf{a},\mathbf{b}) = g(\mathbf{a},\mathbf{b}) - g(\mathbf{n},\mathbf{a})g(\mathbf{l},\mathbf{b}) - g(\mathbf{l},\mathbf{a})g(\mathbf{n},\mathbf{b}); \tag{2.4}$$
$$h_\mathbf{s}(\mathbf{a},\mathbf{b}) = g(\mathbf{a},\mathbf{b}) + g(\mathbf{s},\mathbf{a})g(\mathbf{s},\mathbf{b}).$$

We can also write the **partial inverse metrics**, which are (2,0)-tensors obtained from the above partial metrics by using the map $\hat{g}$, according to

$$h^{-1}(\omega_1,\omega_2) \equiv h(\hat{g}\omega_1, \hat{g}\omega_2),$$

where ω_1 and ω_2 are arbitrary 1-forms. We find

$$h_\mathbf{v}^{-1} = g^{-1} - \mathbf{v}\otimes\mathbf{v},$$
$$h_\mathbf{n}^{-1} = g^{-1} - \mathbf{n}\otimes\mathbf{l} - \mathbf{l}\otimes\mathbf{n},$$
$$h_\mathbf{s}^{-1} = g^{-1} + \mathbf{s}\otimes\mathbf{s}.$$

Note that the partial metrics are not invertible, so we do not talk of "inverse partial metrics."

It is important to investigate the signature the partial metrics on the subspaces where they are nondegenerate (the **partial signature**). The easiest way to deduce the signature is to choose a suitable basis containing the vectors $\mathbf{v}$, $\mathbf{n}$, or $\mathbf{s}$. For the metric $h_\mathbf{v}$ we choose an orthogonal basis $\{\mathbf{v},\mathbf{s}_1,\mathbf{s}_2,\mathbf{s}_3\}$, where $\mathbf{s}_1,\mathbf{s}_2,\mathbf{s}_3$ are all spacelike. Then we find $h_\mathbf{v}(\mathbf{s}_j,\mathbf{s}_k) = -\delta_{jk}$, so the partial signature of $h_\mathbf{v}$ is $(---)$. For the metric $h_\mathbf{n}$, we choose a basis containing the vectors $\{\mathbf{l},\mathbf{n},\mathbf{s}_1,\mathbf{s}_2\}$, where $\mathbf{s}_1,\mathbf{s}_2$ are spacelike and orthogonal to each other and to $\mathbf{l},\mathbf{n}$. Then we find $h_\mathbf{n}(...,\mathbf{l}) = h_\mathbf{n}(...,\mathbf{n}) = 0$, while $h_\mathbf{n}(\mathbf{s}_j,\mathbf{s}_k) = -\delta_{jk}$, and therefore the partial signature of $h_\mathbf{s}$ is $(--)$. Finally, for the metric $h_\mathbf{s}$ we choose an orthogonal basis $\{\mathbf{s},\mathbf{v}_1,\mathbf{s}_2,\mathbf{s}_3\}$, where $\mathbf{v}_1$ is a timelike vector and $\mathbf{s}_2,\mathbf{s}_3$ are spacelike. Thus the partial signature of $h_\mathbf{s}$ is $(+--)$.

2.1.2 Divergence of a null vector field

We have seen that there are *two* spacelike connecting vectors $\mathbf{s}_1,\mathbf{s}_2$ orthogonal to $\mathbf{n}$ within the subspace $\mathbf{n}^\perp$. Thus we can visualize how the 2-volume (i.e. area), $A(\mathbf{s}_1,\mathbf{s}_2)$, of the parallelogram spanned by these two vectors propagates along the null direction $\mathbf{n}$. From elementary geometric considerations, this area can be expressed as

$$A(\mathbf{s}_1,\mathbf{s}_2) = \left|g(\mathbf{s}_1,\mathbf{s}_1)g(\mathbf{s}_2,\mathbf{s}_2) - g(\mathbf{s}_1,\mathbf{s}_2)g(\mathbf{s}_2,\mathbf{s}_1)\right|^{1/2}.$$

The change of the area along the orbits of $\mathbf{n}$ is described by the directional derivative $\nabla_\mathbf{n} A(\mathbf{s}_1,\mathbf{s}_2)$.

It is easier to compute $\nabla_\mathbf{n} A(\mathbf{s}_1,\mathbf{s}_2)$ if we note that $A(\mathbf{s}_1,\mathbf{s}_2)$ is equal to the 4-volume spanned by $\mathbf{s}_1,\mathbf{s}_2$ and two unit vectors $\mathbf{x},\mathbf{y}$ orthogonal to both $\mathbf{s}_1$ and $\mathbf{s}_2$. For instance, we may choose $\mathbf{x}$ to be spacelike and $\mathbf{y}$ to be timelike, $g(\mathbf{y},\mathbf{y}) = 1 = -g(\mathbf{x},\mathbf{x})$. The 2-volume $A(\mathbf{s}_1,\mathbf{s}_2)$ is then interpreted as the area measured in the reference frame of an observer moving along the timelike direction $\mathbf{y}$. (Of course, the area $A(\mathbf{s}_1,\mathbf{s}_2)$ is observer-independent since it is defined regardless of the choice of the auxiliary vectors $\mathbf{x}$, $\mathbf{y}$.) The 4-volume spanned by $\mathbf{s}_1,\mathbf{s}_2,\mathbf{x},\mathbf{y}$ is expressed through the Levi-Civita symbol as

$$A(\mathbf{s}_1,\mathbf{s}_2) = \varepsilon(\mathbf{s}_1,\mathbf{s}_2,\mathbf{x},\mathbf{y}). \tag{2.5}$$

Since ε is totally antisymmetric, the expression $\varepsilon(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d})$ can be interpreted as a linear function of the totally antisymmetric (4,0)-tensor $\mathbf{a}\wedge\mathbf{b}\wedge\mathbf{c}\wedge\mathbf{d}$, where $\mathbf{x}\wedge\mathbf{y}$ is the exterior (antisymmetric) product of vectors. Thus, the vectors $\mathbf{x}$ and $\mathbf{y}$ in Eq. (2.5) can be replaced by their linear combinations $\mathbf{x}'$ and $\mathbf{y}'$ as long as $\mathbf{x}\wedge\mathbf{y} = \mathbf{x}'\wedge\mathbf{y}'$. It turns out to be convenient for the calculation of $\nabla_\mathbf{n} A$ to choose the vectors $\mathbf{x}',\mathbf{y}'$ as $\mathbf{x}' = \mathbf{l}$ and $\mathbf{y}' = \mathbf{n}$, where $\mathbf{l}$ is the null vector orthogonal to both $\mathbf{s}_1$ and $\mathbf{s}_2$ and related to $\mathbf{n}$ by $g(\mathbf{l},\mathbf{n}) = 1$.

Statement: Indeed the choice $\mathbf{x}' = \mathbf{l}$, $\mathbf{y}' = \mathbf{n}$ is possible, i.e. $\mathbf{x}\wedge\mathbf{y} = \mathbf{l}\wedge\mathbf{n}$.

Derivation: Write the vectors $\mathbf{x}' = \mathbf{l}, \mathbf{y}' = \mathbf{n}$ as linear combinations of $\mathbf{x},\mathbf{y}$ with unknown coefficients:

$$\mathbf{l} = a_1\mathbf{x} + b_1\mathbf{y}, \quad \mathbf{n} = a_2\mathbf{x} + b_2\mathbf{y}.$$

Since $g(\mathbf{x},\mathbf{y}) = 0$, the conditions $g(\mathbf{l},\mathbf{l}) = g(\mathbf{n},\mathbf{n}) = 0$ give

$$a_1^2 = b_1^2, \quad a_2^2 = b_2^2,$$

so we could choose $a_1 = b_1, a_2 = -b_2$ (the signs must differ, or else $\mathbf{l}$ will come out parallel to $\mathbf{n}$). Then we have $1 = g(\mathbf{l},\mathbf{n}) = -2a_1a_2$ and

$$\mathbf{l}\wedge\mathbf{n} = -2a_1a_2\mathbf{x}\wedge\mathbf{y} = \mathbf{x}\wedge\mathbf{y}.$$

Thus the area is expressed as

$$A = \varepsilon(\mathbf{s}_1,\mathbf{s}_2,\mathbf{l},\mathbf{n}).$$

Since $\mathbf{s}_{1,2}$ are connecting vectors, the derivative of the area A along the lines of $\mathbf{n}$ is most conveniently computed by using the Lie derivative,

$$\frac{dA}{d\tau} = \mathcal{L}_\mathbf{n} A = \mathcal{L}_\mathbf{n}\varepsilon(\mathbf{s}_1,\mathbf{s}_2,\mathbf{l},\mathbf{n})$$
$$= (\mathrm{div}\mathbf{n})\varepsilon(\mathbf{s}_1,\mathbf{s}_2,\mathbf{l},\mathbf{n}) + \varepsilon(\mathbf{s}_1,\mathbf{s}_2,\mathcal{L}_\mathbf{n}\mathbf{l},\mathbf{n}).$$

We need to retain the last term above because $\mathbf{l}$ is *not necessarily* a connecting vector for $\mathbf{n}$. Let us compute the scalar product of $\mathcal{L}_\mathbf{n}\mathbf{l} \equiv [\mathbf{n},\mathbf{l}]$ with $\mathbf{n}$,

$$g(\mathbf{n},[\mathbf{n},\mathbf{l}]) = g(\mathbf{n},\nabla_\mathbf{n}\mathbf{l}) - g(\mathbf{n},\nabla_\mathbf{l}\mathbf{n})$$
$$= \nabla_\mathbf{n} g(\mathbf{n},\mathbf{l}) - g(\nabla_\mathbf{n}\mathbf{n},\mathbf{l}) - \frac{1}{2}\nabla_\mathbf{l} g(\mathbf{n},\mathbf{n}) = 0.$$

Therefore $[\mathbf{l},\mathbf{n}]$ is a linear combination of $\mathbf{n}$, $\mathbf{s}_1$, and $\mathbf{s}_2$. Hence, $\varepsilon(\mathbf{s}_1,\mathbf{s}_2,\mathcal{L}_\mathbf{n}\mathbf{l},\mathbf{n}) = 0$ regardless of the choice of $\mathbf{l}$. (This is to be expected since $A(\mathbf{s}_1,\mathbf{s}_2)$ and $\nabla_\mathbf{n} A(\mathbf{s}_1,\mathbf{s}_2)$ are defined independently of the choice of $\mathbf{l}$.) Finally, we find

$$\frac{dA}{d\tau} = (\mathrm{div}\mathbf{n})A. \tag{2.6}$$

The divergence $\mathrm{div}\mathbf{n}$ is thus interpreted as the relative rate of change in the area spanned by the vectors $\mathbf{s}_1,\mathbf{s}_2$ as this area is being transported by $\mathbf{n}$.

2.2 Null surfaces

Null surfaces (3-dimensional hypersurfaces orthogonal to a null vector field) play a major role in relativity. Before considering null surfaces, we shall spend some time studying 3-dimensional hypersurfaces is general.

2.2.1 Three-dimensional hypersurfaces

A convenient way to define a 3-dimensional hypersurface is through an equation $f(p) = 0$, where $f(p)$ is an auxiliary function. In this way the hypersurface can be visualized as the locus of constant f. A vector $\mathbf{t}$ is tangent to the hypersurface of constant f if the function f remains constant in the direction of $\mathbf{t}$, i.e. if $\mathbf{t} \circ f = 0$. Since $\mathbf{x} \circ f$ is a linear function of $\mathbf{x}$, there exists a vector $\mathbf{n}$ such that

$$\mathbf{x} \circ f \equiv (\mathrm{d}f) \circ \mathbf{x} = g(\mathbf{n}, \mathbf{x}), \quad \mathbf{n} = \hat{g}^{-1}(\mathrm{d}f), \tag{2.7}$$

where $\hat{g}^{-1}$ is the map from 1-forms into vectors. The vector $\mathbf{n}$ is called the **contravariant gradient** of f and is the normal vector to the hypersurface because it is orthogonal to every tangent vector $\mathbf{t}$,

$$g(\mathbf{n}, \mathbf{t}) = \mathbf{t} \circ f = 0.$$

Note that the vector field $\mathbf{n}$ is well-defined not only on the hypersurface but in the entire space; away from the hypersurface $f(p) = 0$, the vector $\mathbf{n}$ is the normal vector to hypersurfaces $f(p) = f_0$ for other values of f_0.

Below we shall almost always work with 3-dimensional hypersurfaces in 4-dimensional spacetimes, and for brevity we shall call them simply **surfaces**.

2.2.2 Integrable vector fields

Not every vector field $\mathbf{v}$ admits a family of hypersurfaces for which $\mathbf{v}$ is the normal vector at each point. This property is called **hypersurface orthogonality**. A vector $\mathbf{v}$ is hypersurface orthogonal if there exists a scalar function f such that $\mathbf{v}$ is orthogonal to every vector $\mathbf{t}$ tangent to the surfaces of constant f:

$$g(\mathbf{t}, \mathbf{v}) = 0 \text{ for any } \mathbf{t} \text{ such that } \mathbf{t} \circ f = 0. \tag{2.8}$$

A general condition for hypersurface orthogonality is given by the **Frobenius theorem** (see Sec. 2.2.3). We shall now motivate and derive a weaker condition called **integrability**, which is sufficient for many purposes.

In the language of classical mechanics, the "force" field $\mathbf{v}$ has a "potential" f if $\mathbf{v}$ is the **contravariant gradient** of f. In the language of forms, we say that $\mathbf{v}$ is dual to the 1-form df and write $\mathbf{v} = \hat{g}^{-1}(\mathrm{d}f)$. In that case, we have $\mathbf{t} \circ f \equiv g(\mathbf{t}, \mathbf{v})$ for any vector $\mathbf{t}$, and it is clear that $\mathbf{v}$ is the normal vector for surfaces of constant f. A vector field $\mathbf{v}$ is **integrable** if there exists a scalar function f such that $\mathbf{v}$ is dual to $\mathrm{d}f$.

The analogous property for 1-forms is called **exact**ness. A 1-form ω is **exact** if $\omega = \mathrm{d}f$ for some scalar function f. Then we may say that f is the "integral" of ω, and that a vector field is integrable when it is dual to an exact 1-form.

Qualitatively, the orbits of an integrable vector field $\mathbf{v}$ are good coordinate lines. An immediate example of an integrable vector field is a coordinate basis vector field $\mathbf{e}_j \equiv \partial/\partial x^j$, where $j = 0, 1, 2, 3$. The vector $\mathbf{e}_j$ is orthogonal to the surface of constant x^j and is dual to the exact 1-form $\mathrm{d}x^j$.

Example: Let us produce some explicit examples of non-integrable vector fields.

In flat space with polar coordinates, an example of a non-integrable field would be $\partial/\partial\phi$. In Cartesian coordinates, this would be $\mathbf{v} = y\partial/\partial x - x\partial/\partial y$ which is non-integrable because $\hat{g}\mathbf{v} = y\mathrm{d}x - x\mathrm{d}y$ and $\mathrm{d}(\hat{g}\mathbf{v}) = 2\mathrm{d}y \wedge \mathrm{d}x \neq 0$. The field $\mathbf{v}$ is not a potential field since its orbits are closed circles.

A form ω is **closed** if $\mathrm{d}\omega = 0$. Since $\mathrm{dd} = 0$, every exact form is closed, and *locally* (but not necessarily globally!) every closed form is exact. So a field $\mathbf{v}$ is (locally) integrable if it is dual to a 1-form $\hat{g}\mathbf{v}$ which is closed,

$$\mathrm{d}(\hat{g}\mathbf{v}) = 0.$$

Let us now motivate this well-known condition by more intuitive considerations.

We are trying to find a function $f(p)$ such that $\mathbf{v}$ is the gradient of f. Such a function could be found by integrating the "work done by the force field" $\mathbf{v}$ from some initial point p_0 to the point p. The work corresponding to an "infinitesimal" displacement $\mathbf{l}\delta$ in the direction of a vector $\mathbf{l}$ is

$$g(\mathbf{v}, \mathbf{l})\delta = (\hat{g}\mathbf{v}) \circ \mathbf{l}\delta,$$

where $\hat{g}\mathbf{v}$ is the 1-form corresponding to the vector $\mathbf{v}$. [Recall that the 1-form $\hat{g}\mathbf{v}$ acts on vectors $\mathbf{a}$ as $(\hat{g}\mathbf{v}) \circ \mathbf{a} \equiv g(\mathbf{v}, \mathbf{a})$.] So our candidate function f would be defined by integrating this 1-form along a path leading from p_0 to p:

$$f(p) = \int_{p_0}^{p} \hat{g}\mathbf{v}.$$

If the function f is well-defined then we will have $\mathrm{d}f = \hat{g}\mathbf{v}$ as required. However, the function f is well-defined only if the value $f(p)$ is independent of the path of integration, which is the case only if the integral along any closed loop L vanishes, $\oint_L \hat{g}\mathbf{v} = 0$. By the integration theorem (a generalization of Stokes's law),

$$\oint_L \hat{g}\mathbf{v} = \int_A \mathrm{d}(\hat{g}\mathbf{v}),$$

where A is a surface enclosed by the loop L. Thus the two-form $\mathrm{d}\hat{g}\mathbf{v}$ must be everywhere equal to zero (this follows by considering infinitesimal loops around each point).

Let us now obtain a more convenient form of this condition. The 2-form $\mathrm{d}\hat{g}\mathbf{v}$ acts on arbitrary vectors $\mathbf{a}, \mathbf{b}$ according to Eq. (1.23),

$$\begin{aligned}(\mathrm{d}\hat{g}\mathbf{v}) \circ (\mathbf{a}, \mathbf{b}) &= \mathbf{a} \circ \hat{g}\mathbf{v}(\mathbf{b}) - \mathbf{b} \circ \hat{g}\mathbf{v}(\mathbf{a}) - \hat{g}\mathbf{v}([\mathbf{a}, \mathbf{b}]) \\ &= \mathbf{a} \circ g(\mathbf{v}, \mathbf{b}) - \mathbf{b} \circ g(\mathbf{v}, \mathbf{a}) - g(\mathbf{v}, [\mathbf{a}, \mathbf{b}]) \\ &\equiv \nabla_{\mathbf{a}} g(\mathbf{v}, \mathbf{b}) - \nabla_{\mathbf{b}} g(\mathbf{v}, \mathbf{a}) - g(\mathbf{v}, [\mathbf{a}, \mathbf{b}]) \\ &= g(\nabla_{\mathbf{a}}\mathbf{v}, \mathbf{b}) - g(\nabla_{\mathbf{b}}\mathbf{v}, \mathbf{a}).\end{aligned}$$

The condition we are looking for is therefore

$$g(\nabla_{\mathbf{a}}\mathbf{v}, \mathbf{b}) - g(\nabla_{\mathbf{b}}\mathbf{v}, \mathbf{a}) = 0. \tag{2.9}$$

Written in the index notation using the covariant components of $\mathbf{v}$, this condition becomes

$$\nabla_\alpha v_\beta - \nabla_\beta v_\alpha = 0.$$

This condition is reminiscent of the curl (rotation), $\nabla \times \mathbf{v}$, in three-dimensional vector analysis, where a vector field is a gradient if its curl vanishes. By analogy, the 2-form $\mathrm{d}\hat{g}\mathbf{v}$ is called the **rotation** of the vector field $\mathbf{v}$.

Another way to express the integrability condition is to define the $\mathbf{v}$-dependent bilinear form

$$B_{(\mathbf{v})}(\mathbf{a}, \mathbf{b}) \equiv g(\nabla_{\mathbf{a}}\mathbf{v}, \mathbf{b}), \tag{2.10}$$

and to require that $B_{(\mathbf{v})}$ be a *symmetric* bilinear form, $B_{(\mathbf{v})}(\mathbf{a}, \mathbf{b}) = B_{(\mathbf{v})}(\mathbf{b}, \mathbf{a})$. Note that in general, the rotation 2-form $\mathrm{d}\hat{g}\mathbf{v}$ is twice the antisymmetric part of $B_{(\mathbf{v})}$.

Remark: We have already seen the bilinear form $B_{(\mathbf{v})}$ in Eq. (1.52) when we introduced Killing vectors. For a Killing vector $\mathbf{k}$, the form $B_{(\mathbf{k})}$ is *anti*symmetric and thus equal to $\frac{1}{2}\mathrm{d}\hat{g}\mathbf{k}$; for an integrable vector field $\mathbf{v}$, the form $B_{(\mathbf{v})}$ is symmetric. It follows that $B_{(\mathbf{v})} = 0$ for an *integrable* Killing vector field $\mathbf{v}$; this means $\nabla_{\mathbf{a}}\mathbf{v} = 0$ for all $\mathbf{a}$, i.e. the vector $\mathbf{v}$ is "constant everywhere." The existence of such a vector $\mathbf{v}$ is a very special property of the spacetime, indicating a high degree of symmetry. The orbits of $\mathbf{v}$ are geometrically preferred coordinate lines. For instance, any vectors transported by the flow of $\mathbf{v}$ are parallelly transported. If a timelike Killing vector is hypersurface orthogonal, which is a weaker property than integrability, the spacetime is called **static** because in that case there is a natural "time" coordinate t such that $\mathbf{v}$ is parallel to $\partial/\partial t$ and thus orthogonal to surfaces of constant t. A spacetime is called **stationary** if it admits a timelike (but not necessarily hypersurface-orthogonal) Killing vector.

Statement: An integrable vector field $\mathbf{v}$ which is normalized, $g(\mathbf{v}, \mathbf{v}) = \text{const}$, is always geodesic, $\nabla_{\mathbf{v}}\mathbf{v} = 0$.

Derivation: (See also the proof of the statement in Sec. 2.2.8 where forms are not used.) We use Eq. (2.9) to apply the 2-form $\mathrm{d}\omega \equiv \mathrm{d}(\hat{g}\mathbf{v})$ to $\mathbf{v}$ and an arbitrary vector $\mathbf{a}$:

$$\begin{aligned}(\iota_{\mathbf{v}}\mathrm{d}\omega) \circ \mathbf{a} \equiv (\mathrm{d}\omega) \circ (\mathbf{v}, \mathbf{a}) &= g(\nabla_{\mathbf{v}}\mathbf{v}, \mathbf{a}) - \frac{1}{2}\nabla_{\mathbf{a}} g(\mathbf{v}, \mathbf{v}) \\ &= g(\nabla_{\mathbf{v}}\mathbf{v}, \mathbf{a}).\end{aligned}$$

For an integrable field $\mathbf{v}$, we have $\mathrm{d}\omega = 0$ and so $g(\nabla_{\mathbf{v}}\mathbf{v}, \mathbf{a}) = 0$ for all $\mathbf{a}$; thus $\nabla_{\mathbf{v}}\mathbf{v} = 0$. Note that it does not follow from $\nabla_{\mathbf{v}}\mathbf{v} = 0$ that $\mathrm{d}\omega = 0$, but only that $\iota_{\mathbf{v}}\mathrm{d}\omega = 0$. Thus a normalized geodesic field is not necessarily integrable.

Statement: The **acceleration** of a vector field $\mathbf{v}$ is defined as the vector field $\nabla_{\mathbf{v}}\mathbf{v}$; if the field $\mathbf{v}$ is timelike and $g(\mathbf{v}, \mathbf{v}) = 1$ then $\nabla_{\mathbf{v}}\mathbf{v}$ is interpreted as the acceleration of an observer moving along a worldline of $\mathbf{v}$ with respect to the tangent geodesic line. We show that the acceleration of an integrable vector field is again integrable. (In this case, one recovers a limited version of the Newtonian picture, where the acceleration is determined as the gradient of a "gravitational potential," see Sec. 3.1.1.)

Derivation: By assumption, $\mathbf{v}$ is integrable, so $B_{(\mathbf{v})}$ is symmetric and for an arbitrary vector $\mathbf{x}$ we have

$$\begin{aligned}g(\nabla_{\mathbf{v}}\mathbf{v}, \mathbf{x}) = B_{(\mathbf{v})}(\mathbf{v}, \mathbf{x}) &= B_{(\mathbf{v})}(\mathbf{x}, \mathbf{v}) \\ &= g(\nabla_{\mathbf{x}}\mathbf{v}, \mathbf{v}) = \frac{1}{2}\nabla_{\mathbf{x}} g(\mathbf{v}, \mathbf{v}).\end{aligned}$$

So the 1-form ω dual to the field $\nabla_{\mathbf{v}}\mathbf{v}$ is a gradient of the scalar function $\frac{1}{2}g(\mathbf{v}, \mathbf{v})$; this is the definition of integrability.

2.2.3 Frobenius theorem

The Frobenius theorem gives a sufficient and necessary condition for a vector field $\mathbf{v}$ to be hypersurface orthogonal. The condition is written in the language of differential forms as

$$\omega \wedge \mathrm{d}\omega = 0, \quad \text{where} \quad \omega \equiv \hat{g}\mathbf{v},$$

and $\hat{g}\mathbf{v}$ is the 1-form dual to the vector $\mathbf{v}$. Let us motivate the formulation of the Frobenius theorem before we begin to prove it.

A hypersurface orthogonal vector field $\mathbf{v}$ is everywhere parallel to the contravariant gradient $\hat{g}^{-1}(\mathrm{d}f)$ of some function f describing a family of hypersurfaces $f(p) = \text{const}$. The condition of being parallel is most concisely expressed in the language of differential forms: Two 1-forms ω_1 and ω_2 are "parallel" if $\omega_1 \wedge \omega_2 = 0$. So we are prompted to reformulate the condition (2.8) in the following way. Denote by $\omega \equiv \hat{g}^{-1}\mathbf{v}$ the 1-form dual to $\mathbf{v}$; then the vector $\mathbf{v}$ is hypersurface orthogonal if there exists a function f such that $\omega \wedge \mathrm{d}f = 0$.

However, it is inconvenient to involve an unknown function f in the condition $\omega \wedge \mathrm{d}f = 0$. We can try to eliminate the unknown 1-form $\mathrm{d}f$ from this condition by computing

$$\mathrm{d}(\omega \wedge \mathrm{d}f) = \mathrm{d}\omega \wedge \mathrm{d}f = 0,$$

so $\mathrm{d}\omega$ is also "parallel to $\mathrm{d}f$" just as ω is. Heuristically we may say that both ω and $\mathrm{d}\omega$ are "parallel" to an unknown 1-form $\mathrm{d}f$, so the exterior product of ω and $\mathrm{d}\omega$ must vanish, $\omega \wedge \mathrm{d}\omega = 0$. Thus we have eliminated the unknown 1-form $\mathrm{d}f$ and obtained a condition, $\omega \wedge \mathrm{d}\omega = 0$, that conveniently involves only the given 1-form ω (equivalently, the given vector field $\mathbf{v}$). Now we shall begin the proof of the Frobenius theorem.

We say that an n-form ω is **parallel** to a 1-form $\theta \neq 0$ if $\theta \wedge \omega = 0$. The following lemma states a useful property of parallel forms.

Lemma 1: Suppose ω is an n-form, $n \geq 1$, and $\theta \neq 0$ is a 1-form. If ω is parallel to θ, there exists an $(n-1)$-form ψ such that $\omega = \theta \wedge \psi$. (For $n = 1$ this will be a scalar function ψ and we will have $\omega = \psi\theta$.)

Proof: The idea is to think that ω is "parallel to θ" and to obtain the "components" of ω that are "not parallel to θ." To obtain the components of a form, we can use a basis of vectors. Since $\theta \neq 0$, we can choose a basis $\{\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2, ...\}$ such that $\theta \circ \mathbf{v}_0 = 1$ and $\theta \circ \mathbf{v}_j = 0$ for $j \geq 1$. Then for any basis vectors $\mathbf{v}_{s_j} \neq \mathbf{v}_0$ we have

$$\begin{aligned}(\theta \wedge \omega) \circ (\mathbf{v}_0, \mathbf{v}_{s_1}, \mathbf{v}_{s_2}, ...) = \theta(\mathbf{v}_0) = \omega \circ (\mathbf{v}_{s_1}, \mathbf{v}_{s_2}, ...) \\ = \omega \circ (\mathbf{v}_{s_1}, \mathbf{v}_{s_2}, ...) = 0.\end{aligned}$$

Thus ω is zero on any sets of basis vectors that does not contain $\mathbf{v}_0$, but $\omega(\mathbf{v}_0, \mathbf{v}_{s_1}, \mathbf{v}_{s_2}, ...)$ may be nonzero. So we define the $(n-1)$-form

$$\psi \equiv \iota_{\mathbf{v}_0}\omega; \quad \psi(\mathbf{x}, \mathbf{y}, ...) \equiv \omega(\mathbf{v}_0, \mathbf{x}, \mathbf{y}, ...).$$

It is easy to show that $\omega = \theta \wedge \psi$ on any set of basis vectors $\mathbf{v}_j$ and thus $\omega = \theta \wedge \psi$ on any vectors. Note that the form ψ is defined up to a multiple of θ, because a choice of $\mathbf{v}_0'$ instead of $\mathbf{v}_0$ will yield $\theta \circ (\mathbf{v}_0' - \mathbf{v}_0) = 0$ and thus

$$\psi' = \psi + \iota_{\mathbf{v}_0' - \mathbf{v}_0}\omega = \psi + \iota_{\mathbf{v}_0' - \mathbf{v}_0}(\theta \wedge \psi) = \psi - \theta \wedge \iota_{\mathbf{v}_0' - \mathbf{v}_0}\psi.$$

An immediate corollary is: If two forms ω_1, ω_2 are both parallel to the same 1-form θ then $\omega_1 \wedge \omega_2 = 0$.

By Lemma 1, it follows from $\mathrm{d}f \wedge \mathrm{d}\omega = 0$ that $\omega = \theta \wedge \mathrm{d}f$ for some 1-form θ. Thus the forms ω and $\mathrm{d}\omega$ are both "parallel

to" some unknown $\mathrm{d}f$ and then $\omega \wedge \mathrm{d}\omega = 0$. This is one direction of the Frobenius theorem.

It remains to prove the converse statement, namely that if $\omega \wedge \mathrm{d}\omega = 0$ then there exists a surface which is everywhere orthogonal to $\mathbf{v}$.

We can try to construct such a surface by choosing some vector fields $\mathbf{t}$ orthogonal to $\mathbf{v}$ and following the orbits of all such $\mathbf{t}$. It is sufficient to choose a basis of vector fields $\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3$ orthogonal to $\mathbf{v}$. However, the orbits of these vector fields will not necessarily form a single surface. For instance, we may follow an orbit of a vector $\mathbf{t}_1$ and then a line of $\mathbf{t}_2$, and we shall not necessarily arrive at the same point as when we are following first $\mathbf{t}_2$ and then $\mathbf{t}_1$. For infinitesimal distances, the discrepancy is in the direction of the commutator $[\mathbf{t}_1, \mathbf{t}_2]$. This direction must be orthogonal to $\mathbf{v}$ if all the orbits are to lie within a single surface. Therefore, surfaces orthogonal to $\mathbf{v}$ will exist if for any two vector fields $\mathbf{t}_1, \mathbf{t}_2$ orthogonal to $\mathbf{v}$, the commutator $[\mathbf{t}_1, \mathbf{t}_2]$ is again orthogonal to $\mathbf{v}$. We shall now see that this is indeed the case if $\omega \wedge \mathrm{d}\omega = 0$. **This explanation needs to be made more clear and illustrated by a figure.**

Note that a tangent vector $\mathbf{t}$ is orthogonal to $\mathbf{v}$ if $\omega(\mathbf{t}) = 0$, and that $\mathrm{d}\omega(\mathbf{x}, \mathbf{y})$ involves the application of ω to the commutator $[\mathbf{x}, \mathbf{y}]$, according to Eq. (1.23). This motivates us to formulate and prove the following statement.

Lemma 2: If vector fields $\mathbf{x}$ and $\mathbf{y}$ are everywhere orthogonal to $\mathbf{v}$, and if $\omega \wedge \mathrm{d}\omega = 0$, where ω is the 1-form dual to $\mathbf{v}$, then the vector field $[\mathbf{x}, \mathbf{y}]$ is also orthogonal to $\mathbf{v}$.

Proof: The conditions on $\mathbf{x}$ and $\mathbf{y}$ can be written as $\omega(\mathbf{x}) = \omega(\mathbf{y}) = 0$. By Lemma 1, $\mathrm{d}\omega = \theta \wedge \omega$ for some 1-form θ. Then we have

$$(\mathrm{d}\omega) \circ (\mathbf{x}, \mathbf{y}) = (\theta \wedge \omega) \circ (\mathbf{x}, \mathbf{y}) = \theta(\mathbf{x})\omega(\mathbf{y}) - \theta(\mathbf{y})\omega(\mathbf{x}) = 0.$$

On the other hand, from Eq. (1.23) we find

$$(\mathrm{d}\omega) \circ (\mathbf{x}, \mathbf{y}) = \mathbf{x} \circ \omega(\mathbf{y}) - \mathbf{y} \circ \omega(\mathbf{x}) - \omega([\mathbf{x}, \mathbf{y}]) = -\omega([\mathbf{x}, \mathbf{y}]).$$

Thus $\omega([\mathbf{x}, \mathbf{y}]) = 0$.

It follows from Lemma 2 that vector fields orthogonal to $\mathbf{v}$ will have surface-forming orbits if $\omega \wedge \mathrm{d}\omega = 0$. This concludes the proof of the Frobenius theorem.

Here is another useful property of hypersurface-orthogonal geodesic fields.

Statement 1: A hypersurface-orthogonal, normalized, spacelike or timelike, geodesic field $\mathbf{v}$ is always integrable.

Proof: For a normalized and geodesic field $\mathbf{v}$, we have $\nabla_{\mathbf{v}}\mathbf{v} = 0$ and $g(\mathbf{v}, \mathbf{v}) = \text{const}$, so Eq. (2.9) gives

$$(\iota_{\mathbf{v}}\mathrm{d}\omega) \circ \mathbf{x} \equiv (\mathrm{d}\omega) \circ (\mathbf{v}, \mathbf{x}) = g(\nabla_{\mathbf{v}}\mathbf{v}, \mathbf{x}) - g(\nabla_{\mathbf{x}}\mathbf{v}, \mathbf{v}) = 0. \tag{2.11}$$

We say that the 2-form $\mathrm{d}\omega$ is **transverse** to $\mathbf{v}$. The proof can now proceed in one of two ways. If we consider the 2-form $\iota_{\mathbf{v}}(\omega \wedge \mathrm{d}\omega)$ which vanishes since $\omega \wedge \mathrm{d}\omega = 0$, we find

$$0 = \iota_{\mathbf{v}}(\omega \wedge d\omega) = (\iota_{\mathbf{v}}\omega)\, d\omega - \omega \wedge (\iota_{\mathbf{v}} d\omega) = \omega(\mathbf{v}) d\omega.$$

Since the field $\mathbf{v}$ is by assumption not null, $\omega(\mathbf{v}) \equiv g(\mathbf{v}, \mathbf{v}) \neq 0$, and it follows that $\mathrm{d}\omega = 0$, i.e. the field $\mathbf{v}$ is integrable. Another, more constructive way to prove the statement is the following. Since $\mathbf{v}$ is hypersurface-orthogonal, there exists a family of surfaces whose normal vectors are everywhere parallel to $\mathbf{v}$. This family of surfaces can be described by a function ν, and the vector field normal to the surfaces is $\hat{g}^{-1}\mathrm{d}\nu$. Since this field is parallel to $\mathbf{v}$, there exists a scalar function $\lambda \neq 0$ such that $\mathbf{v} = \lambda\hat{g}^{-1}\mathrm{d}\nu$ or equivalently $\omega = \lambda \mathrm{d}\nu$ and $\mathrm{d}\omega = \mathrm{d}\lambda \wedge \mathrm{d}\nu$. Then Eq. (2.11) gives

$$0 = \iota_{\mathbf{v}}\mathrm{d}\omega = \mathrm{d}\lambda(\mathbf{v})\mathrm{d}\nu - \mathrm{d}\nu(\mathbf{v})\mathrm{d}\lambda \equiv (\mathbf{v} \circ \lambda)\mathrm{d}\nu - \frac{1}{\lambda}\omega(\mathbf{v})\mathrm{d}\lambda.$$

Since $\omega(\mathbf{v}) \neq 0$ and $\lambda \neq 0$, we can express

$$\mathrm{d}\lambda = \frac{(\mathbf{v} \circ \lambda)\,\lambda}{\omega(\mathbf{v})}\mathrm{d}\nu,$$

and it follows that $\mathrm{d}\omega = \mathrm{d}\lambda \wedge \mathrm{d}\nu = 0$.

It is interesting to see where the argument fails for *null* fields $\mathbf{v}$: We cannot divide by $\omega(\mathbf{v})$ because $\omega(\mathbf{v}) = g(\mathbf{v}, \mathbf{v}) = 0$. Indeed, we can still find λ and ν such that $\omega = \lambda\mathrm{d}\nu$, but the transversality $\iota_{\mathbf{v}}\mathrm{d}\omega = 0$ yields only $\mathbf{v} \circ \lambda = 0$. This condition can be satisfied by a function λ whose gradients are orthogonal to the null vector $\mathbf{v}$, in other words, $g^{-1}(\mathrm{d}\lambda, \omega) = 0$. However, the gradient $\mathrm{d}\lambda$ does not need to be parallel to ω, and thus we might have $\mathrm{d}\omega = \mathrm{d}\lambda \wedge \mathrm{d}\nu \neq 0$. An explicit example can be easily found in Minkowski spacetime: Choose $\nu = t - x$ and $\lambda = y$, such that $\mathrm{d}\lambda$ is orthogonal to $\mathrm{d}\nu$. Then the vector field dual to $\omega \equiv \lambda\mathrm{d}\nu$ is $\mathbf{v} = y(\partial_t + \partial_x)$, and we have $\iota_{\mathbf{v}}\omega = 0$. So $\mathbf{v}$ is geodesic, null, and hypersurface orthogonal (to surfaces of constant $x - t$), but non-integrable since it has nonvanishing rotation $\mathrm{d}\omega = \mathrm{d}y \wedge (\mathrm{d}t - \mathrm{d}x) \neq 0$.

Statement: A hypersurface-orthogonal field (even normalized and even timelike) does not necessarily have an integrable acceleration $\nabla_{\mathbf{v}}\mathbf{v}$.

Hint: It is not necessarily true that $\mathrm{d}\iota_{\mathbf{v}}\mathrm{d}\omega \neq 0$ even if $\iota_{\mathbf{v}}\omega = 0$.

Derivation: Explicit counterexamples are the 1-forms $\omega = (x + y)\mathrm{d}x$, $\omega = x(\mathrm{d}t - \mathrm{d}x)$, $\omega = \cosh x\,\mathrm{d}t + \sinh x\,\mathrm{d}x$ in the Minkowski spacetime. These 1-forms are dual to hypersurface-orthogonal fields since $\omega \wedge \mathrm{d}\omega = 0$, but they do not have integrable acceleration.

Statement: A geodesic vector field $\mathbf{v}$ is hypersurface-orthogonal *everywhere* if there exists a *single* 3-surface Σ to which $\mathbf{v}$ is orthogonal. The same statement holds also for Killing fields $\mathbf{k}$ instead of geodesic vector fields.

Derivation: Recall that $\nabla_{\mathbf{v}} g(\mathbf{v}, \mathbf{v}) = 0$, thus $g(\mathbf{v}, \mathbf{v})$ is constant along the lines of $\mathbf{v}$, and so the affine tangent vector $\mathbf{v}$ can be rescaled, $\mathbf{v} \to \tilde{\mathbf{v}} = \lambda\mathbf{v}$, where $\lambda \neq 0$ is a scalar function such that $\nabla_{\mathbf{v}}\lambda = 0$ and $g(\tilde{\mathbf{v}}, \tilde{\mathbf{v}}) = \text{const}$ everywhere. The vector field $\tilde{\mathbf{v}}$ is normalized, geodesic, and orthogonal to the 3-surface Σ. Since $\tilde{\mathbf{v}}$ is everywhere parallel to $\mathbf{v}$, it is sufficient to show that $\tilde{\mathbf{v}}$ is hypersurface orthogonal everywhere. Let $\omega = \hat{g}\tilde{\mathbf{v}}$ be the 1-form dual to $\tilde{\mathbf{v}}$; then $\omega \wedge \mathrm{d}\omega = 0$ on Σ, and we need to prove that $\omega \wedge \mathrm{d}\omega$ remains zero away from Σ. We shall now show that $\mathcal{L}_{\tilde{\mathbf{v}}}(\omega \wedge \mathrm{d}\omega) = 0$. This would mean that the Lie-propagated 3-form $\omega \wedge \mathrm{d}\omega$ remains zero along the orbits of $\tilde{\mathbf{v}}$ when applied to a Lie-propagated basis of connecting vectors to $\tilde{\mathbf{v}}$; that property is equivalent to having $\omega \wedge \mathrm{d}\omega = 0$ everywhere. We compute

$$\begin{aligned}(\mathcal{L}_{\tilde{\mathbf{v}}}\omega) \circ \mathbf{x} &= \mathcal{L}_{\tilde{\mathbf{v}}}(\omega \circ \mathbf{x}) - \omega \circ \mathcal{L}_{\tilde{\mathbf{v}}}\mathbf{x} = \\ &= \nabla_{\tilde{\mathbf{v}}} g(\tilde{\mathbf{v}}, \mathbf{x}) - g(\tilde{\mathbf{v}}, [\tilde{\mathbf{v}}, \mathbf{x}]) = g(\tilde{\mathbf{v}}, \nabla_{\mathbf{x}}\tilde{\mathbf{v}}) \\ &= \frac{1}{2}\nabla_{\mathbf{x}} g(\tilde{\mathbf{v}}, \tilde{\mathbf{v}}) = 0,\end{aligned}$$

since $g(\tilde{\mathbf{v}}, \tilde{\mathbf{v}}) = \text{const}$. Recall that $\mathcal{L}$ commutes with d on any n-forms, hence

$$\mathcal{L}_{\tilde{\mathbf{v}}}\mathrm{d}\omega = \mathrm{d}\mathcal{L}_{\tilde{\mathbf{v}}}\omega = 0,$$

and then we have

$$\mathcal{L}_{\tilde{\mathbf{v}}}(\omega \wedge \mathrm{d}\omega) = (\mathcal{L}_{\tilde{\mathbf{v}}}\omega) \wedge \mathrm{d}\omega + \omega \wedge \mathcal{L}_{\tilde{\mathbf{v}}}\mathrm{d}\omega = 0.$$

Alternatively, here is a geometric argument. We can construct a family of 3-surfaces $\Sigma(\tau)$ which are surfaces of constant affine parameter τ measured along orbits of $\tilde{\mathbf{v}}$, where we set $\tau = 0$ on Σ. We can choose a basis of connecting vectors $\mathbf{c}_j$, $j = 1, 2, 3$, such that $[\mathbf{c}_j, \tilde{\mathbf{v}}] = 0$ and $\mathbf{c}_j$ are tangent to Σ and thus orthogonal to $\tilde{\mathbf{v}}$ on Σ. These connecting vectors $\mathbf{c}_j$ will remain orthogonal to $\tilde{\mathbf{v}}$ for all τ (see Lemma 1 in Sec. 2.2.6, where we need to replace $\mathbf{n}$ by $\tilde{\mathbf{v}}$). Therefore, the 3-surfaces $\Sigma(\tau)$ are everywhere orthogonal to $\tilde{\mathbf{v}}$ and thus also to $\mathbf{v}$. Now consider the case of a Killing field $\mathbf{k}$ and define $\omega \equiv \hat{g}\mathbf{k}$. It is sufficient to prove that $\mathcal{L}_{\mathbf{k}}(\omega \wedge \mathrm{d}\omega) = 0$. Since $\mathbf{k}$ is a Killing vector, we have $\mathcal{L}_{\mathbf{k}} g = 0$, so

$$\mathcal{L}_{\mathbf{k}}\omega = \mathcal{L}_{\mathbf{k}}\hat{g}\mathbf{k} = \hat{g}\mathcal{L}_{\mathbf{k}}\mathbf{k} = 0.$$

We again recall that $\mathcal{L}$ commutes with d, thus

$$\mathcal{L}_{\mathbf{k}}\left(\omega \wedge \mathrm{d}\omega\right) = \omega \wedge \mathcal{L}_{\mathbf{k}}\mathrm{d}\omega = \omega \wedge \mathrm{d}\mathcal{L}_{\mathbf{k}}\omega = 0.$$

In summary, we have the following relationships between integrability, hypersurface orthogonality, and metric properties of a vector field. The condition for hypersurface orthogonality, $\omega \wedge \mathrm{d}\omega = 0$, is weaker than the condition for integrability, $\mathrm{d}\omega = 0$. An integrable field $\mathbf{v} = \hat{g}^{-1}(\mathrm{d}f)$ is always hypersurface orthogonal (to the hypersurfaces of constant f). An integrable field $\mathbf{v}$ which is normalized, $g(\mathbf{v}, \mathbf{v}) = \text{const}$, is always geodesic. A hypersurface orthogonal, normalized, geodesic field is integrable unless it is null (a geodesic field can always be rescaled to make it normalized). The acceleration $\nabla_{\mathbf{v}}\mathbf{v}$ of an integrable field $\mathbf{v}$ is again integrable.

2.2.4 Null surfaces

A **null surface** is a 3-surface whose normal vector $\mathbf{n}$ is everywhere null and nonzero, $g(\mathbf{n}, \mathbf{n}) = 0$, $\mathbf{n} \neq 0$. We shall now study the geometry of null surfaces since they have certain special and useful properties.

One unusual fact is that the normal vector is at the same time *tangent* to the null surface since $g(\mathbf{n}, \mathbf{n}) = 0$. The tangent space to a null surface is thus spanned by $\mathbf{n}$ and two other basis vectors. The latter two vectors must be orthogonal to $\mathbf{n}$, and thus both must be spacelike. In fact, all the tangent vectors that are not parallel to $\mathbf{n}$ are spacelike (see the Statement 2.1.0.1). Thus, a null surface is spanned by the null direction $\mathbf{n}$ and two spacelike directions orthogonal to $\mathbf{n}$.

Since the tangent space contains its normal vector, the induced metric in the tangent space is *degenerate* since the vector $\mathbf{n}$ is orthogonal to every vector in the tangent space. This leads to an ambiguity of the induced Levi-Civita connection on the null surface: The formula (1.45), which is the only constraint on the Levi-Civita connection, does not define the vector $\nabla_{\mathbf{x}}\mathbf{y}$ uniquely through its scalar product with an arbitrary vector $\mathbf{z}$. Thus there are many possible Levi-Civita connections on a null surface.

Another peculiarity is that there is no unique projector onto the tangent space (see Sec. 2.1.1). A satisfactory projector (2.2) can be found only after choosing another null vector field $\mathbf{l}$ normalized by $g(\mathbf{n}, \mathbf{l}) = 1$, and then the image of the projection is a two-dimensional subspace within the tangent space.

Self-test question: Do we need separate considerations for surfaces whose tangent space is spanned by (a) a null vector, a timelike vector, and a spacelike vector, (b) two null vectors and a spacelike vector, (c) three null vectors?

Answer: No. Each of the described surfaces is in fact **timelike**, i.e. orthogonal to a timelike vector. Their tangent spaces are also spanned by one timelike and two spacelike vectors.

2.2.5 Examples of null surfaces

The simplest example of a null surface is the null cone of lightrays emitted from a point. Let us describe this surface explicitly in a flat Minkowski spacetime. The lightcone emitted at the origin is specified by the equation

$$t - r \equiv t - \sqrt{x^2 + y^2 + z^2} = 0. \tag{2.12}$$

The vector field $\mathbf{n}$ normal to this surface is

$$\mathbf{n} = \frac{\partial}{\partial t} + \frac{x}{r}\frac{\partial}{\partial x} + \frac{y}{r}\frac{\partial}{\partial y} + \frac{z}{r}\frac{\partial}{\partial z}.$$

It is easy to check that $g(\mathbf{n}, \mathbf{n}) = 0$. The tangent space to the lightcone at a point (t, x, y, z) is spanned by the null vector $\mathbf{n}$ and the two spacelike vectors $\partial/\partial\phi$ and $\partial/\partial\theta$,

$$\frac{\partial}{\partial\phi} \equiv \frac{1}{\sqrt{x^2 + y^2}}\left(y\frac{\partial}{\partial x} - x\frac{\partial}{\partial y}\right),$$
$$\frac{\partial}{\partial\theta} = \frac{z}{r}\left(x\frac{\partial}{\partial x} + y\frac{\partial}{\partial y}\right) - \frac{x^2 + y^2}{r}\frac{\partial}{\partial z}.$$

Another example of a null surface is the horizon of a nonrotating, spherically symmetric black hole (Schwarzschild spacetime). The metric is given by Eq. (1.38). It is well known that the black hole horizon is the surface $r - 2m = 0$. However, it is not easy to analyze vectors at $r = 2m$ since the Schwarzschild coordinates are singular there. To show that the horizon surface is null, we note that its tangent space is spanned by the two spacelike vectors $\partial/\partial\phi$, $\partial/\partial\theta$, and also by the vector $\frac{\partial}{\partial t}$ which is null at $r = 2m$ (but not at other values of r). Since $\partial/\partial t$ is null, it is also the normal vector to the surface and hence the surface is null.

2.2.6 Lightcones are null surfaces

We can consider the surface formed by *all* the null geodesics emitted from a given initial spacetime point p_0 in all directions. This surface is called the **lightcone** emitted at p_0.

Statement: The lightcone emitted at a point p_0 is always a null surface, in an arbitrary curved spacetime.

We begin the proof by a preliminary calculation.

Lemma 1: If $\mathbf{n}$ is normalized, $g(\mathbf{n}, \mathbf{n}) = \text{const}$, and geodesic, $\nabla_{\mathbf{n}}\mathbf{n} = 0$, and if $\mathbf{c}$ is a connecting vector for $\mathbf{n}$, then $g(\mathbf{n}, \mathbf{c})$ is constant along the orbits of $\mathbf{n}$.

Proof: Since $\nabla_{\mathbf{c}}\mathbf{n} = \nabla_{\mathbf{n}}\mathbf{c}$ and $\nabla_{\mathbf{n}}\mathbf{n} = 0$, we have

$$\mathbf{n} \circ g(\mathbf{n}, \mathbf{c}) = g(\mathbf{n}, \nabla_{\mathbf{n}}\mathbf{c}) = g(\mathbf{n}, \nabla_{\mathbf{c}}\mathbf{n}) = \frac{1}{2}\nabla_{\mathbf{c}} g(\mathbf{n}, \mathbf{n}) = 0.$$

Lemma 2: The tangent space to a lightcone at any point p is spanned by one null vector and two spacelike vectors that are both orthogonal to the null vector.

Proof: We first note that in the immediate vicinity of the initial point p_0 the lightcone can be viewed, in suitable local coordinates, as a portion of the Minkowski lightcone (2.12). Thus the tangent space in the vicinity of p_0 is indeed spanned by one null vector and two spacelike vectors orthogonal to it. We now need to extend this property away from the initial point p_0.

By assumption, the lightcone surface is built *entirely* from the congruence of lightrays emitted at p_0. Translating this assumption into the mathematical language, we say that every point on the surface lies on some lightray from the congruence, and thus the surface can be parametrically described by a function

$$\gamma(\tau; s_1, s_2) : \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to M$$

such that $\gamma(\tau; s_1, s_2)$ is a lightray for fixed s_1 and s_2. (Note that there is a two-dimensional sphere of initial directions for emitting the lightrays, hence two parameters $s_{1,2}$.) In particular, $\mathbf{n} = \partial\gamma/\partial\tau$ is the tangent vector field which is null, and the vectors $\mathbf{u}_1 \equiv \partial\tau/\partial s_1$, $\mathbf{u}_2 \equiv \partial\tau/\partial s_2$ are everywhere linearly independent, connecting vectors for $\mathbf{n}$ (see the Statement 13). Thus we have the properties

$$g(\mathbf{n}, \mathbf{n}) = 0, \quad \nabla_{\mathbf{n}}\mathbf{n} = 0,$$
$$[\mathbf{n}, \mathbf{u}_1] = [\mathbf{n}, \mathbf{u}_2] = 0.$$

By construction, the connecting vectors $\mathbf{u}_{1,2}$ are always tangent to the surface, and so is $\mathbf{n}$. Thus the vectors $(\mathbf{n}, \mathbf{u}_1, \mathbf{u}_2)$ are always a basis in the tangent space to the surface. The vector $\mathbf{n}$ is everywhere null by assumption, so it remains to show that $\mathbf{u}_1$ and $\mathbf{u}_2$ remain spacelike throughout the surface, at any point p.

The Minkowski description of the lightcone holds for an infinitesimal vicinity of p_0, and thus we have $g(\mathbf{n}, \mathbf{u}_1) = g(\mathbf{n}, \mathbf{u}_2) = 0$ near p_0. Then it follows from Lemma 1 that $g(\mathbf{n}, \mathbf{u}_1) = g(\mathbf{n}, \mathbf{u}_2) = 0$ everywhere along the orbits of $\mathbf{n}$. Vectors orthogonal to a null vector and not parallel to it must be spacelike (see Statement 2.1.0.1). Since each point p on the surface can be reached by *some* orbit of $\mathbf{n}$ starting from p_0, the vectors $\mathbf{u}_{1,2}$ remain spacelike and orthogonal to $\mathbf{n}$ everywhere on the surface.

Corollary: The null tangent vector $\mathbf{n}$ at a point p is also the normal vector to the surface.

Proof: By Lemma 2, the tangent space at p is spanned by $\mathbf{n}$ and two spacelike connecting vectors $\mathbf{u}_1, \mathbf{u}_2$ that remain orthogonal to $\mathbf{n}$. Thus, $\mathbf{n}$ is orthogonal to every vector in the entire 3-dimensional tangent space, which means that $\mathbf{n}$ is the normal vector to the surface.

We have proved that a lightcone is a null surface. A lightcone is by construction a surface built from lines (the individual lightrays). These lines are called the **null generators** of the surface. Now we shall use the formalism of "null functions" to prove the converse statement: Any null surface is always built from null generators (even though not every null surface is a lightcone emitted from one point).

2.2.7 Null functions

A null surface can be defined through an equation $\nu(p) = 0$, and then the function $\nu(p)$ is called a **null function**. More generally, a null function $\nu(p)$ on the spacetime is any function such that $\mathrm{d}\nu$ is a null covector, $g^{-1}(\mathrm{d}\nu, \mathrm{d}\nu) = 0$. Immediate examples are the functions

$$\nu(t, x, y, z) = t - r \equiv t - \sqrt{x^2 + y^2 + z^2},$$
$$\text{or } \nu(t, x, y, z) = t - x,$$

in the Minkowski spacetime (verify this!). A null function $\nu(p)$ defines a null hypersurface $\nu(p) = \nu_0$ for each value of ν_0.

A null function can be physically interpreted as a "retarded time" carried by lightrays emitted by some sources: If a lightray reaches an event p, it was emitted at the time $\nu(p)$ at its source. (Here we do not assume that all the rays are emitted by one and the same source.) To remember this more easily, recall that time stands still for an observer moving at lightspeed, so a lightray "preserves" its emission time and shows it at a point p as the value of $\nu(p)$.

A peculiar property of null functions $\nu(p)$ is that the normal vector $\mathbf{n} = \hat{g}^{-1}(\mathrm{d}\nu)$ is tangent to the surface of constant ν. Thus a null surface $\nu(p) = \nu_0$ naturally comes with a congruence of null lines, which are the orbits of $\mathbf{n}$ within the surface and the null generators of the surface.

2.2.8 Null functions generate null geodesics

The usefulness of null functions comes from the fact that any null surface consists not just of some null lines, but actually of *null geodesics* (see the following statement). These geodesics are the null generators of the surface. Since any null surface is a surface $\nu(p) = \nu_0$ for some null function ν, it will follow that null functions provide a complete description of any null surface consisting of lightrays.

Statement: If $\nu(p)$ is a null function with the normal vector $\mathbf{n}$, the orbits of $\mathbf{n}$ are null geodesics. More generally, if a null curve $\gamma(\tau)$ is contained within a surface $\nu(p) = \text{const}$ then $\gamma(\tau)$ is a null geodesic.

Proof: A stronger statement was proved in Sec. 2.2.2: If $\mathbf{n}$ is integrable and $g(\mathbf{n}, \mathbf{n}) = \text{const}$, where the constant is not necessarily zero, then $\nabla_{\mathbf{n}}\mathbf{n} = 0$. We shall give a direct derivation that does not use forms. To show that $\nabla_{\mathbf{n}}\mathbf{n} = 0$, we shall prove that $g(\mathbf{x}, \nabla_{\mathbf{n}}\mathbf{n}) = 0$ for an arbitrary vector $\mathbf{x}$. We compute

$$\begin{aligned} g(\mathbf{x}, \nabla_{\mathbf{n}}\mathbf{n}) &= \nabla_{\mathbf{n}} g(\mathbf{x}, \mathbf{n}) - g(\nabla_{\mathbf{n}}\mathbf{x}, \mathbf{n}) \\ &= \mathbf{n} \circ g(\mathbf{x}, \mathbf{n}) - g(\nabla_{\mathbf{x}}\mathbf{n} + [\mathbf{n}, \mathbf{x}], \mathbf{n}). \end{aligned}$$

(We have used torsion-freeness and compatibility of ∇ with the metric.) Since by assumption $g(\mathbf{n}, \mathbf{n}) = \text{const}$ and $\mathbf{n} = \hat{g}^{-1}\mathrm{d}\nu$, we have $g(\mathbf{n}, \nabla_{\mathbf{x}}\mathbf{n}) = 0$ and $g(\mathbf{x}, \mathbf{n}) = \mathbf{x} \circ \nu$ [see also Eq. (2.7)]. Thus we find

$$\begin{aligned} g(\mathbf{x}, \nabla_{\mathbf{n}}\mathbf{n}) &= \mathbf{n} \circ g(\mathbf{x}, \mathbf{n}) - g([\mathbf{n}, \mathbf{x}], \mathbf{n}) \\ &= \mathbf{n} \circ (\mathbf{x} \circ \nu) - ([\mathbf{n}, \mathbf{x}]) \circ \nu = \mathbf{x} \circ (\mathbf{n} \circ \nu) \\ &= \mathbf{x} \circ g(\mathbf{n}, \mathbf{n}) = 0. \end{aligned}$$

It follows that the orbits of $\mathbf{n}$ are geodesics for integrable null $\mathbf{n}$. A tangent space to a null surface contains only one null direction, and that direction is parallel to $\mathbf{n}$. So any null curve contained within the surface must be an orbit of $\mathbf{n}$ and hence a geodesic.

2.2.9 Every lightray comes from null functions

The following statement demonstrates that any null geodesic can be obtained from a suitable null function.

Statement: For any null geodesic curve $\gamma(\tau)$ there exists a null function $\nu(p)$ such that $\dot{\gamma} = \hat{g}^{-1}(\mathrm{d}\nu)$ on the curve, so that $\gamma(\tau)$ is a null generator of $\nu(p) = 0$.

Proof: There are of course many possible null functions satisfying this condition. We shall construct one such null function $\nu(p)$ by choosing an arbitrary timelike curve ζ intersecting γ at a point p_0. This curve will be the worldline of a "light source." Suppose that the source emits lightrays in all directions from each point of the curve ζ, and that $\gamma(\tau)$ is one of these lightrays emitted from p_0. If σ is a suitable parameter on the curve $\zeta(\sigma)$, such that the tangent vector is normalized, $g(\dot{\zeta}, \dot{\zeta}) = 1$, then σ is the observer's "proper time" measured along the curve $\zeta(\sigma)$. We may suppose that the point p_0 corresponds to $\sigma = 0$. Then we define the desired null function $\nu(p)$ as the "retarded time" carried by the lightrays from the source. More precisely, we define $\nu(p) = \sigma$ if the point p is intersected by a lightray emitted from the point $\zeta(\sigma)$. For instance, we then have $\nu(p_0) = 0$. The function $\nu(p)$ is well-defined in some neighborhood of the curve ζ; more precisely, in a neighborhood where the lightcones are well-behaved, i.e. do not form caustics and fill the entire space around the curve ζ, leaving no gaps. By construction, the function $\nu(p)$ is constant along $\gamma(\tau)$ and thus the curve $\gamma(\tau)$ lies within the surface $\nu(p) = 0$, and the null vector $\dot{\gamma}(\tau)$ is everywhere tangent to the surface $\nu(p) = 0$. It remains to show that $\nu(p)$ is a null function; this is so because $\nu(p) = \text{const}$ is by construction a lightcone surface, and we have already proved that any lightcone surface is null.

2.2.10 Conformal invariance

Under a conformal transformation of the metric, $g \to \tilde{g} = e^{2\lambda} g$, null vectors obviously remain null. Therefore null functions remain null under a conformal transformation (i.e. they are conformally invariant).

Also, shapes of null geodesics are conformally invariant: a null geodesic line $\gamma(\tau)$ remains a geodesic after a conformal transformation of the metric. This statement is now very easy to prove: a null geodesic is always an orbit of a gradient of some null function $\nu(p)$, the gradient operation d is independent of the metric, and so the property of being a null function is conformally invariant. It follows that the normal vector $\mathbf{n} \equiv \hat{g}^{-1}(\mathrm{d}\nu)$ will change to $\tilde{\mathbf{n}} = e^{-2\lambda}\mathbf{n}$, and thus the orbits of $\tilde{\mathbf{n}}$ are the same as those of $\mathbf{n}$. The affine parameter τ must of course be changed, but the shape of the lines will remain the same.

2.3 Raychaudhuri equation

We now consider a geodesic congruence with a given tangent vector field. The distortion of the congruence along its own orbits will be determined by the geometry of the given spacetime. The distortion satisfies a purely kinematical relation called the Raychaudhuri equation. We shall derive separate versions of this equation for timelike and for null geodesics. We begin by introducing some geometric quantities characterizing the distortion of geodesic congruences.

2.3.1 Distortion tensor

In this section we consider a normalized, geodesic vector field $\mathbf{u}$ which may be either timelike, spacelike, or null. We have seen in Sec. 1.6.9 that the divergence $\mathrm{div}\mathbf{u}$ of a vector field $\mathbf{u}$ measures the relative rate of change of infinitesimal volume transported by the flow of $\mathbf{u}$. Consider an small region of space around a point p, and imagine that the entire region moves with the flow (we call this a **comoving** region). This region can be visualized as consisting of a number of small, noninteracting particles, each moving along its flow line. The region will be generally distorted by the flow, so that not only its volume but also its shape will change. The distortion of the shape is measured by tensors called the shear and the rotation, which we shall now introduce.

The shape of the comoving region is naturally characterized by connecting vectors $\mathbf{c}$. For convenience, we can choose $\mathbf{c}$ to be orthogonal to $\mathbf{u}$ at the initial point p. It then follows from Lemma 1 in Sec. 2.2.6 that $\mathbf{c}$ remains orthogonal to $\mathbf{u}$ along the orbits of $\mathbf{u}$.

There would be no distortion of the region if every connecting vector $\mathbf{c}$ were parallelly transported by the flow, i.e. if $\nabla_{\mathbf{u}}\mathbf{c} = 0$ for every connecting $\mathbf{c}$. Of course, in general we expect that $\nabla_{\mathbf{u}}\mathbf{c} \neq 0$ at least for some $\mathbf{c}$. Thus, the distortion of the volume around the point p during an infinitesimal time will be described by the derivative $\nabla_{\mathbf{u}}\mathbf{c}$ along the flow. Since

$$\nabla_{\mathbf{u}}\mathbf{c} = \nabla_{\mathbf{c}}\mathbf{u},$$

the quantity $\nabla_{\mathbf{u}}\mathbf{c}$ is a linear function of $\mathbf{c}$ that can be described by a $\mathbf{u}$-dependent transformation $B(\mathbf{u})$, which is a (1,1) rank tensor defined by

$$\nabla_{\mathbf{u}}\mathbf{c} = \nabla_{\mathbf{c}}\mathbf{u} \equiv B(\mathbf{u})\mathbf{c}.$$

In components,

$$B^{\beta}_{\alpha} = \nabla_{\alpha} u^{\beta} \equiv u^{\beta}_{;\alpha}.$$

The transformation $B(\mathbf{u})$ can be also written as

$$B(\mathbf{u}) = \nabla_{\mathbf{v}} - \mathcal{L}_{\mathbf{v}},$$

where it is implied that both sides act on a vector field. We call $B(\mathbf{u})$ the **distortion tensor** since it quantifies the deviation of the shape of a small comoving region from being parallelly transported.

The divergence of $\mathbf{u}$ is equal to the trace of $B(\mathbf{u})$:

$$\mathrm{div}\mathbf{u} \equiv \nabla_{\alpha} u^{\alpha} = \mathrm{Tr}\, B(\mathbf{u}).$$

Even if the divergence is zero, but $B \neq 0$, the shape of the region will be changed by the flow. To analyze the information provided by $B(\mathbf{u})$, it is convenient to consider the bilinear form

$$B_{(\mathbf{u})}(\mathbf{x}, \mathbf{y}) \equiv g(B(\mathbf{u})\mathbf{x}, \mathbf{y}) = g(\nabla_{\mathbf{u}}\mathbf{x}, \mathbf{y}).$$

We have seen in Sec. 2.2.2 that $B_{(\mathbf{u})}$ is a *symmetric* bilinear form iff $\mathbf{u}$ is an integrable vector field.

Practice problem: Compute the divergence of the null vector field corresponding to a lightcone $t = r$ in Minkowski spacetime.

Solution: The vector field is

$$\mathbf{u} = \partial_t + \frac{1}{r}\left(x\partial_x + y\partial_y + z\partial_z\right);$$

using the flat connection, we find $\mathrm{div}\mathbf{u} = 2r^{-1}$.

2.3.2 Rotation

We may decompose $B_{(\mathbf{u})}$ into symmetric and asymmetric parts,

$$B_{(\mathbf{u})}(\mathbf{x},\mathbf{y}) = \tilde{B}_{(\mathbf{u})}(\mathbf{x},\mathbf{y}) + r_{(\mathbf{u})}(\mathbf{x},\mathbf{y}), \tag{2.13}$$
$$\tilde{B}_{(\mathbf{u})}(\mathbf{x},\mathbf{y}) \equiv \frac{1}{2}\left(B_{(\mathbf{u})}(\mathbf{x},\mathbf{y}) + B_{(\mathbf{u})}(\mathbf{y},\mathbf{x})\right),$$
$$r_{(\mathbf{u})}(\mathbf{x},\mathbf{y}) \equiv \frac{1}{2}\left(B_{(\mathbf{u})}(\mathbf{x},\mathbf{y}) - B_{(\mathbf{u})}(\mathbf{y},\mathbf{x})\right).$$

The 2-form $r_{(\mathbf{u})}$ is called the **rotation** (or the **twist** , or the **vorticity**) of the field $\mathbf{u}$. It is easy to see from Eq. (2.9) that

$$r_{(\mathbf{u})} = \frac{1}{2}\mathrm{d}\hat{g}\mathbf{u} \equiv \frac{1}{2}\mathrm{d}\omega,$$

where ω is the 1-form dual to the vector field $\mathbf{u}$. The rotation tensor $r_{(\mathbf{u})}$ is thus the analog of the familiar three-dimensional curl (rotation) $\nabla \times \mathbf{v}$ of a vector field. A nonzero rotation means that the vector field is not a potential field; in our terminology, a nonzero $r_{(\mathbf{u})}$ means that the field $\mathbf{u}$ is not integrable.

To visualize the shape distortion represented by a nonzero $r_{(\mathbf{u})}$, consider a flat (Minkowski) spacetime in which $\mathbf{u}$ is timelike and $r_{(\mathbf{u})}$ is a constant tensor, represented by a constant antisymmetric matrix $r_{\alpha\beta}$ in an orthogonal basis. Denote by $\hat{r}_{(\mathbf{u})}$ the transformation corresponding to the 2-form $r_{(\mathbf{u})}$, so that for all vectors $\mathbf{x}, \mathbf{y}$ we have

$$g(\hat{r}_{(\mathbf{u})}\mathbf{x},\mathbf{y}) = r_{(\mathbf{u})}(\mathbf{x},\mathbf{y}).$$

For simplicity, suppose that the symmetric part of the distortion tensor vanishes, $\tilde{B}_{(\mathbf{u})} = 0$. Then $B(\mathbf{u}) = \hat{r}_{(\mathbf{u})}$ and the connecting vector $\mathbf{c}$ satisfies the equation

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\mathbf{c} = B(\mathbf{u})\mathbf{c} = \hat{r}_{(\mathbf{u})}\mathbf{c},$$

with the solution

$$\mathbf{c}(\tau) = \exp\left(\tau\hat{r}_{(\mathbf{u})}\right)\mathbf{c}_0.$$

Since $r_{(\mathbf{u})}$ is transverse to a timelike direction, it is represented (in an orthogonal basis) by an antisymmetric 3×3 matrix acting in spacelike directions. The exponential of such a matrix is a *rotation* matrix, thus $\mathbf{c}(\tau)$ is related to the initial value $\mathbf{c}_0$ by a rotation. We find that *all* the vectors $\mathbf{c}$ are rotated by the same fixed angle, which is proportional to τ. It follows that the entire shape of the initial region is rigidly rotated around an axis. This motivates the name "rotation" for the tensor $r_{(\mathbf{u})}$.

2.3.3 Introducing Raychaudhuri equation

The Raychaudhuri equation describes the change of the divergence of a geodesic vector field $\mathbf{u}$ along its own orbits. We straightforwardly find

$$\nabla_{\mathbf{u}}(\mathrm{div}\mathbf{u}) = \mathrm{Tr}_{(\mathbf{x},\mathbf{y})}\left(\nabla_{\mathbf{u}}B_{(\mathbf{u})}\right)\circ(\mathbf{x},\mathbf{y});$$
$$\left(\nabla_{\mathbf{u}}B_{(\mathbf{u})}\right)\circ(\mathbf{x},\mathbf{y})$$
$$= \nabla_{\mathbf{u}}\left(B_{(\mathbf{u})}(\mathbf{x},\mathbf{y})\right) - B_{(\mathbf{u})}(\nabla_{\mathbf{u}}\mathbf{x},\mathbf{y}) - B_{(\mathbf{u})}(\mathbf{x},\nabla_{\mathbf{u}}\mathbf{y})$$
$$= \nabla_{\mathbf{u}}g(\nabla_{\mathbf{x}}\mathbf{u},\mathbf{y}) - g(\nabla_{\nabla_{\mathbf{u}}\mathbf{x}}\mathbf{u},\mathbf{y}) - g(\nabla_{\mathbf{x}}\mathbf{u},\nabla_{\mathbf{u}}\mathbf{y}).$$

Using the metricity condition and the definition of the Riemann tensor, we rewrite the first and the third terms as

$$\nabla_{\mathbf{u}}g(\nabla_{\mathbf{x}}\mathbf{u},\mathbf{y}) - g(\nabla_{\mathbf{x}}\mathbf{u},\nabla_{\mathbf{u}}\mathbf{y}) = g(\nabla_{\mathbf{u}}\nabla_{\mathbf{x}}\mathbf{u},\mathbf{y})$$
$$= R(\mathbf{u},\mathbf{x},\mathbf{u},\mathbf{y}) + g(\nabla_{\mathbf{x}}\nabla_{\mathbf{u}}\mathbf{u},\mathbf{y}) + g(\nabla_{[\mathbf{u},\mathbf{x}]}\mathbf{u},\mathbf{y})$$
$$= R(\mathbf{u},\mathbf{x},\mathbf{u},\mathbf{y}) + g(\nabla_{[\mathbf{u},\mathbf{x}]}\mathbf{u},\mathbf{y});$$

finally,

$$\left(\nabla_{\mathbf{u}}B_{(\mathbf{u})}\right)\circ(\mathbf{x},\mathbf{y})$$
$$= R(\mathbf{u},\mathbf{x},\mathbf{u},\mathbf{y}) + g(\nabla_{[\mathbf{u},\mathbf{x}]}\mathbf{u},\mathbf{y}) - g(\nabla_{\nabla_{\mathbf{u}}\mathbf{x}}\mathbf{u},\mathbf{y})$$
$$= R(\mathbf{u},\mathbf{x},\mathbf{u},\mathbf{y}) - g(\nabla_{\nabla_{\mathbf{x}}\mathbf{u}}\mathbf{u},\mathbf{y})$$
$$= R(\mathbf{u},\mathbf{x},\mathbf{u},\mathbf{y}) - g(B(\mathbf{u})B(\mathbf{u})\mathbf{x},\mathbf{y}).$$

The trace of the last expression with respect to $\mathbf{x}, \mathbf{y}$ yields

$$\nabla_{\mathbf{u}}(\mathrm{div}\mathbf{u}) = \mathrm{Ric}(\mathbf{u},\mathbf{u}) - \mathrm{Tr}\left(B(\mathbf{u})B(\mathbf{u})\right), \tag{2.14}$$

where the last term is the trace of the square of the linear transformation $B(\mathbf{u})$. This is the initial form of the **Raychaudhuri equation** which we shall later analyze in particular cases.

At this point we can use the decomposition (2.13) which separates the symmetric and the antisymmetric parts of the distortion tensor. Note that $\mathrm{Tr}\,(AB)$ is a nondegenerate scalar product in the linear space of square matrices. The subspaces of symmetric and antisymmetric matrices are orthogonal with respect to this scalar product. (Verify this!) Therefore with the decomposition $B = \tilde{B} + r$ the trace $\mathrm{Tr}(B^2)$ is simplified to

$$\mathrm{Tr}\left(B(\mathbf{u})B(\mathbf{u})\right) = \mathrm{Tr}\left(\tilde{B}(\mathbf{u})\tilde{B}(\mathbf{u})\right) + \mathrm{Tr}\left(\hat{r}_{(\mathbf{u})}\hat{r}_{(\mathbf{u})}\right),$$

where $\hat{r}_{(\mathbf{u})}$ is the transformation associated with the 2-form $r_{(\mathbf{u})}$. If the vector field $\mathbf{u}$ is integrable, its rotation vanishes and we are left only with the symmetric part $\tilde{B}$ of the distortion tensor.

A further simplification is possible if we separate the trace-free part from $\tilde{B}$. (This symmetric, trace-free part is called the **shear** of the field $\mathbf{u}$.) However, this operation depends sensitively on whether the field $\mathbf{u}$ is null. Therefore we shall consider the relevant cases separately.

2.3.4 Shear for timelike congruences

While the distortion tensor $B_{(\mathbf{u})}$ is not necessarily symmetric, it is always **transverse** to the direction of $\mathbf{v}$:

$$B_{(\mathbf{u})}(\mathbf{u},\mathbf{x}) = g(\nabla_{\mathbf{u}}\mathbf{u},\mathbf{x}) = 0,$$
$$B_{(\mathbf{u})}(\mathbf{x},\mathbf{u}) = g(\mathbf{u},\nabla_{\mathbf{x}}\mathbf{u}) = \frac{1}{2}\nabla_{\mathbf{x}}g(\mathbf{u},\mathbf{u}) = 0,$$

for all $\mathbf{x}$. This fact allows us to reduce $B_{(\mathbf{u})}$ to a bilinear form in a lower-dimensional space.

First we decompose the tangent space $T_p\mathcal{M}$ into the subspace $\mathbf{u}^\perp$ and its complement, the one-dimensional space spanned by $\mathbf{u}$. This decomposition is conveniently expressed using a projector P onto $\mathbf{u}^\perp$ and a complementary projector Q onto $\mathbf{u}$, such that

$$P + Q = \hat{1}, \quad \mathbf{x} = P\mathbf{x} + Q\mathbf{x} \text{ for all } \mathbf{x}, \quad P^2 = P,\ Q^2 = Q.$$

Note that $\mathrm{Tr}\,P = 3$ and $\mathrm{Tr}\,Q = 1$. The above **decomposition of identity** is equivalent to the decomposition of the metric g

into the partial metric $h_{\mathbf{u}}$, see Eq. (2.3), and a suitable complement,

$$g(\mathbf{x},\mathbf{y}) = h_{\mathbf{u}}(\mathbf{x},\mathbf{y}) + g(\mathbf{u},\mathbf{x})g(\mathbf{u},\mathbf{y}).$$

The contravariant metric is correspondingly decomposed as

$$g^{-1} = h_{\mathbf{u}}^{-1} + \mathbf{u}\otimes\mathbf{u}.$$

A bilinear form $B(\mathbf{x},\mathbf{y})$ can be decomposed with respect to the direction $\mathbf{u}$ using the projectors P and Q as follows:

$$B(\mathbf{x},\mathbf{y}) = B(P\mathbf{x},P\mathbf{y}) + B(P\mathbf{x},Q\mathbf{y}) + B(Q\mathbf{x},P\mathbf{y}) + B(Q\mathbf{x},Q\mathbf{y}). \tag{2.15}$$

So far we have not obtained any new information about B. However, if the form B is transverse to the direction of $\mathbf{u}$ then $B(\cdot,Q\mathbf{y}) = 0$, and the above formula simplifies to

$$B(\mathbf{x},\mathbf{y}) = B(P\mathbf{x},P\mathbf{y}).$$

Since the structure of the projectors crucially depends on whether the vector field $\mathbf{u}$ is null, we must consider the relevant cases separately. In this section we treat the easier case when $\mathbf{u}$ is timelike. (The spacelike case is very similar but less important in practice.) Assuming that $g(\mathbf{u},\mathbf{u}) = 1$, the projectors are

$$P\mathbf{x} = \mathbf{x} - \mathbf{u}g(\mathbf{x},\mathbf{u}), \quad Q\mathbf{x} = \mathbf{u}g(\mathbf{x},\mathbf{u}).$$

The distortion tensor is effectively three-dimensional since it is nonzero only on the subspace $\mathbf{u}^\perp$, where it coincides with its projection, denoted by $B^\perp_{(\mathbf{u})}$,

$$B^\perp_{(\mathbf{u})}(\mathbf{x},\mathbf{y}) \equiv B_{(\mathbf{u})}(P\mathbf{x},P\mathbf{y}) = B_{(\mathbf{u})}(\mathbf{x},\mathbf{y}).$$

The **shear** of the vector field $\mathbf{u}$ is, by definition, the symmetric traceless part of the projected distortion tensor. It is convenient to express the shear using the partial metric for the subspace $\mathbf{u}^\perp$,

$$h_{\mathbf{u}}(\mathbf{a},\mathbf{b}) \equiv g(P\mathbf{a},P\mathbf{b}) = g(\mathbf{a},P\mathbf{b}) = g(\mathbf{a},\mathbf{b}) - g(\mathbf{u},\mathbf{a})g(\mathbf{u},\mathbf{b}).$$

Note that the partial signature of this metric is $(---)$ and its trace is equal to 3. If a symmetric bilinear form $T(\mathbf{x},\mathbf{y})$ is nonzero only on the subspace $\mathbf{u}^\perp$, the traceless part of T is expressed as

$$T - \frac{1}{3}(\operatorname{Tr} T)\, h.$$

Since $\operatorname{Tr} B^\perp_{(\mathbf{u})} = \operatorname{Tr} B_{(\mathbf{u})} = \operatorname{div}\mathbf{u}$, we obtain the decomposition

$$B_{(\mathbf{u})} = \frac{1}{3}(\operatorname{div}\mathbf{u})\, h + \sigma_{(\mathbf{u})} + r_{(\mathbf{u})}, \tag{2.16}$$
$$\sigma_{(\mathbf{u})} \equiv \tilde{B}_{(\mathbf{u})} - \frac{1}{3}(\operatorname{div}\mathbf{u})h; \quad \operatorname{Tr}\hat{\sigma}_{(\mathbf{u})} = 0.$$

The name "shear" for the tensor $\sigma_{(\mathbf{u})}$ describes a linear deformation of the comoving region that changes its shape but does not change its volume.

The Raychaudhuri equation (2.14) contains the trace of the square of $B_{(\mathbf{u})}$, and we can now simplify this term by using the decomposition (2.16). Note that every term in that decomposition is transverse to $\mathbf{u}$ and thus is nonzero only on the three-dimensional subspace on which P projects. It is important that this subspace has a metric h with a Euclidean signature; the overall minus sign of h is unimportant for the present consideration. The partial metric h establishes a correspondence between bilinear forms and transformations, and we shall denote transformations by a hat. For example, the bilinear form $\sigma_{(\mathbf{u})}$ corresponds to the transformation $\hat{\sigma}_{(\mathbf{u})}$ such that[1]

$$\sigma_{(\mathbf{u})}(\mathbf{x},\mathbf{y}) = h(\hat{\sigma}_{(\mathbf{u})}\mathbf{x},\mathbf{y}).$$

This relation uniquely determines the vector $\hat{\sigma}_{(\mathbf{u})}\mathbf{x}$ within the subspace $\mathbf{u}^\perp$.

In the space of bilinear forms with the scalar product $\operatorname{Tr}(AB)$, the subspace of symmetric traceless forms is orthogonal to the subspace of "pure trace" (i.e. bilinear forms proportional to the partial metric h). Both these subspaces are also orthogonal to the subspace of antisymmetric forms. Therefore we can decompose the trace term as

$$\operatorname{Tr}(B(\mathbf{u})B(\mathbf{u})) = \frac{1}{3}(\operatorname{Tr} B(\mathbf{u}))^2 + \operatorname{Tr}\left(\hat{\sigma}_{(\mathbf{u})}\hat{\sigma}_{(\mathbf{u})}\right) + \operatorname{Tr}\left(\hat{r}_{(\mathbf{u})}\hat{r}_{(\mathbf{u})}\right).$$

The Raychaudhuri equation for timelike geodesics is then written as

$$\nabla_{\mathbf{u}}(\operatorname{div}\mathbf{u}) = \operatorname{Ric}(\mathbf{u},\mathbf{u}) - \frac{1}{3}(\operatorname{div}\mathbf{u})^2 - \operatorname{Tr}\left(\hat{\sigma}_{(\mathbf{u})}\hat{\sigma}_{(\mathbf{u})}\right) - \operatorname{Tr}\left(\hat{r}_{(\mathbf{u})}\hat{r}_{(\mathbf{u})}\right).$$

Let us now analyze the individual terms in the above equation. Since a trace of a square of a symmetric matrix is nonnegative, we have $\operatorname{Tr}(\hat{\sigma}_{(\mathbf{u})}\hat{\sigma}_{(\mathbf{u})}) \geq 0$. We can demonstrate this fact formally by computing

$$\operatorname{Tr}_{(\mathbf{x},\mathbf{y})} h(\hat{\sigma}\hat{\sigma}\mathbf{x},\mathbf{y}) \equiv \operatorname{Tr}_{(\mathbf{x},\mathbf{y})}\sigma(\hat{\sigma}\mathbf{x},\mathbf{y}) = \operatorname{Tr}_{(\mathbf{x},\mathbf{y})}\sigma(\mathbf{y},\hat{\sigma}\mathbf{x}) \equiv \operatorname{Tr}_{(\mathbf{x},\mathbf{y})} h(\hat{\sigma}\mathbf{y},\hat{\sigma}\mathbf{x}).$$

We now need to use a metric decomposition (see Sec. 1.7.3 for details of trace calculations). If $\{\mathbf{s}_1,\mathbf{s}_2,\mathbf{s}_3\}$ is an orthonormal basis of (spacelike) vectors, so that $h(\mathbf{s}_i,\mathbf{s}_j) = -\delta_{ij}$, then the inverse metric h^{-1} is decomposed as

$$h^{-1} = -\mathbf{s}_1\otimes\mathbf{s}_1 - \mathbf{s}_2\otimes\mathbf{s}_2 - \mathbf{s}_3\otimes\mathbf{s}_3,$$

and we get

$$\operatorname{Tr}_{(\mathbf{x},\mathbf{y})} h(\hat{\sigma}\mathbf{y},\hat{\sigma}\mathbf{x}) = -\sum_{j=1}^{3} h(\hat{\sigma}\mathbf{s}_j,\hat{\sigma}\mathbf{s}_j) \geq 0$$

since h is negative-definite.

Statement: We have $\operatorname{Tr}(\hat{r}_{(\mathbf{u})}\hat{r}_{(\mathbf{u})}) \leq 0$.

Derivation: Same calculation as above, except for

$$\operatorname{Tr}_{(\mathbf{x},\mathbf{y})} r_{(\mathbf{u})}(\hat{r}_{(\mathbf{u})}\mathbf{x},\mathbf{y}) = -\operatorname{Tr}_{(\mathbf{x},\mathbf{y})} r_{(\mathbf{u})}(\mathbf{y},\hat{r}_{(\mathbf{u})}\mathbf{x}).$$

2.3.5 Shear for null congruences

We have seen that for null vector fields $\mathbf{n}$, building a self-adjoint projector (2.2) into $\mathbf{n}^\perp$ requires a choice of an additional null vector field $\mathbf{l}$. It is convenient to normalize $\mathbf{l}$ by $g(\mathbf{l},\mathbf{n}) = 1$. The definition (2.2) of the projector can be interpreted as a decomposition of the identity operator into the projector P and a complementary projector Q,

$$\hat{1} = P + Q, \quad Q\mathbf{x} \equiv \mathbf{l}g(\mathbf{n},\mathbf{x}) + \mathbf{n}g(\mathbf{l},\mathbf{x}).$$

[1] **Why don't we also denote the transformation $B(\mathbf{u})$ by a hat?**

Note that Tr $P = 2$ and Tr $Q = 2$. The above formula is equivalent to a decomposition of the metric g into the partial metric $h_{\mathbf{n}}$, given by Eq. (2.4), and a suitable complement,

$$g(\mathbf{x},\mathbf{y}) = h_{\mathbf{n}}(\mathbf{x},\mathbf{y}) + g(\mathbf{l},\mathbf{x})g(\mathbf{n},\mathbf{y}) + g(\mathbf{n},\mathbf{x})g(\mathbf{l},\mathbf{y}).$$

Note that a tensor decomposition of the inverse metric g^{-1} is

$$g^{-1} = \mathbf{l}\otimes\mathbf{n} + \mathbf{n}\otimes\mathbf{l} - \mathbf{s}_1\otimes\mathbf{s}_1 - \mathbf{s}_2\otimes\mathbf{s}_2, \tag{2.17}$$

where $\mathbf{s}_1$ and $\mathbf{s}_2$ are orthonormal, spacelike vectors, orthogonal to both $\mathbf{n}$ and $\mathbf{l}$, and spanning the image of P.

Let us now decompose the distortion tensor $B_{(\mathbf{n})}$ using the projectors P and Q, analogously to Eq. (2.15). Since $B_{(\mathbf{n})}(\mathbf{x},\mathbf{n}) = 0$ by transversality, we can write

$$\begin{aligned} B_{(\mathbf{n})}(\mathbf{x},\mathbf{y}) &= B_{(\mathbf{n})}(P\mathbf{x},P\mathbf{y}) + B_{(\mathbf{n})}(P\mathbf{x},\mathbf{l})g(\mathbf{n},\mathbf{y}) \\ &\quad + B_{(\mathbf{n})}(\mathbf{l},P\mathbf{y})g(\mathbf{n},\mathbf{x}) + B_{(\mathbf{n})}(\mathbf{l},\mathbf{l})g(\mathbf{n},\mathbf{x})g(\mathbf{n},\mathbf{y}). \end{aligned}$$

The first term, $B_{(\mathbf{n})}(P\mathbf{x},P\mathbf{y}) \equiv B^{\perp}_{(\mathbf{n})}(\mathbf{x},\mathbf{y})$, is effectively a bilinear form on a two-dimensional space spanned by $\mathbf{s}_1,\mathbf{s}_2$. Note that the trace of $B^{\perp}_{(\mathbf{n})}$ is equal to the trace of $B_{(\mathbf{n})}$. This can be seen formally by using the general property

$$\mathrm{Tr}_{(\mathbf{x},\mathbf{y})}g(\mathbf{x},\mathbf{n})A(\mathbf{y},...) = A(\mathbf{n},...);$$

in other words, a trace with $g(\mathbf{n},\mathbf{x})$ substitutes $\mathbf{n}$ into the rest of the expression. However, substituting $\mathbf{n}$ into any term yields zero since $P\mathbf{n} = 0$ and $g(\mathbf{n},\mathbf{n}) = 0$. For instance,

$$\begin{aligned} \mathrm{Tr}_{(\mathbf{x},\mathbf{y})}g(\mathbf{x},\mathbf{n})g(\mathbf{n},\mathbf{y}) &= g(\mathbf{n},\mathbf{n}) = 0, \\ \mathrm{Tr}_{(\mathbf{x},\mathbf{y})}B_{(\mathbf{n})}(P\mathbf{x},\mathbf{l})g(\mathbf{n},\mathbf{y}) &= B_{(\mathbf{n})}(P\mathbf{n},\mathbf{l}) = 0. \end{aligned}$$

Therefore

$$\mathrm{Tr}\, B^{\perp}_{(\mathbf{n})} = \mathrm{div}\mathbf{n}.$$

When we compute the trace needed for the Raychaudhuri equation,

$$\mathrm{Tr}\left(B(\mathbf{n})B(\mathbf{n})\right) \equiv \mathrm{Tr}_{(\mathbf{a},\mathbf{b})(\mathbf{x},\mathbf{y})}B_{(\mathbf{n})}(\mathbf{a},\mathbf{x})\,B_{(\mathbf{n})}(\mathbf{y},\mathbf{b}),$$

it turns out that all the terms vanish except the term Tr $B^{\perp}_{(\mathbf{n})}B^{\perp}_{(\mathbf{n})}$ involving only the "projected" or "transverse" distortion $B^{\perp}_{(\mathbf{n})}$,

$$\mathrm{Tr}\left(B(\mathbf{n})B(\mathbf{n})\right) = \mathrm{Tr}\, B^{\perp}_{(\mathbf{n})}B^{\perp}_{(\mathbf{n})}. \tag{2.18}$$

(Verifying this is left as an easy exercise.) Therefore all the other terms in the decomposition of $B_{(\mathbf{n})}$ are "unphysical," i.e. they do not enter the physically significant equations.

Remark: For null congruences $\mathbf{n}$, the reduction of the distortion tensor $B_{(\mathbf{n})}$ to its transverse $B^{\perp}_{(\mathbf{n})}$ can be also motivated as follows. The distortion tensor describes the change of shape in the comoving 2-volume spanned by two spacelike connecting vectors $\mathbf{s}_{1,2}$. It is clear that these connecting vectors should be orthogonal to $\mathbf{n}$. Also, connecting vectors could be changed by adding a multiple of $\mathbf{n}$, e.g. $\mathbf{s}_1 \to \mathbf{s}_1 + \lambda\mathbf{n}$, where λ is a scalar function; this will not modify the orthogonality of $\mathbf{s}_{1,2}$ and $\mathbf{n}$. However, connecting vectors that differ only by a multiple of $\mathbf{n}$ correspond to a change of the parameter in nearby curves of the congruence. Therefore, the connecting vectors $\mathbf{s}$ and $\mathbf{s}+\lambda\mathbf{n}$ carry the same information about the geometry of the congruence. Thus we should only consider equivalence classes of connecting vectors up to a multiple of $\mathbf{n}$. While there is no canonical choice of representatives for these equivalence classes, a relatively straightforward way is to choose $\mathbf{s}_{1,2}$ within a 2-plane orthogonal to $\mathbf{n}$ and an auxiliary null vector $\mathbf{l}$. This is the construction used in the present section.

We can now separate the symmetric traceless part $\sigma_{(\mathbf{n})}$ of the projected distortion, called the **shear** of $\mathbf{n}$, from the "pure trace" part proportional to $h_{\mathbf{n}}$, and from the antisymmetric part, which is the projected rotation $r^{\perp}_{(\mathbf{n})}$,

$$\begin{aligned} B^{\perp}_{(\mathbf{n})} &= \frac{1}{2}(\mathrm{div}\mathbf{n})h_{\mathbf{n}} + \sigma_{(\mathbf{n})} + r^{\perp}_{(\mathbf{n})}, \\ \sigma_{(\mathbf{n})}(\mathbf{x},\mathbf{y}) &\equiv \frac{1}{2}\left(B^{\perp}_{(\mathbf{n})}(\mathbf{x},\mathbf{y}) + B^{\perp}_{(\mathbf{n})}(\mathbf{y},\mathbf{x})\right) - \frac{1}{2}(\mathrm{div}\mathbf{n})h_{\mathbf{n}}, \\ r^{\perp}_{(\mathbf{n})}(\mathbf{x},\mathbf{y}) &\equiv r_{(\mathbf{n})}(P\mathbf{x},P\mathbf{y}) = \frac{1}{2}\left(B^{\perp}_{(\mathbf{n})}(\mathbf{x},\mathbf{y}) - B^{\perp}_{(\mathbf{n})}(\mathbf{y},\mathbf{x})\right). \end{aligned}$$

Note the factor $\frac{1}{2}$ instead of $\frac{1}{3}$, which is due to the projector P being two-dimensional. The Raychaudhuri equation for null geodesics is then written as

$$\begin{aligned} \nabla_{\mathbf{n}}(\mathrm{div}\mathbf{n}) &= \mathrm{Ric}(\mathbf{n},\mathbf{n}) - \frac{1}{2}(\mathrm{div}\mathbf{n})^2 \\ &\quad - \mathrm{Tr}\left(\hat{\sigma}_{(\mathbf{n})}\hat{\sigma}_{(\mathbf{n})}\right) - \mathrm{Tr}\left(\hat{r}^{\perp}_{(\mathbf{n})}\hat{r}^{\perp}_{(\mathbf{n})}\right). \end{aligned}$$

2.4 Applications of Raychaudhuri equation

2.4.1 Energy conditions

The Einstein equation (1.70) specifies the matter energy-momentum tensor $T_{\mu\nu}$ that would produce any given spacetime. However, not every tensor $T_{\mu\nu}$ corresponds to physically reasonable situations. There are certain conditions that are convenient to impose on $T_{\mu\nu}$; these are summarily called **energy conditions**.

It is reasonable to suppose that the energy density measured locally by an observer is everywhere nonnegative.[2] This gives the **weak energy condition**

$$T_{\mu\nu}u^\mu u^\nu \equiv T(\mathbf{u},\mathbf{u}) \geq 0 \text{ for all timelike } \mathbf{u}.$$

Since null vectors are limits of timelike vectors, a consequence of this condition is the **null energy condition**

$$T(\mathbf{n},\mathbf{n}) \geq 0 \text{ for all null } \mathbf{n}.$$

In some calculations it is necessary to impose conditions directly on the Ricci tensor. One such condition is the **strong energy condition**,

$$\mathrm{Ric}(\mathbf{u},\mathbf{u}) \leq 0 \text{ for all timelike } \mathbf{u}.$$

By the Einstein equation, this is equivalent to

$$T(\mathbf{u},\mathbf{u}) - \frac{g(\mathbf{u},\mathbf{u})}{2}\mathrm{Tr}\,T \geq 0 \text{ for all timelike } \mathbf{u}.$$

Note that (contrary to what the words suggest) the strong energy condition *does not* imply the weak one.

Finally, the **dominant energy condition** is

$$j_\mu \equiv T_{\mu\nu}u^\mu \text{ is future-directed and } j_\mu j^\mu \geq 0,$$

[2] This is true for every known classical field but is sometimes violated by quantum fields.

for all timelike u^μ. This condition means that the energy-momentum flows along future-directed timelike or null lines.

By definition, a **perfect fluid** with the velocity field $\mathbf{v}$, density ρ, and pressure p is described by the energy-momentum tensor

$$T = (\rho + p)\mathbf{v} \otimes \mathbf{v} - pg^{-1};$$
$$T_{\mu\nu} = (\rho + p)\, v_\mu v_\nu - p g_{\mu\nu}.$$

In many situations, the matter EMT can be represented by the perfect fluid form with a suitably chosen "velocity" vector $\mathbf{v}$. It is assumed that $\mathbf{v}$ is timelike and $g(\mathbf{v}, \mathbf{v}) = 1$. Note that the trace of the perfect fluid EMT is

$$\operatorname{Tr} T = \rho - 3p.$$

Energy conditions can be expressed as inequalities for p and ρ as follows: The weak energy condition is

$$\rho \geq 0, \quad \rho + p \geq 0,$$

the null energy condition is just

$$\rho + p \geq 0,$$

the strong energy condition is

$$\rho + 3p \geq 0, \quad \rho + p \geq 0,$$

and the dominant energy condition is

$$\rho \geq 0, \quad \rho \geq |p|\,.$$

Statement: The above formulas represent the energy conditions for the perfect fluid EMT.

Derivation: Since the energy conditions involve an arbitrary timelike vector $\mathbf{u}$, it is convenient to decompose that vector with respect to the fluid velocity $\mathbf{v}$ as

$$\mathbf{u} = a\mathbf{v} + \mathbf{w}, \quad \mathbf{w} \in \mathbf{v}^\perp,\ g(\mathbf{u}, \mathbf{u}) = a^2 - g(\mathbf{w}, \mathbf{w}) = 1,$$

where a is a suitable constant. The value of the vector $\mathbf{w}$ will be irrelevant for the calculations since $T(\mathbf{v}, \mathbf{v}) = p + \rho$, $T(\mathbf{v}, \mathbf{w}) = 0$, and $T(\mathbf{w}, \mathbf{w}) = -p$, while the value of a is constrained only by $|a| \geq 1$. For instance, the weak energy condition gives

$$T(\mathbf{u}, \mathbf{u}) = a^2 T(\mathbf{v}, \mathbf{v}) + T(\mathbf{w}, \mathbf{w}) = a^2(\rho + p) - p \geq 0,$$

which holds for all $a \geq 1$ only if $\rho \geq 0$ (at $a = 1$) and $\rho + p \geq 0$ (the limit of large a). Other energy conditions are treated similarly.

Statement: A massless, minimally coupled scalar field ϕ with the EMT

$$T = \mathrm{d}\phi \otimes \mathrm{d}\phi - \frac{g^{-1}(\mathrm{d}\phi, \mathrm{d}\phi)}{2} g,$$
$$T_{\mu\nu} \equiv \phi_{,\mu}\phi_{,\nu} - \frac{1}{2} g_{\mu\nu} \phi_{,\alpha}\phi^{,\alpha},$$

satisfies the null and the strong energy conditions.

Derivation: For any null vector $\mathbf{n}$, we have

$$T(\mathbf{n}, \mathbf{n}) = (\nabla_{\mathbf{n}}\phi)^2 - \frac{g(\mathbf{n}, \mathbf{n})}{2}(\ldots) = (\nabla_{\mathbf{n}}\phi)^2 \geq 0.$$

For any timelike vector $\mathbf{v}$, we have

$$\operatorname{Tr} T = -g^{-1}(\mathrm{d}\phi, \mathrm{d}\phi),$$
$$T(\mathbf{v}, \mathbf{v}) - \frac{1}{2} g(\mathbf{v}, \mathbf{v}) \operatorname{Tr} T = (\nabla_{\mathbf{v}}\phi)^2 \geq 0.$$

2.4.2 Focusing of timelike geodesics

The divergence $\operatorname{div}\mathbf{u}$ of a timelike vector field $\mathbf{u}$ can be interpreted as the relative rate of change of a comoving 3-volume orthogonal to $\mathbf{u}$. This is so because the 3-volume spanned by vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$ is equal to $\varepsilon(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{x})$, where $\mathbf{x}$ is any normalized vector orthogonal to $\mathbf{a}, \mathbf{b}, \mathbf{c}$. If $\mathbf{a}, \mathbf{b}, \mathbf{c}$ are connecting vectors for $\mathbf{u}$ which are orthogonal to $\mathbf{u}$ initially, they will remain orthogonal to $\mathbf{u}$, so we can use $\mathbf{u}$ as the vector $\mathbf{x}$. Thus

$$\frac{dV}{d\tau} = \mathcal{L}_{\mathbf{u}} \varepsilon(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{u}) = (\operatorname{div}\mathbf{u})\, \varepsilon(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{u}) = (\operatorname{div}\mathbf{u})\, V.$$

Consider a congruence of timelike geodesics such that the tangent vector field $\mathbf{u}$ is hypersurface orthogonal. Then, as we have seen, the rotation of $\mathbf{u}$ vanishes, and the distortion tensor is symmetric. The Raychaudhuri equation is simplified to

$$\frac{d}{d\tau}(\operatorname{div}\mathbf{u}) = R(\mathbf{u}, \mathbf{u}) - \frac{1}{3}(\operatorname{div}\mathbf{u})^2 - \operatorname{Tr}\left(\hat{\sigma}_{(\mathbf{u})}\hat{\sigma}_{(\mathbf{u})}\right).$$

Imposing the strong energy condition, which holds for classical matter, we find that the right-hand side of the above equation is nonpositive, indicating that gravity makes hypersurface-orthogonal geodesics diverge less. This means that gravity is an attractive force. In fact, we can obtain a stronger result,

$$\frac{d}{d\tau}(\operatorname{div}\mathbf{u}) \leq -\frac{1}{3}(\operatorname{div}\mathbf{u})^2.$$

This inequality can be integrated with an initial value $\operatorname{div}\mathbf{u}(\tau_0) = \theta_0$, which yields

$$\operatorname{div}\mathbf{u} \leq \frac{1}{\theta_0^{-1} + \frac{1}{3}\tau}.$$

If the initial value θ_0 is positive, the divergence goes to zero (or becomes negative), and if it is initially negative, it diverges to negative infinity within a finite time. A negative infinite divergence means that the geodesics cross at a **focal point** . The set of all focal points in a given congruence is called a **caustic**. We conclude that the appearance of a focal point is inevitable within a finite proper time once the initial value of the divergence is negative, for hypersurface orthogonal geodesic congruences. This statement is known as the **focusing theorem**.

Here is an example of a hypersurface orthogonal congruence to which the focusing theorem applies.

Statement 1: A congruence of timelike geodesics emitted from one point p is hypersurface orthogonal because it is orthogonal to hypersurfaces of constant proper time τ.

Proof: Let $\mathbf{u}$ be the affine tangent vector field for a congruence of geodesics emitted from p. We may normalize $g(\mathbf{u}, \mathbf{u}) = 1$. Since there are three linearly independent connecting vectors for $\mathbf{u}$ at any point, and since these connecting vectors form a basis of the tangent space to a hypersurface of constant τ, it is sufficient to show that *any* connecting vector field $\mathbf{c}$ is everywhere orthogonal to $\mathbf{u}$. It will then follow that $\mathbf{u}$ is orthogonal to the hypersurface of constant τ. Now, if $\mathbf{c}$ is a connecting vector for $\mathbf{u}$, it is easy to show that $\nabla_{\mathbf{u}} g(\mathbf{c}, \mathbf{u}) = 0$ (see Lemma 1 in Sec. 2.2.6). Thus $g(\mathbf{c}, \mathbf{u})$ is constant along the lines of the congruence, and so it remains to show that $g(\mathbf{c}, \mathbf{u}) = 0$ at any one point along each line. Since obviously all connecting vectors $\mathbf{c}$ must vanish at the point p, we have $g(\mathbf{c}, \mathbf{u}) = 0$ at p for every line of $\mathbf{u}$. (Alternatively, we may

consider an infinitesimal neighborhood of p, where the geometry of lines is approximately the same as in the Minkowski space, where geodesics are straight lines and the statement $g(\mathbf{c}, \mathbf{u}) = 0$ is obviously true.) Thus we have proved the statement.

2.4.3 Repulsive gravity

Suppose that the energy-momentum tensor contains a **cosmological constant** term

$$T = T_{(m)} + \frac{\Lambda}{8\pi} g^{-1},$$

where $T_{(m)}$ is the matter EMT and $\Lambda > 0$ is the cosmological constant. (The cosmological constant is also called **dark energy** if it is implied that $\Lambda = \Lambda(p)$ is a slowly-varying dynamical field and its local value near a point p is perceived as a cosmological "constant".) The effect of introducing a cosmological constant is described by the Raychaudhuri equation, which receives an extra term

$$\frac{d}{d\tau}(\mathrm{div}\mathbf{u}) = -\frac{1}{8\pi}\left[T_{(m)}(\mathbf{u},\mathbf{u}) - \frac{1}{2}\mathrm{Tr}\, T_{(m)}\right] - \frac{1}{3}(\mathrm{div}\mathbf{u})^2 - \mathrm{Tr}(\hat{\sigma}_{(\mathbf{u})}\hat{\sigma}_{(\mathbf{u})}) + \Lambda.$$

The extra term will decrease the rate of focusing, which indicates that a cosmological constant corresponds to a *repulsive* gravitational force.

2.4.4 Focusing of null geodesics

We now consider a congruence of null geodesics with a hypersurface-orthogonal tangent vector field $\mathbf{n}$, and we would like to obtain an analogous focusing theorem. Compared with the timelike case, we have two differences: Firstly, the 3-volume spanned by vectors orthogonal to $\mathbf{n}$ vanishes and so the interpretation of div$\mathbf{n}$ is altered; this issue was treated in Sec. 2.1.2 where we found that div$\mathbf{n}$ describes the rate of change of comoving area. Secondly, the rotation tensor $r_{(\mathbf{n})}$ generally does not vanish for hypersurface-orthogonal null geodesics. However, Raychaudhuri equation involves the trace term $\mathrm{Tr}\,(r_{(\mathbf{n})}r_{(\mathbf{n})})$, and we shall now show that this term vanishes for a hypersurface-orthogonal field $\mathbf{n}$, even though the rotation tensor $r_{(\mathbf{n})}$ itself does not vanish.

We have seen in Sec. 2.3.5 that the trace terms involve only the projected distortion tensor [Eq. (2.18)]. Therefore the trace term will remain the same after we project the rotation tensor onto the subspace spanned by $\mathbf{s}_{1,2}$,

$$\mathrm{Tr}\,\left(r_{(\mathbf{n})}r_{(\mathbf{n})}\right) = \mathrm{Tr}\,\left(r^{\perp}_{(\mathbf{n})}r^{\perp}_{(\mathbf{n})}\right).$$

We now use the fact that the rotation 2-form $r_{(\mathbf{n})}$ is equal to $d\omega$, where $\omega \equiv \hat{g}^{-1}\mathbf{n}$ is the 1-form dual to $\mathbf{n}$. By the assumption of hypersurface orthogonality and the Frobenius theorem, there exist scalar functions λ, ν such that $\omega = e^{\lambda}\mathrm{d}\nu$ and then

$$\mathrm{d}\omega = e^{\lambda}\mathrm{d}\lambda \wedge \mathrm{d}\nu = \mathrm{d}\lambda \wedge \omega.$$

The projected rotation tensor acts on vectors $\mathbf{x}, \mathbf{y}$ as

$$\begin{aligned} r^{\perp}_{(\mathbf{n})}(\mathbf{x},\mathbf{y}) &\equiv (\mathrm{d}\omega)\circ(P\mathbf{x}, P\mathbf{y}) \\ &= \omega(P\mathbf{y})(\mathrm{d}\lambda)\circ P\mathbf{x} - \omega(P\mathbf{x})(\mathrm{d}\lambda)\circ P\mathbf{y}. \end{aligned}$$

Since $\omega(P\mathbf{x}) = 0$ for all $\mathbf{x}$, we have $r^{\perp}_{(\mathbf{n})} = 0$.

Since the projected rotation vanishes, we obtain the Raychaudhuri equation for null geodesics in the form

$$\frac{d}{d\tau}(\mathrm{div}\mathbf{n}) = R(\mathbf{n},\mathbf{n}) - \frac{1}{2}(\mathrm{div}\mathbf{n})^2 - \mathrm{Tr}\left(\hat{\sigma}_{(\mathbf{n})}\hat{\sigma}_{(\mathbf{n})}\right).$$

Imposing the null energy condition, we find $R(\mathbf{n},\mathbf{n}) \leq 0$. Hence, the right-hand side of the above equation is nonpositive and

$$\frac{d}{d\tau}(\mathrm{div}\mathbf{n}) \leq -\frac{1}{2}(\mathrm{div}\mathbf{n})^2.$$

This is the focusing theorem for the null case. The conclusion is that gravity makes null geodesics focus within a finite interval of the affine parameter τ.

2.5 Null tetrad formalism

The null tetrad formalism is due to Newman and Penrose and consists of writing equations in a basis where *all four* vectors are null and "as orthogonal as possible" with respect to each other. This leads to simplifications when writing tensor equations in components. We shall now introduce the null tetrad formalism as an elegant way of expressing the projected distortion tensor for the case of null geodesics.

The idea is to change the previously used basis $\{\mathbf{l}, \mathbf{n}, \mathbf{s}_1, \mathbf{s}_2\}$, where the vectors $\mathbf{l}, \mathbf{n}$ are null and $\mathbf{s}_{1,2}$ are spacelike connecting vectors for $\mathbf{n}$, into a null basis. Since $\mathbf{l}$ must satisfy $g(\mathbf{l},\mathbf{n}) = 1$ and so $\mathbf{l}$ is not (and generally cannot be chosen as) a connecting vector for $\mathbf{n}$, we are motivated to drop the requirement that $\mathbf{s}_{1,2}$ be connecting vectors but instead demand orthonormality, $g(\mathbf{s}_j, \mathbf{s}_k) = -\delta_{jk}$. Then we define the *complex-valued* null vectors

$$\mathbf{m} \equiv \frac{1}{\sqrt{2}}(\mathbf{s}_1 + \mathrm{i}\mathbf{s}_2), \quad \bar{\mathbf{m}} \equiv \frac{1}{\sqrt{2}}(\mathbf{s}_1 - \mathrm{i}\mathbf{s}_2). \tag{2.19}$$

The fact that these vectors are complex-valued is merely a mathematical formality that makes equations more compact; we expect that every physically measurable quantity is real-valued. The null basis $\{\mathbf{l}, \mathbf{n}, \mathbf{m}, \bar{\mathbf{m}}\}$ is called the **Newman-Penrose null tetrad**.

The null tetrad is "nearly orthogonal" since $g(\mathbf{m}, \bar{\mathbf{m}}) = -1$, $g(\mathbf{l},\mathbf{n}) = 1$, and all the other scalar products vanish. Hence, the decomposition (2.17) of the contravariant metric g^{-1} is rewritten through the null tetrad as

$$g^{-1} = \mathbf{l}\otimes\mathbf{n} + \mathbf{n}\otimes\mathbf{l} - \mathbf{m}\otimes\bar{\mathbf{m}} - \bar{\mathbf{m}}\otimes\mathbf{m}.$$

The partial metric $h_{(\mathbf{n})}$ is similarly decomposed as

$$h^{-1}_{(\mathbf{n})} = -\mathbf{m}\otimes\bar{\mathbf{m}} - \bar{\mathbf{m}}\otimes\mathbf{m},$$

and the projector P is

$$P\mathbf{x} = -\mathbf{m}g(\mathbf{x},\bar{\mathbf{m}}) - \bar{\mathbf{m}}g(\mathbf{x},\mathbf{m}).$$

Using the null tetrad decomposition above, the divergence of a hypersurface orthogonal null geodesic field $\mathbf{n}$ can be written as

$$\begin{aligned} \mathrm{div}\mathbf{n} &= \mathrm{Tr}_{(\mathbf{x},\mathbf{y})}B_{(\mathbf{n})}(\mathbf{x},\mathbf{y}) \\ &= B_{(\mathbf{n})}(\mathbf{l},\mathbf{n}) + B_{(\mathbf{n})}(\mathbf{n},\mathbf{l}) - B_{(\mathbf{n})}(\mathbf{m},\bar{\mathbf{m}}) - B_{(\mathbf{n})}(\bar{\mathbf{m}},\mathbf{m}) \\ &= -2B_{(\mathbf{n})}(\mathbf{m},\bar{\mathbf{m}}) \equiv -2g(\nabla_{\mathbf{m}}\mathbf{n},\bar{\mathbf{m}}), \end{aligned}$$

because the tensor $B_{(\mathbf{n})}$ is transverse to $\mathbf{n}$ and its projection onto the subspace spanned by $(\mathbf{m}, \bar{\mathbf{m}})$ is rotation-free.

Note that for a geodesic field $\mathbf{n}$ that is *not* hypersurface orthogonal, the quantity $-2B_{(\mathbf{n})}(\mathbf{m}, \bar{\mathbf{m}})$ is complex-valued, and only its real part is actually equal to the divergence of $\mathbf{n}$. However, the "complex-valued divergence"

$$\theta \equiv -2B_{(\mathbf{n})}(\mathbf{m}, \bar{\mathbf{m}})$$

is still useful for the analysis of such cases.

The other ingredient in the Raychaudhuri equation is the projected shear tensor $\sigma^{\perp}_{(\mathbf{n})}$, which can be expressed as

$$\begin{aligned}\sigma^{\perp}_{(\mathbf{n})}(\mathbf{x}, \mathbf{y}) &= B_{(\mathbf{n})}(P\mathbf{x}, P\mathbf{y}) - \frac{1}{2}\,(\mathrm{div}\mathbf{n})\, h_{(\mathbf{n})}(\mathbf{x}, \mathbf{y}) \\ &= g(\bar{\mathbf{m}}, \mathbf{x})g(\bar{\mathbf{m}}, \mathbf{y})\sigma + g(\mathbf{m}, \mathbf{x})g(\mathbf{m}, \mathbf{y})\bar{\sigma} \qquad (2.20)\end{aligned}$$

through the auxiliary (complex-valued) scalar quantity σ,

$$\sigma \equiv B_{(\mathbf{n})}(\mathbf{m}, \mathbf{m}),$$

also called the **scalar shear**. Another expression for the scalar shear is

$$\sigma \equiv g(\mathbf{m}, \nabla_{\mathbf{m}}\mathbf{n}) = g(\mathbf{m}, [\mathbf{m}, \mathbf{n}]) + g(\mathbf{m}, \nabla_{\mathbf{n}}\mathbf{m}) = g(\mathbf{m}, [\mathbf{m}, \mathbf{n}]).$$

It follows that $\mathbf{m}$ cannot be a connecting vector for $\mathbf{n}$ unless the shear vanishes.

The completion of a given vector $\mathbf{n}$ to a null tetrad is of course not unique. Let us now investigate the freedom in the choice of the vectors $\mathbf{l}, \mathbf{m}$ and the resulting uncertainties in the distortion quantities σ, θ. The tetrad vectors are defined by the requirements $g(\mathbf{l}, \mathbf{l}) = g(\mathbf{m}, \mathbf{m}) = g(\mathbf{l}, \mathbf{m}) = 0$, $g(\mathbf{n}, \mathbf{l}) = -g(\mathbf{m}, \bar{\mathbf{m}}) = 1$. The vector $\mathbf{m}$ is related by Eq. (2.19) to an orthonormal basis $\mathbf{s}_{1,2}$ in the space $(\mathbf{l}, \mathbf{n})^{\perp}$. Note that $\{\mathbf{l}, \mathbf{n}, \mathbf{s}_1, \mathbf{s}_2\}$ is a basis; the set of transformations that preserve the orthonormality of the basis is the Lorentz group; so the admissible transformations of the tetrad vectors are precisely the subgroup of the Lorentz group that leaves the vector $\mathbf{n}$ unchanged. Let us describe these transformations explicitly. Suppose that the vectors $\mathbf{m}$ and $\mathbf{l}$ change to

$$\begin{aligned}\mathbf{m} \to \tilde{\mathbf{m}} &= \alpha\mathbf{m} + \beta\mathbf{n} + \gamma\mathbf{l}, \\ \mathbf{l} \to \tilde{\mathbf{l}} &= \lambda\mathbf{l} + \mu\mathbf{m} + \bar{\mu}\bar{\mathbf{m}} + \nu\mathbf{n},\end{aligned}$$

where $\alpha, \beta, \gamma, \mu$ are complex and λ, ν are real. Then it is straightforward to prove that the orthogonality relations $g(\tilde{\mathbf{m}}, \mathbf{n}) = 0$, $g(\tilde{\mathbf{m}}, \tilde{\mathbf{m}}) = -1$, $g(\tilde{\mathbf{l}}, \mathbf{n}) = 1$, $g(\tilde{\mathbf{l}}, \tilde{\mathbf{m}}) = 0$, and $g(\tilde{\mathbf{l}}, \tilde{\mathbf{l}}) = 0$ require that $\gamma = 0$, $\alpha\bar{\alpha} = 1$, $\lambda = 1$, $\beta = \bar{\mu}\alpha$, and $\nu = \mu\bar{\mu}$. Therefore, the general form of an admissible transformation is

$$\mathbf{m} \to \tilde{\mathbf{m}} = e^{\mathrm{i}\phi}\mathbf{m} + A\mathbf{n}, \qquad (2.21)$$

$$\mathbf{l} \to \tilde{\mathbf{l}} = \mathbf{l} + e^{\mathrm{i}\phi}\bar{A}\mathbf{m} + e^{-\mathrm{i}\phi}A\bar{\mathbf{m}} + A\bar{A}\mathbf{n}, \qquad (2.22)$$

where ϕ and A are arbitrary scalar functions; note that ϕ is real-valued while A is complex-valued. We can easily see that the distortion quantities change under this transformation as

$$\sigma \to \tilde{\sigma} = e^{2\mathrm{i}\phi}\sigma, \quad \theta \to \tilde{\theta} = \theta. \qquad (2.23)$$

It is useful to obtain closed-form equations for the divergence θ and the scalar shear σ of a geodesic congruence. The Raychaudhuri equation (for a hypersurface orthogonal, null, geodesic field $\mathbf{n}$) acquires a more elegant form,

$$\frac{d}{d\tau}\theta \equiv \nabla_{\mathbf{n}}\theta = -\frac{1}{2}\theta^2 + \mathrm{Ric}(\mathbf{n}, \mathbf{n}) - 2\sigma\bar{\sigma}.$$

Since $\sigma\bar{\sigma}$ is invariant under the transformation (2.23), the Raychaudhuri equation does not depend on the completion of $\mathbf{n}$ to a null tetrad.

Statement: It follows from Eq. (2.20) that $\mathrm{Tr}\,\sigma^{\perp}_{(\mathbf{u})}\sigma^{\perp}_{(\mathbf{u})} = 2\sigma\bar{\sigma}$.

Derivation: Calculation **(omitted)**.

Under some additional assumptions about the null tetrad, a similar equation holds for the scalar shear σ,

$$\frac{d}{d\tau}\sigma \equiv \nabla_{\mathbf{n}}\sigma = -\frac{\theta + \bar{\theta}}{2}\sigma + R(\mathbf{n}, \mathbf{m}, \mathbf{n}, \mathbf{m}). \qquad (2.24)$$

We shall now derive this equation and detail the necessary assumptions.

We need to compute the quantity

$$\begin{aligned}\nabla_{\mathbf{n}}\sigma \equiv \nabla_{\mathbf{n}}B_{(\mathbf{n})}(\mathbf{m}, \mathbf{m}) &\equiv \nabla_{\mathbf{n}}g(\nabla_{\mathbf{m}}\mathbf{n}, \mathbf{m}) \\ &= g(\nabla_{\mathbf{n}}\nabla_{\mathbf{m}}\mathbf{n}, \mathbf{m}) + g(\nabla_{\mathbf{m}}\mathbf{n}, \nabla_{\mathbf{n}}\mathbf{m}).\end{aligned}$$

The appearance of nested covariant derivatives suggests that we introduce the Riemann tensor,

$$g(\nabla_{\mathbf{n}}\nabla_{\mathbf{m}}\mathbf{n}, \mathbf{m}) = R(\mathbf{n}, \mathbf{m}, \mathbf{n}, \mathbf{m}) + g(\nabla_{[\mathbf{n},\mathbf{m}]}\mathbf{n}, \mathbf{m}).$$

(We used $\nabla_{\mathbf{n}}\mathbf{n} = 0$.) Thus

$$\nabla_{\mathbf{n}}\sigma = R(\mathbf{n}, \mathbf{m}, \mathbf{n}, \mathbf{m}) + g(\nabla_{[\mathbf{n},\mathbf{m}]}\mathbf{n}, \mathbf{m}) + g(\nabla_{\mathbf{m}}\mathbf{n}, \nabla_{\mathbf{n}}\mathbf{m}). \qquad (2.25)$$

By virtue of the following obvious properties,

$$\begin{aligned}&g(\nabla_{\mathbf{m}}\mathbf{n}, \mathbf{n}) = 0,\ g(\nabla_{\mathbf{m}}\mathbf{n}, \mathbf{m}) \equiv \sigma,\ g(\nabla_{\mathbf{m}}\mathbf{n}, \bar{\mathbf{m}}) \equiv -\frac{1}{2}\theta, \\ &g(\nabla_{\mathbf{n}}\mathbf{m}, \mathbf{n}) = \nabla_{\mathbf{n}}g(\mathbf{m}, \mathbf{n}) - g(\mathbf{m}, \nabla_{\mathbf{n}}\mathbf{n}) = 0,\ g(\nabla_{\mathbf{n}}\mathbf{m}, \mathbf{m}) = 0,\end{aligned}$$

we have

$$\begin{aligned}\nabla_{\mathbf{m}}\mathbf{n} &= -\sigma\bar{\mathbf{m}} + \frac{1}{2}\theta\mathbf{m} + \alpha\mathbf{n}, \\ \nabla_{\mathbf{n}}\mathbf{m} &= \zeta\mathbf{m} + \gamma\mathbf{n},\end{aligned}$$

where we introduced unknown scalar functions

$$\alpha \equiv g(\nabla_{\mathbf{m}}\mathbf{n}, \mathbf{l}),\ \zeta \equiv -g(\nabla_{\mathbf{n}}\mathbf{m}, \bar{\mathbf{m}}),\ \gamma \equiv g(\nabla_{\mathbf{n}}\mathbf{m}, \mathbf{l}).$$

These constants depend on the completion of $\mathbf{n}$ to a null tetrad and should not enter the final expression. Now we can straightforwardly evaluate:

$$\begin{aligned}g(\nabla_{\mathbf{m}}\mathbf{n}, \nabla_{\mathbf{n}}\mathbf{m}) &= \zeta\sigma, \\ [\mathbf{n}, \mathbf{m}] &= \left(\zeta - \frac{1}{2}\theta\right)\mathbf{m} + \sigma\bar{\mathbf{m}} + (\gamma - \alpha)\mathbf{n}, \\ g(\nabla_{[\mathbf{n},\mathbf{m}]}\mathbf{n}, \mathbf{m}) &= B_{(\mathbf{n})}([\mathbf{n}, \mathbf{m}], \mathbf{m}) \\ &= \left(\zeta - \frac{1}{2}\theta\right)B_{(\mathbf{n})}(\mathbf{m}, \mathbf{m}) + \sigma B_{(\mathbf{n})}(\bar{\mathbf{m}}, \mathbf{m}) \\ &= \left(\zeta - \frac{1}{2}\theta\right)\sigma - \frac{1}{2}\bar{\theta}\sigma.\end{aligned}$$

Finally, Eq. (2.25) is rewritten as

$$\nabla_{\mathbf{n}}\sigma = R(\mathbf{n}, \mathbf{m}, \mathbf{n}, \mathbf{m}) - \frac{1}{2}(\theta + \bar{\theta})\sigma + 2\zeta\sigma.$$

This equation still contains the function $\zeta \equiv -g(\nabla_{\mathbf{n}}\mathbf{m}, \bar{\mathbf{m}})$ that depends on the completion of $\mathbf{n}$ to a null tetrad and is thus arbitrary. We would like to eliminate this "gauge dependence." Note that the function ζ always has pure imaginary values and changes under the replacement (2.21) as

$$\zeta \to \tilde{\zeta} = \zeta + \mathrm{i}\nabla_{\mathbf{n}}\phi.$$

Therefore, for a given null tetrad, there exists a suitable function ϕ such that new null tetrad $\{\mathbf{l}, \mathbf{n}, e^{\mathrm{i}\phi}\mathbf{m}, e^{-\mathrm{i}\phi}\bar{\mathbf{m}}\}$ has $\tilde{\zeta} = 0$. (Obviously, the function ϕ can be found by integrating $\mathrm{i}\zeta$ along the orbits of $\mathbf{n}$.) Once we have $\zeta = 0$, the required equation (2.24) follows.

Alternatively, we may *redefine* the scalar shear for a given null tetrad as

$$\sigma_g \equiv e^{2\mathrm{i}\lambda}\sigma,$$

where the function λ is such that

$$\mathrm{i}\nabla_{\mathbf{n}}\lambda = g(\nabla_{\mathbf{n}}\mathbf{m}, \bar{\mathbf{m}}).$$

(There is, of course, a considerable freedom in choosing the function λ.) The redefined "gauge-covariant shear" σ_g still depends on the choice of the tetrad and of the function λ, but is "gauge-covariant" in the sense that a tetrad transformation (2.21), (2.23) yields

$$\nabla_{\mathbf{n}}\tilde{\sigma}_g = e^{2\mathrm{i}\phi}\nabla_{\mathbf{n}}\sigma_g.$$

The quantity $\sigma_g\bar{\sigma}_g$ remains invariant as before, and Eq. (2.24) holds for σ_g with any choice of tetrad.

Calculation: Verify the above equation, where $\tilde{\sigma} = e^{2\mathrm{i}\phi}\sigma$ and $\tilde{\sigma}_g \equiv e^{2\mathrm{i}\tilde{\lambda}}\tilde{\sigma}$.

Derivation: Let λ be the required function for the original tetrad and $\tilde{\lambda}$ for the modified tetrad with $\tilde{\mathbf{m}} = e^{\mathrm{i}\phi}\mathbf{m}$. Then we compute

$$\begin{aligned}
\nabla_{\mathbf{n}}\sigma_g &= (2\mathrm{i}\nabla_{\mathbf{n}}\lambda)\sigma_g + e^{2\mathrm{i}\lambda}\nabla_{\mathbf{n}}\sigma, \\
\mathrm{i}\nabla_{\mathbf{n}}\tilde{\lambda} &= g(\nabla_{\mathbf{n}}e^{\mathrm{i}\phi}\mathbf{m}, e^{-\mathrm{i}\phi}\bar{\mathbf{m}}) = \mathrm{i}\nabla_{\mathbf{n}}\lambda - \mathrm{i}\nabla_{\mathbf{n}}\phi, \\
\tilde{\lambda} &= \lambda - \phi, \\
\nabla_{\mathbf{n}}\tilde{\sigma}_g &= \nabla_{\mathbf{n}}(e^{2\mathrm{i}\tilde{\lambda}}\tilde{\sigma}) = (2\mathrm{i}\nabla_{\mathbf{n}}\tilde{\lambda})\tilde{\sigma}_g + e^{2\mathrm{i}\tilde{\lambda}}\nabla_{\mathbf{n}}\tilde{\sigma} \\
&= 2\mathrm{i}\tilde{\sigma}_g\nabla_{\mathbf{n}}(\lambda - \phi) + (2\mathrm{i}\nabla_{\mathbf{n}}\phi)\tilde{\sigma}_g + e^{2\mathrm{i}\tilde{\lambda}+2\mathrm{i}\phi}\nabla_{\mathbf{n}}\sigma \\
&= e^{2\mathrm{i}\phi}\nabla_{\mathbf{n}}\sigma_g.
\end{aligned}$$

3 Asymptotically flat spacetimes

If a spacetime contains a finite island of matter surrounded by an infinite sea of vacuum, we intuitively expect that the spacetime is "almost flat sufficiently far from the matter." For example, the metric (1.38) for the Schwarzschild spacetime is approximately equal to the Minkowski metric in spherical coordinates in the limit $r \to \infty$. Such spacetimes containing "well-isolated systems" are called **asymptotically flat**. In this chapter we shall formulate and explore the property of "asymptotic flatness" in more detail. We begin by considering an easier case of stationary spacetimes.

3.1 Stationary spacetimes

Suppose a metric g is given in coordinates $\{x^\mu\}$, and the coefficients $g_{\mu\nu}$ are independent of one of the coordinates, say x^1. Then ∂_{x^1} is a Killing vector for the metric g; in other words, the flow of the vector field ∂_{x^1} leaves the metric invariant. Thus, Killing vectors provide a formalization of the concept of "coordinate symmetry" of the metric. Using Killing vectors, the heuristic concept of a "time-independent metric" is made precise in the following way.

A spacetime is called **stationary** if it possesses a timelike Killing vector field. The existence of such a vector field leads to many useful properties that we shall now explore. A stronger property is being *static* rather than stationary. A spacetime is **static** if it is stationary *and* the timelike Killing vector is hypersurface-orthogonal. For now, it suffices to consider stationary spacetimes.

In a stationary spacetime, the orbits of the Killing vector $\mathbf{k}$ can be seen as preferred worldlines along which the geometry does not change. It is useful to consider imaginary observers moving along these worldlines. These are called **stationary observers** and have 4-velocities

$$\mathbf{u} = \frac{1}{\sqrt{g(\mathbf{k},\mathbf{k})}}\mathbf{k}. \tag{3.1}$$

Stationary observers find that the geometry of the spacetime around them is the same at all times. However, these stationary observers are not necessarily geodesic observers, i.e. they are not necessarily freely falling. For example, an observer sitting at the surface of a nonrotating planet observers that the geometry is time-independent. This is a stationary observer who is, obviously, not freely falling.

The following statement lists the most important properties of Killing vectors with regard to being geodesic.

Statement 3.1.0.1: **(a)** For a Killing vector $\mathbf{k}$ the property $\nabla_{\mathbf{k}} g(\mathbf{k},\mathbf{k}) = 0$ holds. **(b)** The function $g(\mathbf{k},\mathbf{k})$ is everywhere constant iff the vector $\mathbf{k}$ is geodesic. **(c)** In the domain where $g(\mathbf{k},\mathbf{k}) \neq 0$, the normalized vector field $\mathbf{u}$ defined by Eq. (3.1) is geodesic iff $\mathbf{k}$ is geodesic.

Idea of proof: Compute the scalar product of $\nabla_{\mathbf{k}}\mathbf{k}$ with an arbitrary vector and use the Killing equation (1.50).

Proof of Statement 3.1.0.1: **(a)** The property $\nabla_{\mathbf{k}} g(\mathbf{k},\mathbf{k}) = 0$ follows directly from the Killing equation (1.50). **(b)** For an arbitrary $\mathbf{x}$, we have

$$g(\nabla_{\mathbf{k}}\mathbf{k},\mathbf{x}) = -g(\mathbf{k},\nabla_{\mathbf{x}}\mathbf{k}) = -\frac{1}{2}\nabla_{\mathbf{x}} g(\mathbf{k},\mathbf{k}).$$

Thus, the scalar product of $\nabla_{\mathbf{k}}\mathbf{k}$ with an arbitrary vector vanishes iff the derivative of $g(\mathbf{k},\mathbf{k})$ in an arbitrary direction vanishes. **(c)** Assuming that $g(\mathbf{k},\mathbf{k}) \neq 0$, we compute

$$\begin{aligned}\nabla_{\mathbf{u}}\mathbf{u} &= \frac{1}{g(\mathbf{k},\mathbf{k})}\nabla_{\mathbf{k}}\mathbf{k} + \frac{\mathbf{k}}{\sqrt{g(\mathbf{k},\mathbf{k})}}\nabla_{\mathbf{k}}\frac{1}{\sqrt{g(\mathbf{k},\mathbf{k})}} \\ &= \frac{1}{g(\mathbf{k},\mathbf{k})}\nabla_{\mathbf{k}}\mathbf{k}.\end{aligned}$$

Thus $\nabla_{\mathbf{u}}\mathbf{u} = 0$ iff $\nabla_{\mathbf{k}}\mathbf{k} = 0$. ■

In general, a Killing vector is not normalized, so stationary worldlines generally correspond to *accelerated* observers rather than to freely falling observers. For example, the stationary worldlines in the Schwarzschild spacetime describe observers remaining at a constant distance from the center of mass. These observers move with a constant acceleration and thus feel a constant pull of gravity. Accordingly, the Killing vector ∂_t is not normalized in the Schwarzschild spacetime, i.e. $g(\mathbf{k},\mathbf{k}) \neq$ const.

3.1.1 Newtonian limit

In the framework of General Relativity, the **Newtonian limit** is the assumption that gravitation is weak and that there exists a globally inertial reference frame—the "absolute spacetime." In other words, a Newtonian setting considers a set of "absolute" observers who see each other as motionless or moving with constant velocities at all times. (Of course, this assumption is only *approximately* correct in the actual curved spacetime; also the clocks at different points run at different rates and cannot be exactly synchronized everywhere at all times.) For example, someone sitting on the surface of the Earth could be such an "absolute" observer. Since "absolute" observers do not necessarily move along geodesic lines, they will perceive the motion of freely falling bodies as an *accelerated* motion occurring in the "absolute" reference frame. The Newtonian theory explains this apparently accelerated motion as being due to "gravitational forces." We now formalize these assumptions and derive the Newtonian equations of gravitation.

A congruence of the world-lines of the "absolute" observers is described by a timelike vector field $\mathbf{v}$, normalized as $g(\mathbf{v},\mathbf{v}) = 1$. The observers measure the "absolute time" τ along the orbits of $\mathbf{v}$. The central assumption of the Newtonian limit is that these observers are "fixed" with respect to each other, so any neighbor observers appear to move without relative acceleration. In other words, *any* connecting vector field $\mathbf{c}$, such that $[\mathbf{c},\mathbf{v}] = 0$, appears (approximately!) unaccelerated, $\nabla_{\mathbf{v}}\nabla_{\mathbf{v}}\mathbf{c} \approx 0$.

Let us find out how the "stationary" observers perceive the motion of freely falling (geodesic) particles. If a geodesic has instantaneously the same tangent vector $\mathbf{v}$ as one of the fixed observers, the observer's acceleration relative to the geodesic

is $\nabla_{\mathbf{v}}\mathbf{v}$, therefore the observer will measure the acceleration of the freely falling particle as $\mathbf{a} \equiv -\nabla_{\mathbf{v}}\mathbf{v}$. Our goal is to derive a formula describing the acceleration $\mathbf{a}$.

The plan of the derivation is the following. First we use the properties of connecting vectors to deduce that the vector field $\mathbf{a}$ is integrable. It will follow that there exists a scalar function ϕ such that $\mathbf{a} = \hat{g}^{-1}\mathrm{d}\phi$. This scalar function is called the **gravitational potential**. Then we shall derive an equation for ϕ, using the Einstein equation with matter sources consisting of "fixed" matter (not moving with respect to the fixed observers) with mass density $\rho(x)$. This equation will coincide with the (three-dimensional) Poisson equation $\Delta\phi = 4\pi\rho$, which is equivalent to Newton's law of gravity. In this way we reproduce the Newtonian theory of gravitation.

The acceleration field $\mathbf{a}$ is integrable if the bilinear form $B_{(\mathbf{a})}(\mathbf{x},\mathbf{y})$ defined by Eq. (2.10) is symmetric. The form of the expression

$$B_{(\mathbf{a})}(\mathbf{x},\mathbf{y}) \equiv -g(\nabla_{\mathbf{x}}\nabla_{\mathbf{v}}\mathbf{v},\mathbf{y})$$

suggests that we should consider the combination

$$g(\nabla_{\mathbf{v}}\nabla_{\mathbf{x}}\mathbf{v} - \nabla_{\mathbf{x}}\nabla_{\mathbf{v}}\mathbf{v} - \nabla_{[\mathbf{v},\mathbf{x}]}\mathbf{v},\mathbf{y}) = R(\mathbf{v},\mathbf{x},\mathbf{v},\mathbf{y})$$

and try to simplify it using the given information about the connecting vectors. We substitute $\mathbf{x} = \mathbf{c}$ and obtain

$$0 = g(\nabla_{\mathbf{v}}\nabla_{\mathbf{v}}\mathbf{c},\mathbf{y}) = g(\nabla_{\mathbf{v}}\nabla_{\mathbf{c}}\mathbf{v},\mathbf{y}) = R(\mathbf{v},\mathbf{c},\mathbf{v},\mathbf{y}) + g(\nabla_{\mathbf{c}}\nabla_{\mathbf{v}}\mathbf{v},\mathbf{y})$$

and therefore

$$B_{(\mathbf{a})}(\mathbf{c},\mathbf{y}) \equiv -g(\nabla_{\mathbf{c}}\nabla_{\mathbf{v}}\mathbf{v},\mathbf{y}) = R(\mathbf{v},\mathbf{c},\mathbf{v},\mathbf{y}). \qquad (3.2)$$

Since $R(\mathbf{v},\mathbf{c},\mathbf{v},\mathbf{y}) = R(\mathbf{v},\mathbf{y},\mathbf{v},\mathbf{c})$, we find

$$B_{(\mathbf{a})}(\mathbf{c},\mathbf{y}) = B_{(\mathbf{a})}(\mathbf{y},\mathbf{c}).$$

In fact, the above relation holds for arbitrary vector fields $\mathbf{x},\mathbf{y}$ at any point, $B_{(\mathbf{a})}(\mathbf{x},\mathbf{y}) = B_{(\mathbf{a})}(\mathbf{y},\mathbf{x})$, because it involves only the *value* of the connecting vector $\mathbf{c}$ at one point, and not the fact that $\mathbf{c}$ is a connecting vector. Hence, the bilinear form $B_{(\mathbf{a})}$ is symmetric, the vector field $\mathbf{a}$ is integrable, and there exists a scalar function ϕ such that $\mathbf{a} = \hat{g}^{-1}\mathrm{d}\phi$.

To derive an equation for ϕ, we use Eq. (3.2) and the definition of divergence,

$$\mathrm{div}\mathbf{a} = \mathrm{Tr}_{(\mathbf{x},\mathbf{y})}g(\nabla_{\mathbf{x}}\mathbf{a},\mathbf{y}).$$

It follows that

$$\mathrm{div}\mathbf{a} = -\mathrm{div}\,(\nabla_{\mathbf{v}}\mathbf{v}) = \mathrm{Tr}_{(\mathbf{x},\mathbf{y})}R(\mathbf{v},\mathbf{x},\mathbf{v},\mathbf{y}) = \mathrm{Ric}(\mathbf{v},\mathbf{v}).$$

On the other hand,

$$\begin{aligned}\mathrm{div}\mathbf{a} &= \mathrm{div}\hat{g}^{-1}\mathrm{d}\phi \\ &= \mathrm{Tr}_{(\mathbf{x},\mathbf{y})}g(\nabla_{\mathbf{x}}\hat{g}^{-1}\mathrm{d}\phi,\mathbf{y}) \\ &= \mathrm{Tr}_{(\mathbf{x},\mathbf{y})}\left(\nabla_{\mathbf{x}}\mathrm{d}\phi\right)\circ\mathbf{y} = \Box\phi,\end{aligned}$$

where the **D'Alembertian** $\Box\phi$ is defined as the trace of the bilinear form $\nabla_{\ldots}\nabla_{\ldots}\phi$,

$$\Box\phi \equiv \mathrm{Tr}_{(\mathbf{x},\mathbf{y})}(\nabla_{\ldots}\nabla_{\ldots}\phi)\circ(\mathbf{x},\mathbf{y}) \equiv \mathrm{Tr}_{(\mathbf{x},\mathbf{y})}(\nabla_{\mathbf{x}}\nabla_{\mathbf{y}}\phi - \nabla_{\nabla_{\mathbf{x}}\mathbf{y}}\phi).$$

In components, $\Box\phi = \nabla_\mu\nabla^\mu\phi$. Hence, we obtain the following equation for ϕ,

$$\Box\phi = \mathrm{Ric}(\mathbf{v},\mathbf{v}).$$

To obtain a concrete expression for $\mathrm{Ric}(\mathbf{v},\mathbf{v})$, we need to relate the curvature to the matter content of the spacetime. In the Newtonian approximation, the matter sources are approximately motionless particles, so the energy-momentum tensor of matter is $T_{\mu\nu} = \rho v_\mu v_\nu$, where $\rho(x)$ is the mass density. The Einstein equation,

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R^\alpha_\alpha = -8\pi T_{\mu\nu},$$

can be rewritten as

$$R_{\mu\nu} = -8\pi\left(T_{\mu\nu} - \frac{1}{2}g_{\mu\nu}T^\alpha_\alpha\right).$$

Since $v_\mu v^\mu = 1$, we find $T^\mu_\mu = \rho$ and finally

$$\Box\phi = R_{\mu\nu}v^\mu v^\nu = -4\pi\rho.$$

This is in fact the familiar Poisson equation relating the Newtonian gravitational potential ϕ to the density ρ of matter, written in a Lorentz-covariant way. In an almost-flat spacetime, we can choose the coordinate t along the orbits of $\mathbf{v}$, and then we have approximately

$$\Box\phi = \left(\partial_t^2 - \Delta\right)\phi = -4\pi\rho, \quad \Delta \equiv \partial_x^2 + \partial_y^2 + \partial_z^2.$$

The time derivatives are negligible since the distribution of matter is almost stationary (in the SI units, this is made more explicit by the factor $\frac{1}{c^2}\partial_t^2$). Hence, we recover the three-dimensional **Poisson equation**

$$\Delta\phi = 4\pi\rho.$$

Note that the three-dimensional acceleration $a_i = -\partial_i\phi$, where $i = x,y,z$, because of the negative sign of the spatial part of the metric.

Example: Schwarzschild spacetime

The metric is given by Eq. (1.38) and the stationary observers follow lines of fixed r,ϕ,θ for $r > 2m$. We expect the Newtonian approximation to hold for large $r \gg 2m$. The vector $\mathbf{v}$ is thus proportional to ∂_t, and normalization $g(\mathbf{v},\mathbf{v}) = 1$ yields

$$\mathbf{v} = \left(1 - \frac{2m}{r}\right)^{-1/2}\partial_t.$$

The vectors $\partial_\theta, \partial_\phi$ are connecting vectors for $\mathbf{v}$, but ∂_r is not,

$$[\mathbf{v},\partial_r] = \frac{m}{r(r-2m)}\mathbf{v}.$$

However, the approximation $[\mathbf{v},\partial_r] \approx 0$ is good for large r, namely for $r \gg m$.

A calculation shows that the acceleration field $\nabla_{\mathbf{v}}\mathbf{v}$ is *exactly* integrable (not only in the Newtonian approximation). Let us compute the scalar products of $\nabla_{\mathbf{v}}\mathbf{v}$ with the basis vectors $\mathbf{v},\partial_r,\partial_\theta,\partial_\phi$:

$$\begin{aligned}g(\nabla_{\mathbf{v}}\mathbf{v},\mathbf{v}) &= \frac{1}{2}\nabla_{\mathbf{v}}g(\mathbf{v},\mathbf{v}) = 0,\\ g(\nabla_{\mathbf{v}}\mathbf{v},\partial_\theta) &= -g(\nabla_{\mathbf{v}}\partial_\theta,\mathbf{v}) = -g(\nabla_{\partial_\theta}\mathbf{v},\mathbf{v}) = 0,\\ g(\nabla_{\mathbf{v}}\mathbf{v},\partial_\phi) &= 0,\\ g(\nabla_{\mathbf{v}}\mathbf{v},\partial_r) &= -g(\nabla_{\mathbf{v}}\partial_r,\mathbf{v}) = -g(\mathbf{v},[\mathbf{v},\partial_r]) - g(\mathbf{v},\nabla_{\partial_r}\mathbf{v})\\ &= -\frac{m}{r(r-2m)}.\end{aligned}$$

Therefore

$$\nabla_{\mathbf{v}}\mathbf{v} = \frac{g(\nabla_{\mathbf{v}}\mathbf{v}, \partial_r)}{g(\partial_r, \partial_r)}\partial_r = \frac{m}{r^2}\partial_r.$$

The potential is easier to find from the 1-form $\hat{g}\nabla_{\mathbf{v}}\mathbf{v}$, which is expressed as

$$\hat{g}\nabla_{\mathbf{v}}\mathbf{v} = -\frac{m}{r(r-2m)}\mathrm{d}r \equiv \mathrm{d}\Phi(r),$$
$$\Phi(r) \equiv \frac{1}{2}\ln\left(1 - \frac{2m}{r}\right). \tag{3.3}$$

Thus $\Phi(r)$ is the "potential" for the gravitational force measured by the stationary observers in the Schwarzschild spacetime. For large $r \gg 2m$ the potential becomes $-m/r$, which is the familiar Newtonian potential due to a point mass m, while for $r \to 2m$ the gravitational force approaches infinity. (Note that the coordinate r does not coincide with the physical distance.)

Self-test question: Is the motion of test particles in the Schwarzschild spacetime actually the same as the Newtonian motion in the potential $\Phi(r)$ or in some other potential? *Answer:* No. (The "gravitational force" depends on velocity as well.) **Not sure how to explain! Need to make this clear.**

Remark: The fact that the Schwarzschild spacetime admits a "potential" $\Phi(r)$ is a consequence of spherical symmetry and time-independence, i.e. the metric depends only on the radius r. (More formally: The presence of an integrable Killing vector ∂_t shows that the spacetime is static; additionally, the "angular" Killing vectors ∂_θ and ∂_ϕ indicate the spherical symmetry.) Thus the acceleration of an observer at constant r is a vector directed along ∂_r whose magnitude is a function only of r. Therefore this acceleration must be equal to a gradient of some function $\Phi(r)$. In general, all stationary spacetimes admit such a "potential" although it may not be spherically symmetric; see Sec. 3.1.4.

Example: de Sitter spacetime

The metric is given by Eq. (1.39) and the stationary observers follow lines of fixed x, y, z. The vector $\mathbf{v}$ is simply equal to ∂_t and the basis of connecting vectors is $\partial_x, \partial_y, \partial_z$. Since the field $\mathbf{v}$ is integrable and normalized, it is geodesic and $\nabla_{\mathbf{v}}\mathbf{v} = 0$. Hence, stationary observers are at the same time freely falling and do not feel any gravity at all. However, the Newtonian approximation breaks down at sufficiently large distances. This can be seen by computing the relative acceleration of two nearby observers (see the following calculation),

$$\nabla_{\mathbf{v}}\nabla_{\mathbf{v}}\partial_x = H^2\partial_x.$$

At small distances $L \ll H^{-1}$, we are allowed to use the approximation that stationary observers are not accelerated relative to each other.

Calculation: For a connecting vector $\mathbf{c}$ which is one of $\partial_x, \partial_y, \partial_z$, show that

$$\nabla_{\mathbf{v}}\mathbf{c} = H\mathbf{c} \tag{3.4}$$

and thus derive the above equation. The relation (3.4) is called the **Hubble law**. (In a uniformly expanding spacetime, the velocity of a comoving observer is proportional to the distance to the observer.)

Solution: We can compute $\nabla_{\mathbf{v}}\mathbf{c}$ by taking scalar products; for instance,

$$g(\nabla_{\mathbf{v}}\mathbf{c}, \mathbf{c}) = \frac{1}{2}\nabla_{\mathbf{v}}g(\mathbf{c}, \mathbf{c}) = -He^{2Ht},$$

while scalar products of $\nabla_{\mathbf{v}}\mathbf{c}$ with other basis vectors vanish. Thus we obtain Eq. (3.4) and as a trivial consequence, $\nabla_{\mathbf{v}}\nabla_{\mathbf{v}}\mathbf{c} = H^2\mathbf{c}$.

3.1.2 Redshift

In a stationary spacetime, the magnitude of the Killing vector defines a scalar function, $z \equiv \sqrt{g(\mathbf{k}, \mathbf{k})}$. The function z is called the **redshift factor** because, as we shall now show, the local values of z characterize the variation of the photon frequency measured by different stationary observers.

A null geodesic with a null tangent vector $\mathbf{n}$ represents a path of a photon. The vector $\mathbf{n}$ may be rescaled in such a way that $g(\mathbf{n}, \mathbf{u})$ represents the energy E of the photon in the rest frame of a timelike observer with 4-velocity $\mathbf{u}$. From the Planck relation, $E = h\nu$, where h is Planck's constant, we find that the measured frequency of the photon is $\nu = h^{-1}g(\mathbf{n}, \mathbf{u})$. In a stationary spacetime, stationary observers have worldlines with 4-velocity

$$\mathbf{u} = \frac{1}{\sqrt{g(\mathbf{k}, \mathbf{k})}}\mathbf{k} = z^{-1}\mathbf{k}.$$

Therefore, a stationary observer at a given location will measure the frequency of a given lightray with a (null) worldline $\gamma(\tau)$ as

$$h\nu = g(z^{-1}\mathbf{k}, \dot{\gamma}).$$

By Statement 3.1.2.1 below, $g(\mathbf{k}, \dot{\gamma})$ remains constant along lightrays. Thus, for a given single photon moving along a null geodesic, the frequency measured by a stationary observer is proportional to the local value of z^{-1}. This justifies the name "redshift factor" for the function z.

Statement 3.1.2.1: If $\mathbf{k}$ is a Killing vector, so that Eq. (1.50) holds, and $\gamma(\tau)$ is a geodesic, then $g(\mathbf{k}, \dot{\gamma})$ is constant along γ.

Proof of Statement 3.1.2.1: Let $\mathbf{v}$ be a geodesic vector field containing $\dot{\gamma}$; then $\nabla_{\mathbf{v}}\mathbf{v} = 0$ and we compute

$$\nabla_{\mathbf{v}}g(\mathbf{k}, \mathbf{v}) = g(\nabla_{\mathbf{v}}\mathbf{k}, \mathbf{v}) = -g(\mathbf{v}, \nabla_{\mathbf{v}}\mathbf{k}),$$

thus $\nabla_{\mathbf{v}}g(\mathbf{k}, \mathbf{v}) = 0$. ■

It follows from Statement 3.1.0.1 that $\nabla_{\mathbf{k}}z = 0$; in other words, the redshift factor is constant for a given stationary observer. It also follows that the function z is *everywhere* constant (i.e. there is no redshift) iff the stationary observers are geodesic.

3.1.3 Conformal Killing vectors

In cosmology, one usually considers spacetimes that are not stationary, so no timelike Killing vectors are present. Nevertheless, sometimes there exists a "conformal Killing vector" and then the redshift function $z \equiv \sqrt{g(\mathbf{k}, \mathbf{k})}$ is still useful.

By definition, a **conformal Killing vector** is a vector field $\mathbf{k}$ such that

$$\mathcal{L}_{\mathbf{k}}g = 2\lambda g,$$

where λ is some scalar function. It follows directly from Eq. (1.48) that a conformal Killing vector $\mathbf{k}$ satisfies the equation

$$g(\nabla_{\mathbf{a}}\mathbf{k}, \mathbf{b}) + g(\mathbf{a}, \nabla_{\mathbf{b}}\mathbf{k}) = 2\lambda g(\mathbf{a}, \mathbf{b}),$$

where $\mathbf{a}, \mathbf{b}$ are arbitrary vectors. This is a modified version of Eq. (1.50).

When a spacetime admits a conformal Killing vector, the redshift function is again useful, as demonstrated by the following statement.

Statement 3.1.3.1: Assume that a spacetime admits a timelike conformal Killing vector $\mathbf{k}$, and call observers moving along $\mathbf{k}$ "conformally stationary." Then the frequency of a photon measured by conformally stationary observers is inversely proportional to the redshift function $z \equiv \sqrt{g(\mathbf{k}, \mathbf{k})}$.

Proof of Statement 3.1.3.1: For an arbitrary geodesic field $\mathbf{v}$ we have

$$\nabla_{\mathbf{v}} g(\mathbf{k}, \mathbf{v}) = g(\nabla_{\mathbf{v}}\mathbf{k}, \mathbf{v}) = \lambda g(\mathbf{v}, \mathbf{v}).$$

If the field $\mathbf{v}$ is null ($g(\mathbf{v}, \mathbf{v}) = 0$) then

$$\nabla_{\mathbf{v}} g(\mathbf{k}, \mathbf{v}) = 0,$$

so $g(\mathbf{k}, \mathbf{v})$ remains constant along the geodesic lines, just as in the case of ordinary Killing vectors. Defining $z = \sqrt{g(\mathbf{k}, \mathbf{k})}$ and $\mathbf{u} \equiv z^{-1}\mathbf{k}$, we again find

$$h\nu \equiv g(\mathbf{u}, \dot{\boldsymbol{\gamma}}) = z^{-1} g(\mathbf{k}, \dot{\boldsymbol{\gamma}}) \propto z^{-1}.$$

The photon's frequency is inversely proportional to the redshift factor. ■

Note however that $\nabla_{\mathbf{k}} g(\mathbf{k}, \mathbf{k}) = 2\lambda g(\mathbf{k}, \mathbf{k})$, so the redshift factor z is *not constant* along the worldlines of stationary observers.

Example 3.1.3.2: For a de Sitter spacetime with the metric (1.39), it is reasonable to guess that the vector $\mathbf{k} = f(t)\partial_t$ is a conformal Killing vector if the scalar function $f(t)$ is chosen correctly. Let us determine $f(t)$, the conformal factor λ, and the redshift z.

It suffices to compute $g(\nabla_{\mathbf{a}}\partial_t, \mathbf{b})$ for $\mathbf{a}, \mathbf{b}$ either ∂_t or ∂_x. We find that $\nabla_{\partial_x}\partial_t = H\partial_x$ is the only nonzero derivative (similarly for ∂_y and ∂_z). Hence

$$g(\nabla_{\partial_t} f\partial_t, \partial_t) = \dot{f} g(\partial_t, \partial_t),$$

and $\dot{f} = \lambda$. Further,

$$g(\nabla_{\partial_x} f\partial_t, \partial_x) = fH\, g(\partial_x, \partial_x) \equiv \lambda g(\partial_x, \partial_x),$$

thus $\lambda = fH = \dot{f}$ and $f(t) = e^{Ht}$, $\lambda = He^{Ht}$, $z = \sqrt{g(\mathbf{k}, \mathbf{k})} = e^{Ht}$. The frequency of photons decays with time t as e^{-Ht}. ■

Practice problem: For a flat Friedmann-Robertson-Walker metric,

$$g = dt^2 - a^2(t)(dx^2 + dy^2 + dz^2),$$

where $a(t)$ is a nonzero function, show that $\mathbf{k} = a(t)\partial_t$ is a conformal Killing vector with the factor $\lambda = \dot{a}(t)$. ■

Unlike the case of null geodesics, the behavior of timelike geodesics in the presence of a conformal Killing vector is not as straightforward to describe without additional assumptions.

Suppose that a metric g has a conformal Killing vector $\mathbf{k}$, and consider particles moving along a timelike geodesic field $\mathbf{v}$ normalized as $g(\mathbf{v}, \mathbf{v}) = 1$. The relative velocity $\delta\vec{v}$ of the particles with respect to the conformally stationary observers is characterized by the scalar product $g(\mathbf{v}, \mathbf{u})$ according to the standard special relativistic formula for the "gamma factor,"

$$\gamma \equiv \frac{1}{\sqrt{1 - (\delta\vec{v})^2}} = g(\mathbf{v}, \mathbf{u}).$$

We would like to derive an equation for $\gamma(\tau)$ along a particle's wordline parameterized by the proper time τ. This is possible in the following special case.

Calculation 3.1.3.3: Consider stationary observers whose worldlines are parallel to the conformal Killing vector $\mathbf{k} = e^{Ht}\partial_t$ for de Sitter spacetime with the metric (1.39). A point particle of mass m is moving along a timelike geodesic; the motion is slow relative to the conformally stationary observers. Show that the relative 3-velocity $\delta\vec{v}$ of the particle decreases exponentially at late times. In the Newtonian approximation, the slowdown can be ascribed to a "gravitational friction force" $\vec{F}$ which is proportional to the velocity according to the law

$$\vec{F} = -mH\delta\vec{v}.$$

(Details on page 179.) ■

In a more general situation, a closed equation for $\gamma(\tau)$ can be derived if we assume additionally that $\mathbf{k}$ is an *integrable* vector field.

Statement 3.1.3.4: Suppose $\mathbf{k}$ is an integrable, timelike conformal Killing vector field, $\mathcal{L}_{\mathbf{k}} g = 2\lambda g$; the vector field $\mathbf{u}$ is defined by $\mathbf{u} = z^{-1}\mathbf{k}$, where z is the redshift function, $z \equiv \sqrt{g(\mathbf{k}, \mathbf{k})}$; and $\mathbf{v}$ is a timelike geodesic normalized by $g(\mathbf{v}, \mathbf{v}) = 1$. Then the scalar product $\gamma \equiv g(\mathbf{u}, \mathbf{v})$ satisfies the equation

$$\nabla_{\mathbf{v}}\gamma = \left(1 - \gamma^2\right)\lambda z^{-1}. \tag{3.5}$$

(Proof on page 179.) ■

For a particle moving with a small relative velocity $\delta\vec{v}$ with respect to conformally stationary observers, Eq. (3.5) can be interpreted in the Newtonian limit. We have $\gamma \approx 1 + \frac{1}{2}(\delta\vec{v})^2$, so Eq. (3.5) can be rewritten as the equation for the "local kinetic energy,"

$$\frac{d}{d\tau} E_{\text{kin}} = \delta\vec{v} \cdot \vec{F}, \quad \vec{F} \equiv -m\lambda z^{-1}\delta\vec{v}, \quad E_{\text{kin}} \equiv \frac{1}{2}(\delta\vec{v})^2.$$

According to the Newtonian description from the viewpoint of a conformally stationary observer, the kinetic energy of the particle is decreased due to the "force of gravitational friction" $\vec{F}$ that acts opposite and proportional to the velocity $\delta\vec{v}$ of the particle.

3.1.4 Gravitational potential

In Sec. 3.1.1 we have seen that the Schwarzschild spacetime possesses a "gravitational potential" Φ such that the 3-acceleration of freely falling bodies in stationary reference frames is described by the Newtonian formula, $\vec{a} = -\vec{\nabla}\Phi$. This fact is a manifestation of a more general property: Every stationary spacetime admits a "gravitational potential" in this sense.

Statement:[1] If $\mathbf{k}$ is a Killing vector, show that $\nabla_{\mathbf{k}}\mathbf{k}$ is integrable. If $\mathbf{k}$ is timelike (a stationary spacetime), show that the

[1] After problem 4a in chapter 6 of [36].

acceleration field $\nabla_{\mathbf{u}}\mathbf{u}$ is integrable for a stationary observer with the 4-velocity $\mathbf{u} = z^{-1}\mathbf{k}$, where $z \equiv \sqrt{g(\mathbf{k},\mathbf{k})}$ is the redshift factor. Derive the "gravitational potential" for the acceleration field $\nabla_{\mathbf{u}}\mathbf{u}$.

Derivation: For an arbitrary vector $\mathbf{x}$,

$$g(\nabla_{\mathbf{k}}\mathbf{k},\mathbf{x}) = -g(\mathbf{k},\nabla_{\mathbf{x}}\mathbf{k}) = -\mathbf{x}\circ\frac{1}{2}g(\mathbf{k},\mathbf{k}) = -\frac{1}{2}\nabla_{\mathbf{x}}z^2,$$

thus $\nabla_{\mathbf{k}}\mathbf{k} = \hat{g}^{-1}\mathrm{d}f$ with $f \equiv -\frac{1}{2}z^2$. Since $\mathbf{u} = z^{-1}\mathbf{k}$ and $\nabla_{\mathbf{k}}z = 0$, we find

$$\begin{aligned}\nabla_{\mathbf{u}}\mathbf{u} &= \frac{1}{z}\nabla_{\mathbf{k}}\left(\frac{\mathbf{k}}{z}\right) = \frac{1}{z^2}\nabla_{\mathbf{k}}\mathbf{k} + \frac{\mathbf{k}}{z}\nabla_{\mathbf{k}}\frac{1}{z}\\ &= \frac{1}{z^2}\nabla_{\mathbf{k}}\mathbf{k} = -\frac{1}{2}\hat{g}^{-1}\frac{1}{z^2}\mathrm{d}\left(z^2\right)\\ &= -\hat{g}^{-1}\mathrm{d}\ln z.\end{aligned}$$

Thus the "gravitational potential" Φ is

$$\Phi = \ln z = \frac{1}{2}\ln g(\mathbf{k},\mathbf{k}), \quad \nabla_{\mathbf{u}}\mathbf{u} = -\hat{g}^{-1}\mathrm{d}\Phi.$$

(Note the minus sign in the above formula: The 4-acceleration of a freely falling body in a stationary frame is $-\nabla_{\mathbf{u}}\mathbf{u}$, and the 3-acceleration, $\vec{a} = -\vec{\nabla}\Phi$, has the opposite sign relative to the 4-acceleration. ***Alarm: what about the minus sign we get while computing $\hat{g}^{-1}\mathrm{d}\Phi$ from the minus sign in the spatial components of g?) This general expression agrees with the result (3.3) derived above for the Schwarzschild spacetime with the Killing vector $\mathbf{k} \equiv \partial_t$,

$$\Phi = \frac{1}{2}\ln g(\partial_t,\partial_t) = \frac{1}{2}\ln\left(1-\frac{2m}{r}\right).$$

According to our (so far) heuristic picture of an asymptotically flat spacetime as something generated by a "well-isolated" clump of gravitating matter, we expect that at large distances the acceleration of stationary observers goes to zero, thus the gravitational potential Φ becomes constant. Since $\Phi = \ln z$, this is equivalent to assuming that the redshift z becomes constant at infinite distances. Then it is convenient to multiply the Killing vector $\mathbf{k}$ by a constant to achieve $z \to 1$ and $\Phi \to 0$ at infinity. After this rescaling, the redshift function z describes the redshift of lightrays with respect to observers at infinite distances.

Remark: A non-asymptotically flat spacetime may involve unbounded redshifts at infinity. An example is the de Sitter spacetime with the metric (1.39), which possesses a conformal Killing vector $e^{2Ht}\partial_t$. The redshift function is $z = e^{2Ht}$ and grows unboundedly along any lightray.

Note that the potential Φ is stationary:

$$\mathbf{k}\circ\Phi = -g(\mathbf{k},\nabla_{\mathbf{u}}\mathbf{u}) = 0.$$

On physical grounds, we expect that the 3-acceleration of a freely falling mass points "towards the center of gravity," at least when the test mass is outside of the clump of gravitating matter. Thus, the gravitational potential decreases "away from infinity," and outside of the matter sources we have

$$\Phi < 0, \quad z < 1.$$

3.1.5 Energy

In a stationary spacetime, the Killing vector $\mathbf{k}$ generates time translations along the stationary worldlines. The local geometry of the spacetime is invariant under these time translations: for instance, the scalar product of connecting vectors remains constant. The existence of such time translations, in turn, leads to a conserved quantity, interpreted as the total energy of the spacetime.

We know that the distortion $B_{(\mathbf{k})}$ of a Killing vector $\mathbf{k}$ is a 2-form,

$$B_{(\mathbf{k})}(\mathbf{x},\mathbf{y}) \equiv g(\nabla_{\mathbf{x}}\mathbf{k},\mathbf{y}) = -g(\nabla_{\mathbf{y}}\mathbf{k},\mathbf{x}).$$

In the index notation, we have

$$B_{(\mathbf{k})\alpha\beta} = \nabla_\alpha k_\beta.$$

The divergence of this 2-form is a 1-form e,

$$e_\beta \equiv \nabla^\alpha B_{(\mathbf{k})\alpha\beta} = \nabla^\alpha\nabla_\alpha k_\beta \equiv \Box k_\beta; \quad \mathbf{e} = \Box\mathbf{k}.$$

Written in the index-free notation, the expression for e is significantly more cumbersome,

$$\begin{aligned}e\circ\mathbf{z} &\equiv \mathrm{Tr}_{(\mathbf{x},\mathbf{y})}\left[\nabla_{\mathbf{x}}B_{(\mathbf{k})}\right]\circ(\mathbf{y},\mathbf{z})\\ &= \mathrm{Tr}_{(\mathbf{x},\mathbf{y})}\left[\nabla_{\mathbf{x}}g(\nabla_{\mathbf{y}}\mathbf{k},\mathbf{z}) - g(\nabla_{\nabla_{\mathbf{x}}\mathbf{y}},\mathbf{z}) - g(\nabla_{\mathbf{y}}\mathbf{k},\nabla_{\mathbf{x}}\mathbf{z})\right].\end{aligned}$$

So we shall use the index notation in the following calculations. The 1-form e is divergence-free,[2]

$$\nabla^\beta e_\beta = \nabla^\beta\nabla^\alpha B_{(\mathbf{k})\alpha\beta} = 0,$$

therefore it defines a conserved vector current $\mathbf{e} \equiv \hat{g}^{-1}e$. The flux of $\mathbf{e}$ through a (spacelike) 3-volume V with a (timelike) normal vector $\mathbf{v}$ is constant, up to boundary terms. Assuming that the 3-surface V "extends to infinity" and that the current $\mathbf{e}$ vanishes sufficiently quickly at infinity so that all the boundary terms vanish, we find that the quantity

$$\tilde{E} \equiv \int_V (e\circ\mathbf{v})\,\mathrm{d}^3V \equiv \int_V e_\beta v^\beta \mathrm{d}^3V = \int_V (\Box k_\beta)\,v^\beta\mathrm{d}^3V$$

is independent of the choice of the 3-volume V. The quantity $\tilde{E}$ is the "conserved charge" that corresponds to the conserved current $\mathbf{e}$. Alternatively, $\tilde{E}$ can be written as an integral over a closed spacelike 2-surface Σ which is the boundary of V,

$$\tilde{E} = \oint_\Sigma (\nabla_\alpha k_\beta)\,v^\alpha n^\beta \mathrm{d}^2\Sigma, \tag{3.6}$$

where the vectors $\mathbf{v},\mathbf{n}$ are normal to the 2-surface Σ. The 2-surface Σ lies outside of any matter sources and encircles the entire matter distribution.

We shall now express $\tilde{E}$ through the matter distribution using the Einstein equation. It is again more convenient to work in the index notation. We have seen in Statement 1.8.4.1 that

$$\nabla_\mu\nabla_\nu k_\alpha = R_{\nu\alpha\mu}{}^\beta k_\beta.$$

Therefore

$$\begin{aligned}e_\beta &= \Box k_\beta = g^{\alpha\gamma}R_{\alpha\beta\gamma}{}^\delta k_\delta = R_{\beta\delta}k^\delta,\\ e\circ\mathbf{x} &= \mathrm{Ric}(\mathbf{x},\mathbf{k}),\\ \tilde{E} &= \int_V \mathrm{Ric}(\mathbf{v},\mathbf{k})\mathrm{d}^3V = \int_V R_{\mu\nu}v^\mu k^\nu \mathrm{d}^3V.\end{aligned}$$

[2] The fact that e is divergence-free can be also understood as a consequence of the identity $\mathrm{d}\mathrm{d}*B_{(\mathbf{k})} = 0$, where $*$ is the Hodge star operation. Thus the integral of the 3-form $\mathrm{d}*B_{(\mathbf{k})}$ over a 3-volume can be expressed through an integral of $*B_{(\mathbf{k})}$ over a closed 2-surface.

We shall now show that $\tilde{E}$ is proportional to the total energy E of the gravitating system in the Newtonian limit, where the gravitating matter is an isolated, stationary distribution of dust with density ρ. The energy of such a system is

$$E = \int_V \rho \mathrm{d}^3 V,$$

and the energy-momentum tensor is

$$T_{\mu\nu} = \rho v_\mu v_\nu,$$

where $\mathbf{v} \approx \mathbf{k}$ are the stationary worldlines of matter particles. By the Einstein equation (1.70), we have

$$R_{\mu\nu} = -8\pi(T_{\mu\nu} - \frac{1}{2} T g_{\mu\nu}),$$

therefore

$$R_{\mu\nu} v^\mu k^\nu = -4\pi\rho v_\mu v_\nu v^\mu k^\nu \approx -4\pi\rho$$

and

$$\tilde{E} = \int_V R_{\mu\nu} v^\mu k^\nu \mathrm{d}^3 V \approx -4\pi \int_V \rho \mathrm{d}^3 V = -4\pi E.$$

Thus the total energy of the system is, in the Newtonian approximation,

$$E \approx -\frac{1}{4\pi}\tilde{E}.$$

We are therefore motivated to *define* the total energy contained in a stationary spacetime by

$$E = -\frac{1}{4\pi}\tilde{E} = -\frac{1}{4\pi}\int_V (\Box k_\beta)\, v^\beta \mathrm{d}^3 V.$$

Calculation: Compute the total energy of the Schwarzschild spacetime with the metric (1.38). The Killing vector is $\mathbf{k} \equiv \partial_t$.

Solution: It is convenient to use Eq. (3.6) and integrate over a 2-sphere Σ of a large radius R. Choosing the normal vectors $\mathbf{v} = z^{-1}\mathbf{k}$ and $\mathbf{n} = z\partial_r$, we have

$$E = -\frac{1}{4\pi}\oint_\Sigma g(\nabla_{\mathbf{v}}\mathbf{k}, \mathbf{n}) \mathrm{d}^2\Sigma = -\frac{1}{4\pi}\oint_\Sigma g(\nabla_{\mathbf{k}}\mathbf{k}, \partial_r) \mathrm{d}^2\Sigma.$$

Since (see Sec. 3.1.4)

$$z = z(r) = \sqrt{1 - \frac{2m}{r}},$$
$$g(\nabla_{\mathbf{k}}\mathbf{k}, \partial_r) = -\frac{1}{2}\partial_r z^2 = -\frac{m}{r^2},$$

we find

$$E = \frac{1}{4\pi}\oint_\Sigma \frac{m}{r^2}\mathrm{d}^2\Sigma = m.$$

Thus the total energy of the Schwarzschild spacetime is equal to the mass m of the source of gravity.

Remark: The total momentum and the total angular momentum of an isolated system are so far undefined. (There is no preferred reference frame with respect to which one would measure the total momentum or angular momentum.) A satisfactory definition of these quantities requires more information than the existence of a timelike Killing vector. For instance, symmetry with respect to rotations around an axis leads to conservation of the corresponding component of the angular momentum. In a curved spacetime, there are no preferred axes, so instead the rotational symmetry is formulated as the existence of a spacelike Killing vector with closed orbits (these orbits play the role of "circles of constant r, θ around the z axis"). In a stationary spacetime there exists also a timelike Killing vector $\mathbf{k}$. If these two Killing vectors commute, the spacetime is called **azimuthally symmetric**. In such a spacetime, the total angular momentum is well-defined and conserved. In the Newtonian limit, this quantity will correspond to the total angular momentum with respect to the z axis.

3.2 Conformal infinity

In the previous section, we considered a stationary spacetime where the total energy of an isolated system can be defined through an integral over an infinitely remote 2-surface [see Eq. (3.6)]. It appears that such a definition only requires that a spacetime be "asymptotically stationary," that is, almost stationary sufficiently far from the matter sources. In fact, it would be convenient if we could evaluate the 2-surface integral "at infinity" where the spacetime is "exactly" stationary (in a heuristic sense). The notion of conformal infinity allows one to make these ideas precise. We shall approach this notion by considering the elementary example of a flat Minkowski spacetime.

3.2.1 Conformal infinity for Minkowski spacetime

We shall be able to evaluate quantities "at infinity" if we extend the spacetime manifold $\mathcal{M}$ by adding "points at infinity" to it. A convenient way to produce such an "extended" spacetime manifold is by using a conformal transformation of the metric, $g \to \tilde{g} = e^{2\lambda} g$, such that the conformal factor $e^{2\lambda}$ vanishes at infinity. Then the infinitely remote points can be located at finite distances according to the new metric $\tilde{g}$. Of course, the new metric $\tilde{g}$ is *unphysical* since it does not describe actual distances or time intervals in the spacetime. However, the metric $\tilde{g}$ is related to the physical metric g by a conformal transformation, which preserves important causal properties of the spacetime. Here we describe such a conformal transformation for the Minkowski metric; later, we will consider more general spacetimes.

Consider the **lightcone coordinates** $\{u, v, \theta, \phi\}$ in the Minkowski spacetime, which are defined through the standard spherical coordinates $\{t, r, \theta, \phi\}$ via

$$u \equiv t + r, \quad v \equiv t - r.$$

The **radial lightrays** are then lines of constant u or v. The metric has the form

$$g = du\, dv - \frac{(u-v)^2}{4} dS^2, \quad dS^2 \equiv d\theta^2 + \sin^2\theta d\phi^2.$$

A conformal transformation of the metric,

$$g \to \tilde{g} = e^{2\lambda(u,v)} g, \tag{3.7}$$

where $\lambda(u, v)$ is an arbitrary function of u and v (but not of θ, ϕ), leaves the radial lightrays invariant (see Statement 3.2.1.1). The transformed (unphysical) metric can be written as

$$\tilde{g} = d\tilde{u}\, d\tilde{v} - e^{2\lambda}\frac{(u-v)^2}{4} dS^2 \equiv d\tilde{u}\, d\tilde{v} - \tilde{r}^2 dS^2,$$

where $\tilde{u}, \tilde{v}$ are new lightcone coordinates,

$$\tilde{u} = f_1(u), \quad \tilde{v} = f_1(v).$$

For simplicity, let us choose $f_1(s) = f_2(s) = f(s)$. (Choosing more complicated functions for $\tilde{u}$ and $\tilde{v}$ is certainly admissible but will make our work harder.) Then we have

$$e^{2\lambda} = f'(u)f'(v); \quad \tilde{r}^2 = f'(u)f'(v)\frac{(u-v)^2}{4}.$$

It is obvious that the lines of constant u are at the same time lines of constant $\tilde{u}$. Thus, the radial lines of constant u are null generators of a 3-surface $u = \text{const}$, which are always geodesics (see Sec. 2.2.8). In this case, the function $\lambda(u,v)$ has a particularly simple form; the following statement shows that radial lightrays remain geodesics in the metric $\tilde{g} = e^{2\lambda(u,v)}g$ even with an *arbitrary* function $\lambda(u,v)$.

Statement 3.2.1.1: **(a)** In a two-dimensional spacetime, *any* null curve is a geodesic curve. **(b)** The radial null geodesics (lightrays) are invariant under a conformal transformation of the form (3.7), where g is a flat metric. (Proof on page 179.) ■

The next important step is the following: the function f can be chosen such that the new lightcone coordinates $\tilde{u}, \tilde{v}$ have a *finite range*. There are many possible choices of f; for example,

$$f(s) = \frac{2}{\pi}\arctan s; \quad f(s) = \tanh s; \quad f(s) = \operatorname{erf} s;$$

etc. With any of the above choices, $f(s)$ tends to ± 1 as $s \to \pm\infty$. So at large values of u, v the unphysical metric $\tilde{g}$ is flat on the sections of constant angles ϕ, θ and describes the points $\tilde{u}, \tilde{v} = \pm 1$ as being at finite (unphysical) distances, although physically these points correspond to infinitely remote events $u, v = \pm\infty$. For convenience, one can introduce the coordinates $\{\tilde{t}, \tilde{r}, \theta, \phi\}$ by defining

$$\tilde{u} \equiv \tilde{t} + \tilde{r}, \quad \tilde{v} \equiv \tilde{t} - \tilde{r}.$$

The $\{\tilde{t}, \tilde{r}\}$ (**time-radial**) section of the spacetime (with the angular coordinates omitted) is sketched in Fig. 3.1.

As we described above, the unphysical manifold $\tilde{\mathcal{M}}$ is constructed from the original spacetime manifold $\mathcal{M}$ by changing the metric from g to $\tilde{g}$ and adding the "infinitely far away" domains $\tilde{u} = \pm 1$, $\tilde{v} = \pm 1$. Thus $\tilde{\mathcal{M}}$ is a *manifold with boundary*. The boundary of $\tilde{\mathcal{M}}$ consists of null lines (future null infinity and past null infinity), the spacelike infinity domain i^0, and timelike future infinity/past infinity points $i^{\pm}$. Note that the conformal factor $e^{2\lambda}$ vanishes on the boundary; this is what brings infinity to finite distances in the unphysical manifold.

To see that the choice of the function $f(s)$ is not entirely arbitrary, consider the spacelike infinity domain i^0 corresponding to $r \to \infty$ ($u \to \infty, v \to -\infty$). This domain is formally a 2-sphere of radius

$$\tilde{r}^2(i^0) = \lim_{r\to\infty} e^{\lambda} r^2 = \lim_{\substack{v\to-\infty \\ u\to\infty}} f'(u)f'(v)\frac{(u-v)^2}{4}.$$

Since $f(s) \to \pm 1$ as $s \to \infty$, the derivative $f'(s)$ must go to zero at large s. Suppose that $f'(s)$ tends to zero as $\sim s^{-n}$ for $s \to \infty$; then the limit $\tilde{r}^2(i^0)$ depends on the value of n. If $n \geq 2$ then the limit is zero, meaning that the domain of a "spacelike infinity" is a sphere of zero radius, i.e. a *point*. In the opposite case, $n < 2$, the limit is infinite, which means that the metric $\tilde{g}$ diverges (is not regular) at the point i^0. Thus we are motivated to consider only the regular case $n \geq 2$.

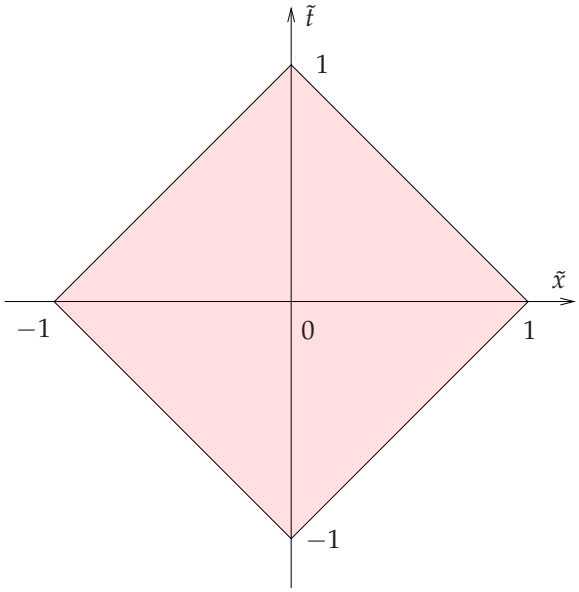

Figure 3.1: The time-radial section of the unphysical manifold $\tilde{\mathcal{M}}$ for the Minkowski spacetime. The coordinates $\{\tilde{t}, \tilde{x}\}$ have finite ranges. The boundary of the diagram represents points "at infinity."

We may restrict the choice of the function $f(s)$ further. Let us consider the limit of the metric $\tilde{g}$ along a fixed lightray, say $u = u_0$, $v \to \infty$. The coefficient $\tilde{r}^2$ of the metric is then

$$\tilde{r}^2 = \lim_{v\to\infty} f'(u_0)f'(v)\frac{(u_0-v)^2}{4}.$$

This limit is equal to zero when $n > 2$ but is finite and nonzero when $n = 2$. For $f'(s) \sim f_0 s^{-2}$, we find

$$\tilde{r}^2(\mathscr{I}^+) = \frac{1}{4}f_0 f'(u_0) > 0.$$

We expect that the structure of the null infinity should reflect the fact that null rays can be emitted in all directions spanning a 2-sphere. Thus, the null infinity at a fixed value of $\tilde{u} = f(u_0)$ should have a metric structure of a 2-sphere with a nonzero radius. This is the case only if $f'(s) \propto s^{-2}$ for $s \to \pm\infty$. Hence, we shall only consider these choices of $f(s)$. Then the future null infinity, denoted $\mathscr{I}^+$ ("scri-plus"), has the topological structure $\mathbb{R} \times S^2$ (a 2-sphere for each value $\tilde{u} = \alpha$), while a spacelike infinity is represented by a single point i^0. The asymptotic form of the unphysical metric near a point $(u = u_0, v = \infty)$ on $\mathscr{I}^+$ is

$$\tilde{g} \approx d\tilde{u}\, d\tilde{v} - \frac{f_0 f'(u_0)}{4} dS^2,$$

while the conformal factor Ω near $\mathscr{I}^+$ is

$$\Omega \approx \frac{1}{f_0}(1-\tilde{u})(1-\tilde{v}).$$

Therefore, $\Omega \sim r^{-1}$ at large distances. From the above expression for Ω, it is also easy to see that the null infinity $\mathscr{I}^+$ is a null surface: The normal vector $\mathbf{n} \equiv \hat{\tilde{g}}^{-1}\mathrm{d}\Omega$ is null on $\mathscr{I}^+$ (but not away from it!) since

$$\tilde{g}(\mathbf{n}, \mathbf{n}) = \tilde{g}^{-1}(\mathrm{d}\Omega, \mathrm{d}\Omega) \propto (1-\tilde{u})(1-\tilde{v}) \propto \Omega = 0 \text{ on } \mathscr{I}^+.$$

Note that there still remains a considerable freedom in choosing the conformal transformations. However, within the

present scheme every lightray will always be represented in a diagram by a straight line drawn at 45° angles, and the null future boundary $\mathscr{I}^+$ will be a null surface.

Calculation: Show that the Riemann tensor has only one independent component in a 2-dimensional spacetime. Then use the following expression for the curvature scalar in the unphysical metric,

$$\tilde{R} = \Omega^{-2} R + 6\Omega^{-3} \Box \Omega, \tag{3.8}$$

to show that any 2-dimensional metric can be transformed into a *flat* metric by a choice of a conformal factor Ω^2. (The above equation is derived e.g. in [36] and [9]. See also Statement 1.8.4.1 on page 48.)

Solution: The Riemann tensor $R(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d})$ is a symmetric bilinear function of $\mathbf{a} \wedge \mathbf{b}$ and $\mathbf{c} \wedge \mathbf{d}$; for a 2-dimensional vector space with a basis $\{\mathbf{e}_1, \mathbf{e}_2\}$, any exterior product $\mathbf{x} \wedge \mathbf{y}$ is proportional to the basis product $\mathbf{e}_1 \wedge \mathbf{e}_2$, so the space of exterior products $\mathbf{x} \wedge \mathbf{y}$ is one-dimensional. Thus $R(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d})$ is effectively a symmetric bilinear function in one-dimensional space. Such a function is fully specified by one coefficient, say $R(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_1, \mathbf{e}_2)$, or equivalently by the curvature scalar $R = 2R(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_1, \mathbf{e}_2)$. It is clear from the given relation that the curvature scalar can be made to vanish by a choice of the function Ω. Then the unphysical spacetime will have an identically vanishing Riemann tensor, i.e. it will be flat.

3.2.2 Conformal diagrams

In the previous section we explored the construction of a conformal infinity for Minkowski spacetime. The result was an unphysical 4-dimensional manifold $\tilde{\mathcal{M}}$ containing all of $\mathcal{M}$ and additionally some "infinitely remote points." The metric on $\tilde{\mathcal{M}}$ was "unphysical" since it did not represent actual distances. However, this unphysical manifold allows one to evaluate quantities "at infinity" and hence is useful for the analysis of asymptotic behavior of fields in the real spacetime.

Another important application of the unphysical manifold $\tilde{\mathcal{M}}$ is as a tool to help visualize the causal structure of the physical spacetime $\mathcal{M}$. Such a visualization is possible if one considers a 1+1-dimensional slice of the physical spacetime. This slice is a 1+1-dimensional spacetime $\mathcal{M}_{(1+1)}$ having a certain "reduced" metric. This metric can be transformed into a flat Minkowski metric by a conformal transformation. Then, a coordinate transformation can be performed such that the manifold $\mathcal{M}_{(1+1)}$ is mapped into into a 1+1-dimensional unphysical manifold $\tilde{\mathcal{M}}$ having a finite extent in all directions. The resulting manifold $\tilde{\mathcal{M}}$ can be simply drawn as a figure on a plane. (Implicitly, the plane carries the unphysical flat Minkowski metric; in particular, straight lines drawn at 45° angles are null geodesics.) An example of such a drawing is Fig. 3.1. Figures of this kind are called **conformal diagrams** (also **Carter-Penrose diagrams**). We now explore conformal diagrams in more detail.[3]

A conformal diagram provides a visualization of the causal structure of a spacetime because the null geodesics remain straight lines after the conformal transformation, and because the entire spacetime is drawn as a finite figure in a plane. For instance, it is easy to see the entire domain that can receive signals from a given spacetime point.

Remark: There is no analog of conformal diagrams for general 3+1-dimensional spacetimes. The reason is that a general spacetime is not conformally flat and cannot be mapped into a region of Minkowski spacetime by a conformal transformation. For instance, even if we demand that the spatial sections of the spacetime be flat and assume a metric of the form

$$g = \mathrm{d}t \otimes \mathrm{d}t - a^2(t, x_1, x_2, x_3) \sum_{i=1}^{3} \mathrm{d}x_i \otimes \mathrm{d}x_i,$$

the spacetime will be conformally flat only if $a(t, x_1, x_2, x_3)$ has the special form[4]

$$a(t, x_1, x_2, x_3) = \frac{1}{\eta(t) + \xi(t) \sum_{i=1}^{3} (x_i - \beta_i(t))^2},$$

where β_i, η, ξ are arbitrary functions of time. Nevertheless, in many cases a 3+1-dimensional spacetime can be adequately represented by a suitable 1+1-dimensional slice, at least for the purpose of qualitative illustration. For instance, a spherically symmetric spacetime is visualized using the time-radial half-plane $\{t, r\}, r \geq 0$, where each point stands for a 2-sphere of radius r. A conformal diagram is then drawn for the reduced 1+1-dimensional spacetime.

The reduced conformal diagram is meaningful only if the null geodesics in the 1+1-dimensional section are also geodesics in the physical 3+1-dimensional spacetime. In this case, the 1+1-dimensional section can be visualized as the set of events accessible to an observer who sends and receives signals only along a fixed spatial direction. Then a conformal diagram provides information about the causal structure of spacetime along this line of sight. ■

After recapitulating the standard construction of conformal diagrams, I develop an easier method.

Standard procedure

A conformal diagram is defined for a 1+1-dimensional spacetime with a given line element $g_{ab}dx^a dx^b$. The coordinates $\{x^a\}$ (where $a = 0, 1$) must cover the *entire* spacetime, and several sets of overlapping coordinate patches may be used if necessary. The standard construction of the conformal diagram may be formulated as follows (see e.g. **[?]**, chapter 3). One first finds a change of coordinates $x \to \tilde{x}$ such that the new variables $\tilde{x}^a$ have a *finite* range of variation; the components of the metric change according to $g_{ab}(x)dx^a dx^b = \tilde{g}_{ab}(\tilde{x})d\tilde{x}^a d\tilde{x}^b$. One then chooses a conformal transformation of the metric (in the new coordinates),

$$\tilde{g}_{ab} \to \gamma_{ab}(\tilde{x}) = \Omega^2(\tilde{x})\, \tilde{g}_{ab}, \quad \Omega(\tilde{x}) \neq 0, \tag{3.9}$$

such that the new metric γ_{ab} is flat, i.e. has zero curvature. A suitable function $\Omega(\tilde{x})$ always exists because all two-dimensional metrics are conformally flat. The new metric γ_{ab} describes an unphysical, auxiliary flat spacetime. Since the metric γ_{ab} is flat, a further change of coordinates $\tilde{x} \to \tilde{\tilde{x}}$ (the new variables $\tilde{\tilde{x}}$ again having a finite extent) can be found to transform γ_{ab} explicitly into the Minkowski metric η_{ab},

$$\gamma_{ab}(\tilde{x})\, d\tilde{x}^a d\tilde{x}^b = \eta_{ab} d\tilde{\tilde{x}}^a d\tilde{\tilde{x}}^b, \quad \eta_{ab} \equiv \mathrm{diag}(1, -1). \tag{3.10}$$

For brevity we incorporate all the required coordinate changes into one, $x \to \tilde{x}(x)$, and summarize the transformation of the metric as

$$\Omega^2(x)\, g_{ab} dx^a dx^b = \eta_{ab} d\tilde{x}^a d\tilde{x}^b. \tag{3.11}$$

[3] This and the following sections are adapted from the paper [38].

[4] The author is grateful to Matthew Parry for figuring this out.

Thus the new coordinates $\tilde{x}^a$ map the initial spacetime onto a *finite* domain within a 1+1-dimensional Minkowski plane. This finite domain is a **conformal diagram** of the initial (physical) 1+1-dimensional spacetime. The diagram is drawn on a sheet of paper which implicitly carries the fiducial Minkowski metric η_{ab}, the vertical axis usually representing the timelike coordinate $\tilde{x}^0$.

The coordinate and conformal transformations severely distort the geometry of the spacetime since they bring infinite spacetime points to finite distances in the diagram plane. So one cannot expect in general that straight lines in the diagram correspond to geodesics in the physical spacetime. However, it is well known that straight lines drawn at 45° angles in a conformal diagram represent null geodesics in the physical spacetime. This follows from the fact that any null trajectory $x^a(\tau)$ in 1+1 dimensions, i.e. any solution of

$$g_{ab}\dot{x}^a\dot{x}^b = 0, \quad \dot{x}^a \equiv \frac{dx^a}{d\tau}, \tag{3.12}$$

is necessarily a geodesic (this is not true in higher dimensions), and Eq. (3.12) is invariant under conformal transformations of the metric. By drawing lightrays emitted from various points in the diagram, one can illustrate the causal structure of the spacetime.

A textbook example is the conformal diagram for the flat Minkowski spacetime with the metric $g_{ab} \equiv \eta_{ab}$. Calculations are conveniently done in the lightcone coordinates

$$u \equiv x^0 - x^1, \; v \equiv x^0 + x^1, \; \eta_{ab}dx^a dx^b = du\, dv, \tag{3.13}$$

and a suitable coordinate transformation is

$$\tilde{u} = \tanh u, \; \tilde{v} = \tanh v, \; du\, dv = \frac{d\tilde{u}\, d\tilde{v}}{(1-\tilde{u}^2)\,(1-\tilde{v}^2)}. \tag{3.14}$$

The new coordinates $\tilde{u}, \tilde{v}$ extend from -1 to 1. Multiplying the metric by the conformal factor $\Omega^2(\tilde{u},\tilde{v}) \equiv (1-\tilde{u}^2)\,(1-\tilde{v}^2)$, we obtain the fiducial spacetime,

$$\Omega^2 du\, dv = d\tilde{u}\, d\tilde{v} = \eta_{ab}d\tilde{x}^a d\tilde{x}^b, \tag{3.15}$$

$$\tilde{u} \equiv \tilde{x}^0 - \tilde{x}^1, \quad \tilde{v} \equiv \tilde{x}^0 + \tilde{x}^1. \tag{3.16}$$

The new coordinates $\tilde{x}^a$ have a finite extent, namely $\left|\tilde{x}^0 \pm \tilde{x}^1\right| < 1$, and the resulting diagram has a diamond shape shown in Fig. 3.2. To appreciate the distortion of the spacetime geometry, we can draw the worldline of an inertial observer moving with a constant velocity. Note that the angle at which this trajectory enters the endpoints depends on the chosen conformal transformation and thus cannot serve as an indication of the observer's velocity.

There is no analog of conformal diagrams for general 3+1-dimensional spacetimes. Nevertheless, in many cases a 3+1-dimensional spacetime can be adequately represented by a suitable 1+1-dimensional slice, at least for the purpose of qualitative illustration. For instance, a spherically symmetric spacetime is visualized as the (t,r) half-plane ($r \geq 0$) where each point stands for a 2-sphere of radius r. A conformal diagram is then drawn for the reduced 1+1-dimensional spacetime.

The reduced conformal diagram is meaningful only if the null geodesics in the 1+1-dimensional section are also geodesics in the physical 3+1-dimensional spacetime. In this case, the 1+1-dimensional section can be visualized as the set of events accessible to an observer who sends and receives signals only along a fixed spatial direction. Then a conformal diagram provides information about the causal structure of spacetime along this line of sight.

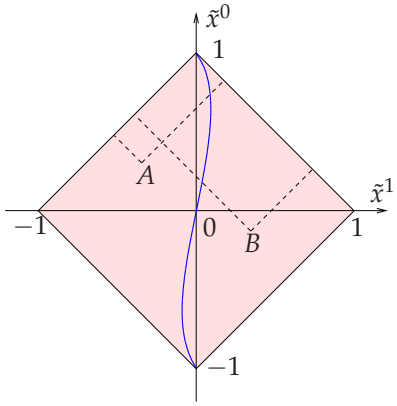

Figure 3.2: A conformal diagram of the 1+1-dimensional Minkowski spacetime. Dashed lines show lightrays emitted from points A, B. The curved line is the trajectory of an inertial observer moving with a constant velocity, $x^1 = 0.3x^0$.

Method of lightrays

The standard construction of conformal diagrams involves an explicit transformation of the metric to new coordinates that have a finite extent, and usually a further transformation to bring the metric to a manifestly conformally flat form. Finding these transformations requires a certain ingenuity. If the spacetime manifold is covered by several coordinate patches, a different transformation must be used in each patch. However, a conformal diagram typically consists of just a few lines and one would expect that the required computations should not be so cumbersome.

I now describe a method of drawing conformal diagrams that avoids the need for performing explicit transformations of the metric. The method is based on a qualitative analysis of intersections of lightrays. This approach is particularly suitable for the analysis of stochastic spacetimes encountered in models of eternal inflation. Such spacetimes have no symmetries and their metric is not known in closed form, so one cannot apply the standard construction of conformal diagrams.

Another motivation for the new method is the apparent redundancy involved in the standard method. It is clear that the transformations used in the standard construction are not unique. For instance, one may replace the lightcone coordinates in Eq. (3.14) by

$$\tilde{u} \to f(\tilde{u}), \quad \tilde{v} \to g(\tilde{v}), \tag{3.17}$$

where f, g are arbitrary monotonic, bounded, and continuous functions. The shape of the diagram will vary with each possible choice of the transformations, but all resulting diagrams are equivalent in the sense that they contain the same information about the causal structure of the spacetime. One thus expects to be able to extract this information without involving specific explicit transformations of the coordinates and the metric.

The crucial observation is that this information is unambiguously represented by the geometry and topology of lightrays and their intersections. I shall now develop this idea into

a self-contained approach to drawing conformal diagrams that does not involve explicit transformations.

New definition of conformal diagrams

A conformal diagram is a figure in the fiducial Minkowski plane satisfying certain conditions, and I first formulate a definition of a conformal diagram in terms of such conditions. A constructive procedure for drawing conformal diagrams will be presented subsequently.

A finite open domain of the plane is a **conformal diagram** of a given 1+1-dimensional spacetime S if there exists a one-to-one correspondence between all maximally extended lightrays in S and all straight line segments drawn at 45° angles within the domain of the diagram. This correspondence must be **intersection-preserving**, i.e. any two lightrays intersect in the physical spacetime exactly as many times as the corresponding lines intersect in the diagram. It is assumed that all null geodesics in the spacetime S are either infinitely extendible or end at singularities or at explicitly introduced spacetime boundaries. Similarly, the straight line segments drawn at 45° angles in a conformal diagram must be limited only by the boundary of the diagram. Note that there are only two spatial directions in a 1+1-dimensional spacetime S, and that two lightrays emitted in the same direction cannot intersect.

For example, the diamond $\left|\tilde{x}^0 \pm \tilde{x}^1\right| < 1$ is a conformal diagram for the flat spacetime due to the intersection-preserving one-to-one correspondence of null lines $\tilde{x}^0 \pm \tilde{x}^1 = \text{const}$ in the diagram and lightrays $x^0 \pm x^1 = \text{const}$ in the physical spacetime.

For spacetimes having a nontrivial topology, appropriate topological features need to be introduced also into the fiducial Minkowski plane. At this point I do not consider such cases.

I shall now demonstrate the equivalence of the proposed definition to the standard procedure for drawing conformal diagrams. It suffices to find a conformal transformation of the form (3.11) in some neighborhood of an arbitrary (nonsingular) spacetime point. Given a diagram with an intersection-preserving correspondence of lightrays, we can introduce local lightcone coordinates u, v in the diagram such that the null geodesics are locally the lines $u = \text{const}$ or $v = \text{const}$. By assumption, each null geodesic uniquely corresponds to a lightray in the physical spacetime. Since the correspondence is intersection-preserving, the local configuration of the null geodesics in the physical spacetime can be visualized as in Fig. 3.3. Hence the local lightcone coordinates u, v become well-defined local coordinates in the physical spacetime, and again the null geodesics are the lines $u = \text{const}$ or $v = \text{const}$. On the other hand, these null geodesics must be solutions of Eq. (3.12), therefore

$$g_{uu}\dot{u}^2 + 2g_{uv}\dot{u}\dot{v} + g_{vv}\dot{v}^2 = 0 \text{ if } \dot{u} = 0 \text{ or } \dot{v} = 0. \qquad (3.18)$$

It follows that $g_{uu} = g_{vv} = 0$. Thus the metric in the local lightcone coordinates is of the form $g_{ab}dx^a dx^b = 2g_{uv}(u,v)du\,dv$ which is explicitly conformally flat. This demonstrates the existence of a local conformal transformation bringing the physical metric g_{ab} into the fiducial Minkowski metric $du\,dv = \eta_{ab}d\tilde{x}^a d\tilde{x}^b$ in the diagram.

Before presenting examples, I comment on the proposed definition of conformal diagrams. The definition may appear to be too broad, allowing many geometric shapes to represent the same spacetime. However, the old procedure also does not specify a particular conformal transformation of the metric and in effect admits precisely as much freedom. According to the new definition, any two different conformal diagrams of the same spacetime are equivalent in the sense that all the lightrays in those two diagrams will be in an intersection-preserving one-to-one correspondence. In the old language, there must exist a conformal transformation bringing one diagram into the other. Equivalent diagrams carry identical information about the causal structure of the physical spacetime. In the next sections I shall give examples illustrating the arbitrary and the necessary choices involved in drawing conformal diagrams.

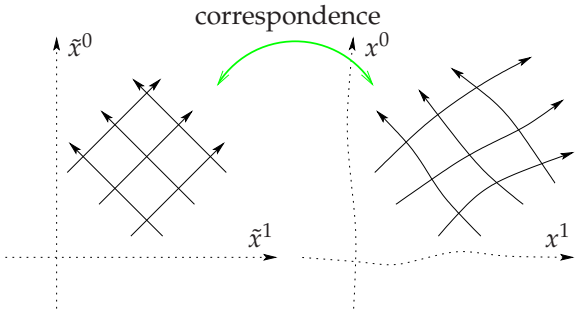

Figure 3.3: The local correspondence between straight lines drawn at 45° angles in the conformal diagram (left) and lightrays in the physical spacetime (right).

A conformal diagram delivers information mainly through the shape of its boundary line. The boundary of a conformal diagram generally contains points representing a spacelike, timelike, or null infinity, and points belonging to explicit boundaries of the spacetime manifold (e.g. singularities) where lightrays end in the physical spacetime. The latter boundaries will be called **physical boundaries** to distinguish them from putative **infinite boundaries** whose points do not correspond to any events in the physical spacetime. (The definition of conformal diagrams contains the requirement that the diagram domain be topologically open, and so the boundary points are not supposed to belong to the diagram.) The infinite boundary is of course the most interesting feature of a diagram.

Lastly, I would like to emphasize that conformal diagrams can be drawn not only for geodesically complete spacetimes but also for "artificially incomplete" spacetimes, i.e. for selected subdomains of larger manifolds. In fact such "artificially incomplete" spacetimes are often needed in cosmological applications. Examples are a description of a collapsing star using a subdomain of the Kruskal spacetime and a description of an inflationary universe using a subdomain of the de Sitter spacetime.

Minkowski spacetime

The new definition merely lists the conditions to be satisfied by a conformal diagram. Based on these conditions, I now develop a practical procedure for drawing the diagrams, using the Minkowski spacetime as the first example.

We begin by choosing a Cauchy surface in the physical spacetime. In 1+1 dimensions, a Cauchy surface is a line L such that intersections of lightrays emitted from L entirely cover the part of the spacetime to the future of L. In the Minkowski spacetime, we may choose the line $x^0 = 0$ as the

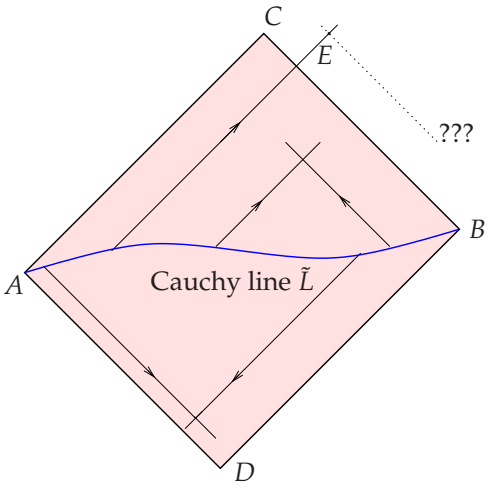

Figure 3.4: Construction of the conformal diagram for the Minkowski spacetime. The point E cannot belong to the diagram because there are no lightrays intersecting at E.

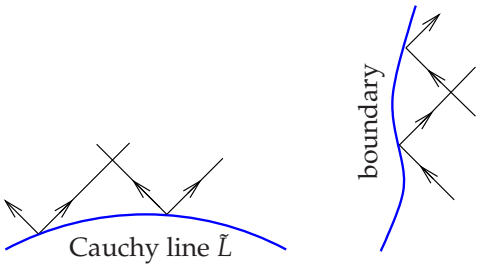

Figure 3.5: A Cauchy line must have a slope between $-45°$ and $45°$ (left). A timelike boundary must be a curve with a slope between $45°$ and $135°$ (right).

Cauchy surface L. The image of the Cauchy surface L in the conformal diagram must be a finite curve $\tilde{L}$ from which lightrays can be emitted in both spatial directions. Therefore the slope of the curve $\tilde{L}$ must not exceed $\pm 45°$ but otherwise $\tilde{L}$ may be drawn arbitrarily, e.g. as the curve AB in Fig. 3.4. The endpoints A, B represent a spatial infinity in the two directions.

In the Minkowski spacetime, two lightrays emitted towards each other from any two points on L will eventually intersect. Therefore the domain of the conformal diagram must contain at least the triangular region ABC. On the other hand, any point outside ABC, such as the point E in Fig. 3.4, cannot belong to the diagram domain because the point E cannot be reached by any left-directed lightray emitted from L, and we know that all points in the Minkowski spacetime are intersection points of some lightrays. (More formally, the existence of the point E within the diagram would violate the condition that the correspondence between lightrays is intersection-preserving.) Therefore the future-directed part of the conformal diagram is bounded by the lines AC and BC.

A completely analogous consideration involving past-directed lightrays leads to the conclusion that the past-directed part of the diagram is the region ABD. Thus a possible diagram for the Minkowski spacetime is the interior of the rectangle $ACBD$. This diagram differs from the square-shaped diagram in Fig. 3.2 by a (finite) conformal transformation of the form (3.17).

We can also ascertain that the points C and D are the future and the past timelike infinity points. For instance, the point C is the intersection of the lines AC and BC; these lines are interpreted as putative lightrays emitted from infinitely remote points of L. At sufficiently late times, any inertial observer in the Minkowski spacetime will catch lightrays emitted from arbitrarily far points. The same holds for observers moving non-inertially as long as their velocity does not approach that of light. Hence all trajectories of such observers must finish at C.

Cauchy surfaces and artificial boundaries

Drawing Cauchy surfaces is a convenient starting point in the construction of conformal diagrams.

A **Cauchy surface** for a domain of spacetime is a 3-surface S such that every timelike curve (without endpoints) within the region intersects S exactly once. (It follows that a Cauchy surface must be spacelike.)

It is clear that in any 1+1-dimensional spacetime a sufficiently small neighborhood of a Cauchy surface has the same causal properties as the line $x^0 = 0$ in the Minkowski plane: namely, two nearby lightrays emitted toward each other will cross, while rays emitted from a point in opposite directions will diverge. Therefore the line $\tilde{L}$ representing a Cauchy surface in a conformal diagram must have a slope between $-45°$ and $45°$ (see Fig. 3.5, left). We shall call such lines **horizontally-directed**. Other than this, there are no restrictions on the shape of the line $\tilde{L}$ and it may be drawn as an arbitrary horizontally-directed curve. (In some spacetimes, one needs to use several disconnected Cauchy surfaces, but I shall not consider such cases here.)

Another frequently occurring feature in conformal diagrams is a *timelike boundary*. For example, a spherically symmetric 3+1-dimensional spacetime is usually reduced to the 1+1-dimensional (r, t) plane, where $0 < r < +\infty$. From the 1+1-dimensional point of view, the line $r = 0$ is an artificially introduced timelike boundary that can absorb and emit lightrays. The local geometry of lightrays near $r = 0$ is shown in Fig. 3.5 (right). It is clear from the figure that in a conformal diagram the timelike boundary must be represented by a line with a slope between $45°$ and $135°$. Let us call such lines **vertically-directed**.

As an example of using timelike boundaries, let us consider the subdomain $(t_0 < t < \infty, x_1 < x < x_2)$ of a de Sitter spacetime with flat spatial sections, described by the metric

$$g_{ab}dx^a dx^b = dt^2 - e^{2Ht}dx^2. \tag{3.19}$$

This subdomain can be visualized as the future of a selected initial comoving region. The Cauchy line $t = t_0$ is connected to the two timelike boundary lines, $x = x_{1,2}$. In the coordinate system (3.19), the null geodesics are solutions of $dx/dt = \pm e^{-Ht}$ and it is easy to see that a lightray emitted at $x = 0$, $t = 0$ only reaches the values $|x| < H^{-1}$ (the limit value is the de Sitter horizon). Null geodesics emitted from the Cauchy surface and from the boundary lines are sketched in Fig. 3.6 where it is assumed that the comoving domain $x_1 < x < x_2$ contains several de Sitter horizons. A lightray emitted in the positive direction, such as the ray A, intersects left-directed lightrays emitted from the point B or from nearer points but does not intersect lightrays emitted further away, such as the ray C. We call the ray B the **rightmost ray** intersecting A. It is clear that the intersection of the correspond-

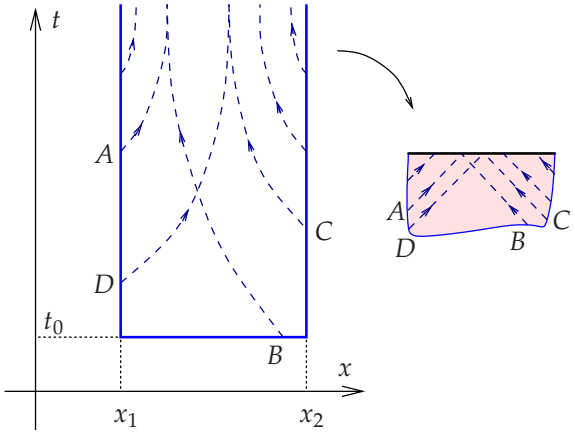

Figure 3.6: Construction of a conformal diagram for a part of de Sitter spacetime delimited by thick lines (left). The upper boundary of the diagram (right) is a horizontal line.

ing lines A and B in the conformal diagram must occur *at the boundary* of the diagram, otherwise there would exist further lines intersecting A to the right of B. Considering a ray D to the right of A, we find that the rightmost ray for D is C. The intersection of C and D is thus another point on the boundary of the conformal diagram. It follows that the boundary line must have a slope between $-45°$ and $45°$; for simplicity, we draw a straight horizontal line (Fig. 3.6, right). This line represents the (timelike and null) infinite future.

A timelike boundary can be interpreted as the trajectory of an observer who absorbs or emits lightrays and thus participates in the exploration of the causal structure of the spacetime. The lightrays emitted by the boundary, together with those emitted from the Cauchy surface, form the totality of all lightrays that must be bijectively mapped into straight lines in the conformal diagram. The role of timelike boundaries and Cauchy surfaces is to provide a physically motivated boundary for the part of the spacetime we are interested in.

An artificial timelike boundary may also be introduced into the spacetime with the purpose of simplifying the construction of the diagram. Below we shall show that conformal diagrams can be simply pasted together along a common timelike boundary.

3.2.3 How to draw conformal diagrams

We can now outline a general procedure for building a conformal diagram for a given 1+1-dimensional spacetime using the method of lightrays. The procedure can be applied not only to geodesically complete spacetimes but also to spacetimes with explicitly specified boundaries.

One starts by considering the future-directed part of the spacetime and by choosing a suitable Cauchy surface and, possibly, some timelike boundaries. These lines are represented in the diagram by arbitrarily drawn horizontal and vertical curves of finite extent. The endpoints of these curves correspond either to the intersection points of Cauchy lines and timelike boundaries, or to imaginary points at spacelike and timelike infinity.

Note that Cauchy surfaces and timelike boundaries are the lines on which boundary conditions for e.g. a wave equation must be specified to obtain a unique solution within a domain. One expects that any physically relevant spacetime should

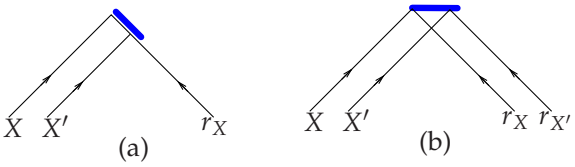

Figure 3.7: The local direction of the infinite boundary (thick line) is found by determining the rightmost rays $r_X, r_{X'}$ for nearby rays X, X'. The direction is at 45° angle (a) when $r_X = r_{X'}$ and horizontal (b) when $r_X \neq r_{X'}$.

contain a suitable set of Cauchy surfaces and timelike boundaries, if the classical field theory is to have predictive power. Qualitative knowledge of the geometry and topology of these boundaries is required for building a conformal diagram using the method of lightrays.

After drawing the curves for the physical boundaries, it remains to determine the shape of the infinite (timelike and null) boundaries of the diagram. To this end, one can first consider *right-directed* lightrays emitted from various points on the Cauchy line and from the boundaries, including the limit points at infinity. For each right-directed lightray X there exists a certain subset $\mathcal{S}_X$ of (left-directed) lightrays that intersect X. Since the subset $\mathcal{S}_X$ has a finite extent, there exists a rightmost ray $r_X \in \mathcal{S}_X$. (In the de Sitter example above, the rightmost ray r_A is the ray B and the rightmost ray r_D is C.) By definition of the conformal diagram, the lightrays are straight lines limited only by the boundary of the conformal diagram, therefore the intersection point of X and r_X must belong to the boundary. In this way we have established the location of one point of the unknown boundary line, namely the endpoint of the ray X.

To determine the local direction of the boundary line at that point, we use the following argument. For each right-directed lightray X, we can consider a right-directed ray X' infinitesimally close and to the right of X (if no ray X' can be found to the right of X, it means that X itself belongs to the boundary of the diagram). Then there are two possibilities (see Fig. 3.7a,b): either the rightmost ray r_X is also the rightmost ray $r_{X'}$ for X', or the ray $r_{X'}$ is located to the right of r_X. In the first case, the boundary line has a 45° slope and locally coincides with r_X, while in the second case the boundary line is horizontally-directed. Thus we can draw a right-directed fragment of the infinite boundary that limits the ray X. We then continue by moving further to the right and consider the endpoint of the ray X', etc.

The same procedure is then repeated for left-directed lightrays, until one finishes drawing all unknown lines in the future-directed part of the diagram. In this way the infinite boundary of the conformal diagram is constructed as the locus of "last intersections" of lightrays. Finally one applies the same considerations to past-directed lightrays and so completes the diagram.

It follows that the infinite boundary of a conformal diagram can always be drawn as a sequence of either straight line segments directed at 45° angles, or horizontally- and vertically-directed curves. In our convention, Cauchy surfaces and artificial timelike boundaries are drawn as curved lines and infinite boundaries as straight lines (when possible).

Let us briefly consider the pasting of conformal diagrams in general. When two spacetime domains are separated by

a timelike worldline, the corresponding conformal diagrams can be pasted together along the boundary line. To verify this almost obvious statement more formally, we begin by drawing the timelike boundary as a vertically-directed line in the diagram. The shape of the conformal diagram to the right of the boundary is determined solely by the intersections of lightrays within the right half of the spacetime. Hence, the conformal diagram for the right half of the spacetime can be pasted to the right of the timelike boundary line. The same holds for the left half of the spacetime. This justifies pasting of diagrams along a common timelike boundary.

Further examples

Collapsing star The method of lightrays does not require explicit formulae for the spacetime metric if the qualitative behavior of lightrays is known. As another example of using the lightray method, let us consider an asymptotically flat spacetime with a star collapsing to a black hole (BH).

To reduce the spacetime to 1+1 dimensions, we assume spherical symmetry and consider only the (r, t) plane, where $0 < r < +\infty$ and the line $r = 0$, the center of the star, is an artificial boundary. As before, we start with the future-directed part of the diagram. We must first choose a Cauchy surface; a suitable Cauchy surface is the line $t = t_0$ where t_0 is a time chosen before the collapse of the star. We represent the Cauchy surface by the curve AB in the diagram (Fig. 3.8). The point A corresponds to $(r = 0, t = t_0)$, while B is a spatial infinity $(r = \infty, t = t_0)$. The artificial boundary $r = 0$ is represented by the vertical line AC.

To investigate the shape of the future part of the diagram, we need to analyze the intersections of lightrays emitted from faraway points of the Cauchy surface. We know from qualitative considerations of the black hole formation that lightrays can escape from the star interior only until the appearance of the BH horizon. Shortly thereafter the star center becomes a singularity that cannot emit any lightrays. Hence, among all the rays emitted from the star center at various times, there exists a "last ray" not captured by the BH, while rays emitted later are captured. We arbitrarily choose a point E on the boundary line $r = 0$ to represent the emission of this "last ray." Any lightray emitted from $r = 0$ before E will propagate away from the BH and so will intersect all left-directed lightrays emitted from arbitrarily remote points of the Cauchy line AB. Since all the intersection points are outside of the BH horizon, it follows that the conformal diagram contains the polygon $AEFB$ which represents the spacetime outside the BH. The line FB is the infinite null boundary of the diagram, while the line EF is the BH horizon.

It is clear that the point C is the last point from which lightrays can be emitted from the star center and thus C is the beginning of the BH singularity. It remains to determine the shape of the diagram between the points C and F. We know that a lightray emitted from the center after E will be recaptured by the BH singularity and thus will not intersect with lightrays entering the BH horizon sufficiently late. For instance, the lightray emitted at E' will intersect with the lightray F' but not with a later ray F'', as shown in Fig. 3.8. Therefore the diagram boundary line connecting C and F is locally horizontally-directed. This line consists of final intersection points of lightrays emitted from the star center and those entering the BH horizon from outside. These final intersection points are located at the BH singularity which is therefore represented by the entire line CF.

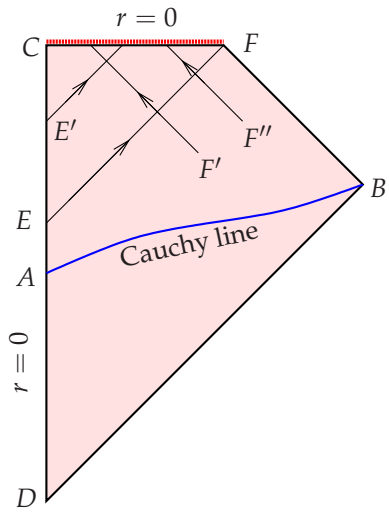

Figure 3.8: Construction of the conformal diagram for the spacetime of a collapsing star. The thick dotted line CF is a Schwarzschild singularity.

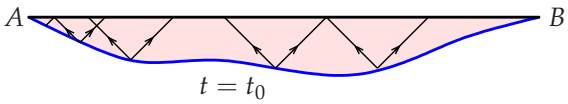

Figure 3.9: A conformal diagram for the future part of the de Sitter spacetime with flat spatial sections. The curved line represents the spatially infinite Cauchy surface $t = t_0$. The local behavior of lightrays is the same as in Fig. 3.6.

The past-directed part of the conformal diagram is easy to construct. Since any two past-directed lightrays intersect, the past-directed part is similar to that for the Minkowski spacetime with a boundary at $r = 0$, namely the triangular domain ABD. Thus the diagram in Fig. 3.8 is complete.

Future part of de Sitter spacetime In Fig. 3.6 we have drawn a conformal diagram for the subdomain of a de Sitter spacetime delimited by two timelike boundaries. We shall now construct the diagram for the future part of a de Sitter spacetime with spatially unlimited sections. (Note that the past half of the de Sitter spacetime is not covered by the flat coordinates because of incompleteness of past-directed geodesics. In this paper we shall not use the complete de Sitter spacetime but only the future of an arbitrarily chosen, unbounded, spacelike Cauchy hypersurface.)

The Cauchy surface $t = t_0$, $-\infty < x < \infty$ is drawn as a finite horizontal curve in the conformal diagram (the curved line AB in Fig. 3.9). The points A, B in the diagram represent a spacelike infinity in the two directions and do not correspond to any points in the physical spacetime. It remains to establish the shape of the infinite future boundary which must be a line connecting the points A, B to the future of the Cauchy surface. We already know from the construction of the diagram in Fig. 3.6 that this future boundary is locally horizontal. Since the behavior of lightrays emitted from all points of the Cauchy surface is the same, the future boundary may be represented by a horizontal straight line connecting the points A, B. (To keep the convention of having straight infinite boundaries, we have drawn the Cauchy surface $t = t_0$ as a curve extending downward from the straight line AB.)

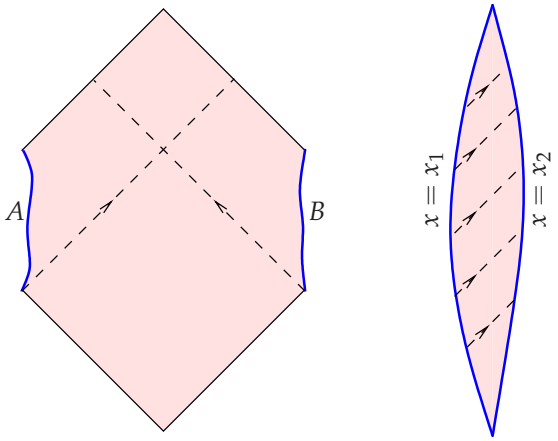

Figure 3.10: Left: The domain of Minkowski spacetime between two causally separated observers A, B moving with a constant proper acceleration in opposite directions. The dashed lines show that the two observers cannot see each other. Right: A closed Minkowski universe. The thick lines represent the identified boundaries $x = x_{1,2}$. A lightray (dashed line) crosses the line $x = x_{1,2}$ infinitely many times.

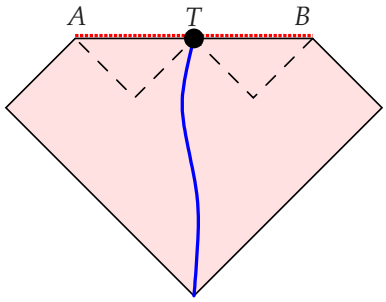

Figure 3.11: A spacetime with two collapsing stars. The point T is a future timelike infinity reached by an observer (thick curve) remaining between the two black holes. The lines AT and TB represent BH singularities and the dashed lines are the BH horizons.

Note that the same diagram also represents an inhomogeneous spacetime consisting of de Sitter-like regions with different values of the Hubble constant H. This is so because a difference in the local values of H does not change the qualitative behavior of lightrays at infinity: the rays will intersect only if emitted from sufficiently near points. The infinite future boundary remains a horizontally-directed line as long as $H \neq 0$ everywhere.

The construction of conformal diagrams for the following spacetimes is left as an exercise for the reader.

- A subdomain of the Minkowski spacetime between two causally separated observers moving with a constant proper acceleration in opposite directions (Fig. 3.10, left).
- A flat closed universe: the subdomain ($-\infty < t < \infty$, $x_1 < x < x_2$) of a Minkowski spacetime with the lines $x = x_1$ and $x = x_2$ identified (Fig. 3.10, right).
- An asymptotically flat spacetime with two stars collapsing into two black holes; the line of sight crosses the two star centers (Fig. 3.11).
- The future part of a de Sitter spacetime with a star collapsing into a black hole (Fig. 3.12).
- A maximally extended Schwarzschild-de Sitter spacetime (Fig. 3.13). It is interesting to note that this diagram is usually drawn with all Schwarzschild and de Sitter regions having the same size, which makes the figure unbounded (Fig. 3.13, top) despite the intention to represent the spacetime by a *finite* figure in the fiducial Minkowski plane. To adhere to the definition of a conformal diagram as a bounded figure, one can use a suitable conformal transformation reducing the diagram to a finite size (e.g. Fig. 3.13, bottom).

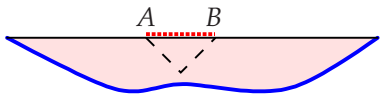

Figure 3.12: The future part of a de Sitter spacetime with a collapsing star. The line AB represents the BH singularity; the dashed line is the BH horizon.

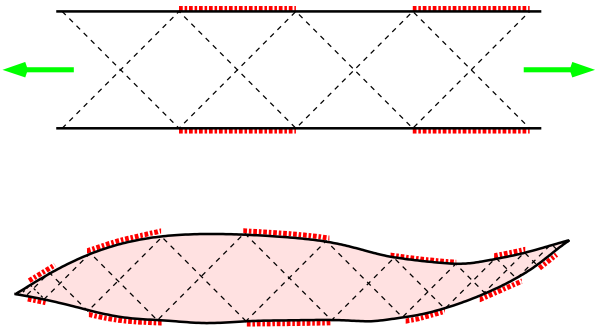

Figure 3.13: Diagrams for a maximally extended Schwarzschild-de Sitter spacetime. The conventional, unbounded diagram (top) and an equivalent diagram having a finite extent (bottom) are related by a conformal transformation. The thick dotted lines represent Schwarzschild singularities. Both diagrams contain infinitely many Schwarzschild and de Sitter regions.

3.3 Asymptotic flatness

Additional literature:
[23] A concise derivation of the peeling property.
[9] Good conceptual explanations of the asymptotic description of spacetimes in GR.
[7] A recent review.

The concept of an asymptotically flat spacetime was invented to formalize the picture of a spacetime which is "similar to Minkowski near infinity." Informally, a spacetime is asymptotically flat if it can be extended via a conformal transformation to include the infinity domain, and then if the structure of the infinity is the same as that of the Minkowski spacetime. A definition can be given more formally as follows:

A spacetime manifold $\mathcal{M}$ is **asymptotically flat** in the null future direction if:

1. There exists an extension of $\mathcal{M}$ to an (unphysical) manifold $\tilde{\mathcal{M}}$ with an (unphysical) regular metric $\tilde{g}$ and boundary $\mathscr{I}^+$, such that $\tilde{g} = \Omega^2 g$, where Ω is a smooth function such that $\Omega = 0$ on the boundary, but $\tilde{\nabla}\Omega \neq 0$ everywhere (we denote by $\tilde{\nabla}$ the covariant derivative with respect to the unphysical metric).

2. The boundary $\mathscr{I}^+$ is a null surface (in the unphysical metric), i.e. Ω is a null function on the boundary.

3. The boundary surface $\mathscr{I}^+$ has topology $\mathbb{R} \times S^2$.

4. The vector $-\tilde{\nabla}\Omega$ is future-pointing at $\mathscr{I}^+$ (as appropriate for the future boundary).

5. The energy-momentum tensor of matter (equivalently, the Ricci tensor) vanishes in a neighborhood of $\mathscr{I}^+$. (This condition may be relaxed to vanishing only as Ω^4 or faster.)

From these conditions, we shall now derive the existence of Minkowski-like coordinates "at large distances." Let $\mathbf{n}$ be the null vector dual to $-d\Omega$. First, we note that $\tilde{g}(\mathbf{n},\mathbf{n}) = 0$ on $\mathscr{I}^+$ implies that the auxiliary function $f \equiv \Omega^{-1}\tilde{g}(\mathbf{n},\mathbf{n})$ is smooth on $\tilde{M}$. Secondly, there is a freedom to rescale the unphysical metric by a nowhere vanishing factor $e^{2\lambda}$. Under such a rescaling, the function f changes as

$$f \to e^{-\lambda}\left(f + 2\tilde{\nabla}_{\mathbf{n}}\lambda\right) \quad \text{on } \mathscr{I}^+. \tag{3.20}$$

Therefore, a suitable choice of λ will make f vanish on $\mathscr{I}^+$. There will remain the freedom to perform conformal transformations $\tilde{g} \to e^{2\lambda}\tilde{g}$ with $\tilde{\nabla}_{\mathbf{n}}\lambda = 0$, i.e. with the factor λ constant along the orbits of $\mathbf{n}$.

Calculation: Derive Eq. (3.20).
Solution: A replacement $\tilde{g} \to e^{2\lambda}\tilde{g}$ requires $\Omega \to e^{\lambda}\Omega$ because the physical metric $\Omega^{-2}\tilde{g}$ must remain fixed. Since $\tilde{g}(\mathbf{n},\mathbf{n}) = \tilde{g}^{-1}(\mathrm{d}\Omega,\mathrm{d}\Omega)$ and the differential d is metric-independent, we find that the function f transforms under the above replacements as

$$\begin{aligned} f &= \Omega^{-1}\tilde{g}^{-1}(\mathrm{d}\Omega,\mathrm{d}\Omega) \\ &\to e^{-\lambda}\Omega^{-1}e^{-2\lambda}\tilde{g}^{-1}(\mathrm{d}(e^{\lambda}\Omega),\mathrm{d}(e^{\lambda}\Omega)) \\ &= e^{-\lambda}f + 2e^{-2\lambda}\tilde{g}^{-1}(\mathrm{d}e^{\lambda},\mathrm{d}\Omega) + O(\Omega) \\ &= e^{-\lambda}f + 2e^{-2\lambda}\tilde{\nabla}_{\mathbf{n}}e^{\lambda} + O(\Omega). \end{aligned}$$

Thus we obtain the desired relation on $\mathscr{I}^+$.

Need to check all conformal transformation equations against Appendix D of Wald!

A calculation shows that the unphysical Ricci tensor $\tilde{R}_{\mu\nu}$ is related to the physical one, $R_{\mu\nu}$, by (see [36] Appendix D)

$$\begin{aligned} R_{\mu\nu} = \tilde{R}_{\mu\nu} &- 2\Omega^{-1}\tilde{\nabla}_{\mu}\tilde{\nabla}_{\nu}\Omega \\ &- \tilde{g}_{\mu\nu}\left(\Omega^{-1}\tilde{\nabla}_{\alpha}\tilde{\nabla}^{\alpha}\Omega - 3\Omega^{-2}\tilde{g}(\mathbf{n},\mathbf{n})\right). \end{aligned} \tag{3.21}$$

We now multiply both sides of Eq. (3.21) by Ω. By assumption, $R_{\mu\nu}$ vanishes in a neighborhood of $\mathscr{I}^+$, and $\tilde{R}_{\mu\nu}$ is regular everywhere on $\tilde{\mathcal{M}}$ since it is a nonsingular manifold. Using $\Omega^{-1}\tilde{g}(\mathbf{n},\mathbf{n}) = O(\Omega)$, we find

$$2\tilde{\nabla}_{\mu}\tilde{\nabla}_{\nu}\Omega + \tilde{g}_{\mu\nu}\tilde{\nabla}_{\alpha}\tilde{\nabla}^{\alpha}\Omega = 0 \text{ on } \mathscr{I}^+.$$

Contracting with $\tilde{g}^{\mu\nu}$ gives $\tilde{\nabla}_{\mu}\tilde{\nabla}_{\nu}\Omega = 0$ or equivalently $\tilde{\nabla}\mathbf{n} = 0$, which means that $\mathbf{n}$ is a (null) geodesic, shear-free, and divergence-free vector field on $\mathscr{I}^+$. (The entire distortion tensor $B_{(\mathbf{n})}$ vanishes on $\mathscr{I}^+$.) Note that the condition $g(\mathbf{n},\mathbf{n}) = 0$ could be also obtained from Eq. (3.21) if we multiply both sides by Ω^2.

Since $\mathscr{I}^+$ has the topology $\mathbb{R} \times S^2$, we can select coordinates θ, ϕ on the S^2 sections such that the partial metric is that of a sphere,

$$\mathrm{d}S^2 = \left(\mathrm{d}\theta \otimes \mathrm{d}\theta + \sin^2\theta \mathrm{d}\phi \otimes \mathrm{d}\phi\right) V.$$

Since the vector field $\mathbf{n}$ is divergence-free, the cross-section area of the S^2 sections remains constant along the orbits of $\mathbf{n}$, thus the partial metric will have equal values of V on every S^2 section. So the remaining conformal freedom ($\tilde{g} \to e^{2\lambda}\tilde{g}$ with $\lambda = \text{const}$ along the orbits of $\mathbf{n}$) can be used to set $V = 1$. The function Ω is a well-defined coordinate near $\mathscr{I}^+$ since $\mathbf{n} \equiv \hat{\tilde{g}}^{-1}d\Omega \neq 0$. Finally, we can select a fourth coordinate u parametrizing the $\mathbb{R}$ part of $\mathscr{I}^+ = \mathbb{R} \times S^2$, such that $\mathbf{n} \circ u = 1$. The coordinate system $(u, \Omega, \theta, \phi)$ is well-defined in a neighborhood of $\mathscr{I}^+$, where the metric has the form

$$\tilde{g} = \frac{1}{2}\left(\mathrm{d}u \otimes \mathrm{d}\Omega + \mathrm{d}\Omega \otimes \mathrm{d}u\right) - \mathrm{d}\theta \otimes \mathrm{d}\theta - \sin^2\theta \mathrm{d}\phi \otimes \mathrm{d}\phi.$$

A coordinate system in the physical manifold $\mathcal{M}$ away from $\mathscr{I}^+$ can be obtained by the replacement $r = \Omega^{-1}$. The function r will then serve as a "radial coordinate near infinity" since $r \to \infty$ "at infinity."

The coordinate system (u, r, θ, ϕ) is extended away from $\mathscr{I}^+$ into the interior of $\mathcal{M}$ as follows. The coordinate u is chosen as a null function, so that $\mathbf{l} \equiv \hat{g}^{-1}\mathrm{d}u$ is a null vector such that $g(\mathbf{n},\mathbf{l}) = 1$. Then, r is defined as an affine parameter along the orbits of $\mathbf{l}$, such that $\mathbf{l} \circ r = 1$. The angular coordinates θ, ϕ are kept constant along the orbits of $\mathbf{l}$, so $\nabla_{\mathbf{l}}\phi = \nabla_{\mathbf{l}}\theta = 0$. The resulting coordinate system asymptotically corresponds to spherical Minkowski coordinates $\{t, r, \theta, \phi\}$ if we define $u \equiv t + r$.[5] This local coordinate system is called the **Bondi coordinate system**.

Thus, we conclude that the assumptions of asymptotic flatness are a precise way of formulating the notion of a spacetime that is "almost Minkowski at large distances."

[5] This statement is justified in more detail in [36], p. 280.

3.4 Conformal radiation fields

Very far from an isolated system, the spacetime is almost flat but there may still remain radiation (gravitational, electromagnetic, or other). We are interested in studying the structure of the radiation field at large distances. The construction of conformal infinity is especially helpful for this task, because radiation fields are *conformally invariant*. Let us review some examples of conformally invariant fields.

3.4.1 Scalar field in 1+1 dimensions

The action for a massless scalar field in a 1+1-dimensional spacetime is

$$S[g,\phi] = \frac{1}{2}\int \mathrm{d}^2x\sqrt{-g}\,g^{-1}(\mathrm{d}\phi,\mathrm{d}\phi),$$

where $\sqrt{-g} \equiv \sqrt{-\det g}$ is the determinant of the *covariant* metric, and so $\sqrt{-g}\mathrm{d}^2x$ is the invariant 2-volume element. This action is invariant under the conformal transformation of the metric, $g \to \tilde{g} = \Omega^2 g$.

Calculation: Show that the action is invariant under the above transformation.

Hint: In a 1+1-dimensional spacetime, $\sqrt{-\tilde{g}} = \Omega^2\sqrt{-g}$. Substitute $g = \Omega^{-2}\tilde{g}$ into the action and use the independence of d from the metric.

3.4.2 Scalar field in 3+1 dimensions

The action for a massless, conformally coupled scalar field is

$$S[g,\phi] = \frac{1}{2}\int \mathrm{d}^4x\sqrt{-g}\left(g^{-1}(\mathrm{d}\phi,\mathrm{d}\phi) - \frac{1}{6}R\phi^2\right),$$

where R is the scalar curvature. Again one can verify that the action is invariant under a conformal transformation $g \to \tilde{g} = \Omega^2 g$ (up to boundary terms), provided that the field is also transformed as

$$\phi \to \tilde{\phi} = \Omega^{-1}\phi.$$

The power of the conformal factor needed for the correct transformation law is called the **conformal weight** of the field. Thus a scalar field has conformal weight 0 in 1+1 dimensions and weight -1 in 3+1 dimensions. Conformal invariance means that a solution of the equation of motion for ϕ in the metric g also gives a solution for the field $\tilde{\phi}$ in the metric $\tilde{g}$.

Calculation: Using Eq. (3.8), show that the action is invariant under the above transformations (up to boundary terms).

Hint: An integration by parts yields an equivalent form of the action,

$$S[g,\phi] = \frac{1}{2}\int \mathrm{d}^4x\sqrt{-g}\left(-\phi\Box\phi - \frac{1}{6}R\phi^2\right), \quad \Box \equiv \nabla_\alpha\nabla^\alpha.$$

Substitute $g = \Omega^{-2}\tilde{g}$, $\phi = \Omega\tilde{\phi}$ into this action, and replace R according to Eq. (3.8). Note that

$$\Box(\Omega\tilde{\phi}) = \Omega\Box\tilde{\phi} + \tilde{\phi}\Box\Omega + 2(\nabla_\alpha\Omega)(\nabla^\alpha\tilde{\phi}).$$

3.4.3 Electromagnetic field

The electromagnetic field is described by the Maxwell tensor which is a 2-form F,

$$F = \mathrm{d}A, \quad F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu,$$

where the 1-form A is the electromagnetic potential. The action for the Maxwell field is

$$S[g,F] = \frac{1}{16\pi}\int \sqrt{-g}\mathrm{d}^4x\, g^{\lambda\nu}g^{\mu\rho}F_{\lambda\mu}F_{\nu\rho}.$$

It is easy to see that this action is conformally invariant (in 3+1-dimensional spacetimes), the Maxwell field F having conformal weight 0.

3.4.4 Gravitational radiation field

The Einstein equation relates the Ricci tensor to the distribution of matter in an algebraic way (if there is no matter at a point p, the Ricci tensor is zero at p). The Ricci tensor, however, is only the trace of the full Riemann tensor. Therefore, the curvature of spacetime contains degrees of freedom not algebraically related to the distribution of matter. These degrees of freedom are described by the trace-free part of the Riemann tensor called the **Weyl tensor**,

$$C^{\alpha\beta}{}_{\gamma\delta} = R^{\alpha\beta}{}_{\gamma\delta} + \frac{1}{2}\delta^{[\alpha}_{[\gamma}S^{\beta]}_{\delta]},$$

where

$$S_{\alpha\beta} \equiv R_{\alpha\beta} - \frac{1}{6}Rg_{\alpha\beta}$$

is an auxiliary tensor carrying equivalent information to the Ricci tensor. Note that $R_{\alpha\beta\gamma\delta} = C_{\alpha\beta\gamma\delta}$ in vacuum (where $R_{\alpha\beta} = 0$). It can be shown that the second Bianchi identity implies the constraints

$$\nabla^\mu C_{\mu\alpha\beta\gamma} + \nabla_{[\alpha}S_{\beta]\gamma} = 0,$$
$$\nabla^\mu S_{\alpha\mu} - \nabla_\alpha S^\mu{}_\mu = 0.$$

In the absence of matter, $S_{\alpha\beta} = 0$ and thus we obtain the equation $\nabla^\mu C_{\mu\alpha\beta\gamma} = 0$ which determines the dynamics of pure gravitational waves.

The Weyl tensor has the simple and useful conformal transformation property

$$\tilde{C}_{\alpha\beta\gamma\delta} = \Omega^2 C_{\alpha\beta\gamma\delta} \quad \text{for} \quad \tilde{g} = \Omega^2 g.$$

Another way to express this property is to say that the tensor $C^\alpha{}_{\beta\gamma\delta}$ is conformally invariant.

3.4.5 Asymptotic behavior of radiation

Radiation fields are massless and propagate along null geodesics (lightrays). Let us consider a lightray $\gamma(\tau)$ that approaches null infinity $\mathscr{I}^+$ at a point p. We shall now characterize a radiation field along the lightray γ, in the asymptotically far region near infinity.

Denote by $\mathbf{l}$ the affine tangent vector to the lightray $\gamma(\tau)$. In the unphysical metric ($\tilde{g} = \Omega^2 g$), the affine parameter τ should be replaced by $\Omega^2\tau$, so that the corresponding affine tangent vector is $\tilde{\mathbf{l}} = \Omega^{-2}\mathbf{l}$ (see Statement 1.9.2.2). Since the tangent vector $\tilde{\mathbf{l}}$ is regular in the unphysical manifold, it should have a well-defined value $\tilde{\mathbf{l}}_0$ on $\mathscr{I}^+$. Then the

physical-space tangent vector $\mathbf{l}$ must decay at infinity as $\mathbf{l} \sim \tilde{\mathbf{l}}_0\Omega^2$. We shall now complete $\mathbf{l}$ to a null tetrad $\{\mathbf{l}, \mathbf{m}, \mathbf{n}\}$ (see Sec. 2.5), obtain similar asymptotic properties for the tetrad at infinity, and study the asymptotic behavior of radiation fields using the tetrad components.

The vector $\tilde{\mathbf{l}}_0$ can be completed to a null tetrad using the vector $\tilde{\mathbf{n}} \equiv -\hat{\tilde{g}}^{-1}\mathrm{d}\Omega$ and a complex-valued vector $\mathbf{m}$ is chosen in the orthogonal complement $\{\tilde{\mathbf{n}}, \tilde{\mathbf{l}}_0\}^\perp$. The (unphysical) null tetrad $\{\tilde{\mathbf{l}}_0, \tilde{\mathbf{m}}, \tilde{\mathbf{n}}\}$ can be parallelly transported from $\mathscr{I}^+$ along γ using the unphysical metric, which produces null vector fields $\{\tilde{\mathbf{l}}, \tilde{\mathbf{m}}, \tilde{\mathbf{n}}\}$ defined along γ and satisfying $\tilde{g}(\tilde{\mathbf{n}}, \tilde{\mathbf{l}}) = 1$ and $\tilde{g}(\tilde{\mathbf{m}}, \bar{\tilde{\mathbf{m}}}) = -1$ at every point. However, we need to find a "physical" tetrad, $\{\mathbf{l}, \mathbf{m}, \mathbf{n}\}$, normalized and parallelly transported along γ using the physical metric. We have already seen that $\mathbf{l} = \Omega^2\tilde{\mathbf{l}}$. The vector $\mathbf{n}$ of the physical tetrad can be chosen as $\mathbf{n} = \tilde{\mathbf{n}}$ since we must have $\tilde{g}(\tilde{\mathbf{n}}, \tilde{\mathbf{l}}) = g(\mathbf{n}, \mathbf{l}) = 1$ everywhere along γ. Finally, the complex-valued vector $\mathbf{m}$ is selected from the subspace $\{\mathbf{n}, \mathbf{l}\}^\perp$ to be proportional to $\tilde{\mathbf{m}}$. The normalization $g(\mathbf{m}, \bar{\mathbf{m}}) = \tilde{g}(\tilde{\mathbf{m}}, \bar{\tilde{\mathbf{m}}}) = -1$ shows that the unphysical-space vector $\tilde{\mathbf{m}}$ must be chosen as $\tilde{\mathbf{m}} = \Omega^{-1}\mathbf{m}$. Thus, the physical null tetrad $\{\mathbf{l}, \mathbf{m}, \mathbf{n}\}$ and the unphysical one, $\{\tilde{\mathbf{l}}, \tilde{\mathbf{m}}, \tilde{\mathbf{n}}\}$, must be related by

$$\{\Omega^2\tilde{\mathbf{l}}, \Omega\tilde{\mathbf{m}}, \tilde{\mathbf{n}}\} = \{\mathbf{l}, \mathbf{m}, \mathbf{n}\}$$

everywhere along the lightray γ.

The convenience of the null tetrads is in the concise way they describe components of tensors. For example, the antisymmetric Maxwell tensor $F(\mathbf{a}, \mathbf{b})$ for the electromagnetic field is represented by three **tetrad components**, usually chosen as follows,

$$\begin{aligned}
\Phi_0 &\equiv F(\mathbf{n}, \mathbf{m}),\\
\Phi_1 &\equiv \frac{1}{2}F(\mathbf{n}, \mathbf{l}) - \frac{1}{2}F(\mathbf{m}, \bar{\mathbf{m}}),\\
\Phi_2 &\equiv F(\bar{\mathbf{m}}, \mathbf{l}).
\end{aligned}$$

(The properties of the tetrad components Φ_j are studied in the Statement below.) The unphysical components $\tilde{\Phi}_j$, $j = 0, 1, 2$, are given by the same equations with the unphysical tetrad. The relation between the physical and the unphysical components is therefore

$$\Phi_0 = \Omega\tilde{\Phi}_0, \quad \Phi_1 = \Omega^2\tilde{\Phi}_1, \quad \Phi_2 = \Omega^3\tilde{\Phi}_2.$$

Since the Maxwell field is conformally invariant, the unphysical components will also satisfy the Maxwell equations and will have regular solutions on $\mathscr{I}^+$. It is now easy to see that the component Φ_0 will be the slowest-decaying one,

$$\Phi_0 \sim \Omega \sim \frac{1}{r},$$

while other components will decay faster. Note that F and Φ_j represent the field strength, which should decay as r^{-1} for a radiation field. Hence, it is clear that the the intensity of *outgoing radiation* (i.e. energy carried away to infinity) must be independent of the two other components $\Phi_{1,2}$. In fact, the radiated power is proportional to $|\Phi_0|^2$. Since the unphysical component $\tilde{\Phi}_0$ is constant at infinity, the differential intensity of radiation in a given (null) direction can be obtained directly by evaluating $|\tilde{\Phi}_0|^2$ at a point of $\mathscr{I}^+$ that corresponds to the required asymptotic direction.

Other radiation fields exhibit analogous properties. For example, a scalar field ϕ has the "radiation" component $\nabla_{\mathbf{n}}\phi$, while the gravitational radiation is described by the component

$$\Psi_0 \equiv C_{\alpha\beta\gamma\delta}n^\alpha m^\beta n^\gamma m^\delta = R(\mathbf{n}, \mathbf{m}, \mathbf{n}, \mathbf{m});$$

note that this component of the Weyl tensor coincides with that of the Riemann tensor. The component Ψ_0 decays as $\sim \Omega = r^{-1}$ at infinity, and the intensity of gravitational radiation is proportional to $|\Psi_0|^2$. All the other components of the Weyl tensor decay faster at infinity.

The splitting of components according to their falloff behavior along null geodesics is called the **peeling property**. This property holds in general (even-dimensional) asymptotically flat spacetimes and for all massless fields (of arbitrary spin).

Statement: Consider a Lorentz transformation of the tetrad that leaves the vector $\mathbf{l}$ constant,

$$\begin{aligned}
\mathbf{m} &\to e^{\mathrm{i}\phi}\mathbf{m} + A\mathbf{l},\\
\mathbf{n} &\to \mathbf{n} + e^{\mathrm{i}\phi}\bar{A}\mathbf{m} + e^{-\mathrm{i}\phi}A\bar{\mathbf{m}} + A\bar{A}\mathbf{l},
\end{aligned}$$

where ϕ is real and A is complex. Under this transformation, the tetrad components Φ_j of the Maxwell tensor change as

$$\begin{aligned}
\Phi_0 &\to e^{\mathrm{i}\phi}\Phi_0 + 2A\Phi_1 + A^2e^{-\mathrm{i}\phi}\Phi_2,\\
\Phi_1 &\to \Phi_1,\\
\Phi_2 &\to e^{-\mathrm{i}\phi}\Phi_2.
\end{aligned}$$

It follows that the leading falloff behavior of $\Phi_0 \sim r^{-1}$ is independent of the choice of the null tetrad. The quantities $F(\mathbf{n}, \mathbf{l})$ and $F(\mathbf{m}, \bar{\mathbf{m}})$, taken separately, do not carry tetrad-independent information (i.e. one of these may be set to zero by a choice of tetrad).

4 Global techniques

Under **global techniques** we understand considerations that involve entire manifolds or large domains, rather than *local* considerations (e.g. calculations involving properties of tensors and their derivatives at a single point). Global techniques can be used to answer questions such as whether there is a singularity anywhere, or whether a certain interesting subdomain of the spacetime is finite or infinite. In this chapter we shall introduce some of these techniques and consider some applications.

4.1 Singularity theorems

Additional literature: [8], [26]. There are not many textbooks that explain singularity theorems.

[12]: Contains all the derivations but requires you to invest a significant time in reading large portions of the book.

[21]: Contains some introductory material on singularity theorems; Chapter 34 contains a sufficiently detailed proof of Hawking's area theorem.

[36]: Devotes two chapters to mathematical developments and explains the main ideas behind singularity theorems. However, refers to [12] for proofs of some crucial statements.

4.1.1 Singularities and geodesic incompleteness

It is well known that some spacetimes occurring in general relativity contain points where the force of gravity becomes "infinitely large" (more precisely, the geometric description of the spacetime breaks down). Such spacetimes are called **singular**. Two major examples of singular spacetimes are the Schwarzschild spacetime and a generic Friedmann-Robertson-Walker (FRW) spacetime. In the Schwarzschild spacetime, it is relatively easy to understand that the black hole center $r = 0$ represents a singularity because the curvature becomes infinite as $r \to 0$, indicating the presence of unboundedly large tidal forces. Also, every timelike or null geodesic entering the Schwarzschild radius cannot escape hitting the center within a finite time. No lightrays or timelike geodesics can be continued past that point because the geometry is singular and the geodesic equation becomes undefined (i.e. it does not predict how a body will move past the center). Since the tidal forces will tear apart any material objects approaching the center, it is reasonable to conclude that there is no sensible way to describe the evolution of an observer past the center of a black hole. Thus the center is a place where the geometry of the spacetime cannot be described by general relativity. In an FRW spacetime with the scale factor $a(t) \to 0$ as $t \to 0$, the curvature is unbounded near $t = 0$. Similarly to the Schwarzschild case, geodesics cannot be extended to the past beyond $t = 0$, so one cannot describe what happened to an observer before $t = 0$. In this sense, $t = 0$ is the "beginning of time" for an FRW spacetime. It is reasonable to say that the 3-surface $t = 0$ in an FRW spacetime is singular.

Since the spacetime manifold is not smooth at "singular" points, it is natural to excise such points from the spacetime and to restrict the calculations to regular (nonsingular) points.[1] Hence, one finds a curious situation: gravitational singularities are not "places" and are not present "anywhere" in the spacetime manifold. Since all the singular points are removed, it is difficult to formulate a general definition of a singularity in terms of internal properties of the remaining nonsingular manifold. Nevertheless, the "presence" of singularities is usually manifested in several ways. For instance, the curvature may become infinite along a certain geodesic worldline, indicating that a freely falling observer following that worldline will experience unbounded tidal forces. Or, certain geodesics may have only a finite range of the affine parameter, because these geodesics "hit a singularity" and cannot be extended any further.

An accepted way to define a singular spacetime is through the non-extensibility of geodesics. Intuitively, a geodesic should be able to extend "arbitrarily far" (for an infinite range of the affine parameter) unless there is a singularity in the way. A spacetime is called **geodesically complete** if any geodesic is extendible to arbitrary values of the affine parameter. A geodesically *in*complete spacetime is interpreted as "singular" in some way. Usually, only timelike and null geodesics are used to determine geodesic completeness. (Note that a spacetime can be intuitively "singular" without containing any incomplete timelike and null geodesics because the singularity may be "infinitely far away," approachable after finite proper time only by a sufficiently highly accelerated observer, or by a spacelike geodesic.)

The presently known **singularity theorems** are mathematical statements about conditions for a spacetime to be geodesically incomplete. Thus, these theorems only indicate the presence of singularities but neither predict their "location" nor characterize their physical properties.

Statement:[2] Let $\mathbf{k}$ be a timelike Killing vector and $\gamma(\tau)$ a (non-geodesic) timelike curve with tangent vector $\mathbf{v}$, such that $g(\mathbf{v}, \mathbf{v}) = 1$. Show that

$$|\nabla_{\mathbf{v}} g(\mathbf{k}, \mathbf{v})| \leq |g(\mathbf{k}, \mathbf{v})|\,|\mathbf{a}|\,,$$

where the vector $\mathbf{a}(\tau) \equiv \nabla_{\mathbf{v}}\mathbf{v}$ is the proper acceleration of the curve and $|\mathbf{a}| \equiv \sqrt{-g(\mathbf{a}, \mathbf{a})}$. Deduce that $g(\mathbf{v}, \mathbf{k})$ can change only by a finite amount along the curve γ as long as $\int_{\tau_0}^{\infty} |\mathbf{a}(\tau)|\, d\tau < \infty$, which means that the worldline $\gamma(\tau)$ can be realized by a rocket with a finite amount of fuel. Show that the curve γ cannot reach any spacetime regions where $g(\mathbf{k}, \mathbf{k})$ becomes unbounded.

Solution: By the Killing property, $g(\nabla_{\mathbf{v}}\mathbf{k}, \mathbf{v}) = 0$, thus $\nabla_{\mathbf{v}} g(\mathbf{k}, \mathbf{v}) = g(\mathbf{k}, \mathbf{a})$. It remains to show that $|g(\mathbf{k}, \mathbf{a})| < |g(\mathbf{k}, \mathbf{v})\mathbf{a}|$ for timelike $\mathbf{k}, \mathbf{v}$ and spacelike $\mathbf{a} \neq 0$ such that $g(\mathbf{v}, \mathbf{a}) = 0$. We may decompose $\mathbf{k} = \alpha\mathbf{a} + \beta\mathbf{v} + \mathbf{s}$, where

$$\alpha \equiv \frac{g(\mathbf{k}, \mathbf{a})}{g(\mathbf{a}, \mathbf{a})}, \quad \beta \equiv g(\mathbf{k}, \mathbf{v}),$$

[1] It is conjectured that a quantum theory of gravity (to be developed) will replace "singularities" by some well-defined and finite new physics. At this point, it is safe to say that the classical theory of general relativity breaks down at singularities.

[2] After [36], Problem 9.1.

and $\mathbf{s}$ is a spacelike vector orthogonal to both $\mathbf{a}$ and $\mathbf{v}$. Then the inequality $g(\mathbf{k},\mathbf{k}) > 0$ yields

$$g(\mathbf{k},\mathbf{k}) = \beta^2 - \alpha^2 |\mathbf{a}|^2 - |\mathbf{s}|^2 > 0,$$

hence

$$|\beta| = |g(\mathbf{k},\mathbf{v})| > |\alpha|\,|\mathbf{a}| = \frac{1}{|\mathbf{a}|}\,|g(\mathbf{k},\mathbf{a})|\,,$$

which is equivalent to the desired inequality. Rewriting it as

$$|\nabla_{\mathbf{v}} \ln g(\mathbf{k},\mathbf{v})| < |\mathbf{a}|\,,$$

and integrating both sides, we have

$$\left|\int_{\tau_0}^{\infty} \mathrm{d}\tau \nabla_{\mathbf{v}} \ln g(\mathbf{k},\mathbf{v})\right| \le \int_{\tau_0}^{\infty} \mathrm{d}\tau\, |\nabla_{\mathbf{v}} \ln g(\mathbf{k},\mathbf{v})| \le \int |\mathbf{a}|\,\mathrm{d}\tau < \infty.$$

Therefore the total change of $\ln g(\mathbf{k},\mathbf{v})$ is bounded. Since for any timelike vectors $\mathbf{x},\mathbf{y}$ we have the inequality

$$g(\mathbf{x},\mathbf{x})g(\mathbf{y},\mathbf{y}) \le [g(\mathbf{x},\mathbf{y})]^2\,,$$

it follows that $g(\mathbf{k},\mathbf{k}) \le |g(\mathbf{k},\mathbf{v})|$ and thus the total change of $g(\mathbf{k},\mathbf{k})$ along the curve $\gamma(\tau)$ is bounded for all τ.

4.1.2 Past-incompleteness of inflation

To get a taste of singularity theorems, we start with a simple statement about the past timelike geodesic incompleteness of an inflationary spacetime.[3]

Cosmological **inflation** is a period of accelerated expansion of spacetime. For instance, a spatially flat FRW metric with a given scale factor $a(t)$, namely

$$g = \mathrm{d}t \otimes \mathrm{d}t - a^2(t)\mathrm{d}^2 r, \qquad \mathrm{d}^2 r \equiv \mathrm{d}x \otimes \mathrm{d}x + \mathrm{d}y \otimes \mathrm{d}y + \mathrm{d}z \otimes \mathrm{d}z, \tag{4.1}$$

represents inflation for those times t when $\ddot{a} > 0$ and $\dot{a} > 0$. The **Hubble expansion rate** is defined as

$$H(t) = \frac{d}{dt} \ln a,$$

so the conditions for the presence of inflation are $\dot{H} + H^2 > 0$ and $H > 0$. A full discussion of possible varieties of cosmological inflation is beyond the scope of this text. We shall restrict our attention to a class of spacetimes that are qualitatively similar to the inflating FRW case, namely, to spacetimes with a metric of approximately the form (4.1), in suitable local coordinates. Locally, the parameters a and H will be well-defined and should satisfy the conditions $H > 0$, $\dot{H} + H^2 > 0$. The processes that generate inflation may result in local fluctuations of the Hubble rate H, but these fluctuations will occur on distance and time scales typically much larger than the horizon size H^{-1}.

It is clear that inflation can continue forever to the future. The main statement is that such an inflationary spacetime cannot be also eternal to the past. The way to prove this is to show that a past-directed timelike geodesic will have only a finite range of proper time. Events that preceded that time are not described by the inflationary spacetime and so may be interpreted as the presence of an initial singularity (the "beginning of time").

We first describe the inflationary character of the spacetime in a coordinate-free manner. A salient property of the spacetime (4.1) is the existence of a congruence of timelike geodesics with tangent vectors $\mathbf{v} \equiv \partial_t$ and connecting vectors $\mathbf{c}_{1,2,3} = \partial_x, \partial_y, \partial_z$ that satisfy the Hubble law (3.4). These geodesics have everywhere positive divergence (see the following calculation).

Calculation: For the metric (4.1), show that the vectors $\mathbf{c}_{1,2,3} \equiv \partial_x, \partial_y, \partial_z$ are connecting for $\mathbf{v} \equiv \partial_t$ and satisfy the Hubble law (3.4) with $H = \dot{a}/a$. Then show that the divergence $\operatorname{div} \mathbf{v} = 3H$ and hence is everywhere positive during inflation.

Solution: The Hubble law, $\nabla_{\mathbf{v}}\mathbf{c}_j = H\mathbf{c}_j$, follows from

$$g(\nabla_{\mathbf{v}}\mathbf{c}_j, \mathbf{c}_j) = \frac{1}{2}\nabla_{\mathbf{v}} g(\mathbf{c}_j, \mathbf{c}_j) = -\frac{1}{2}\partial_t a^2 = -a^2 H.$$

Then we compute the divergence using the decomposition

$$g^{-1} = \mathbf{v} \otimes \mathbf{v} - a^{-2} \sum_{j=1}^{3} \mathbf{c}_j \otimes \mathbf{c}_j$$

that follows from the fact that $\{\mathbf{v}, a^{-1}\mathbf{c}_1, a^{-1}\mathbf{c}_2, a^{-1}\mathbf{c}_3\}$ is an orthonormal basis. Hence,

$$\begin{aligned}\operatorname{div} \mathbf{v} &= \operatorname{Tr}_{(\mathbf{x},\mathbf{y})} g(\mathbf{x}, \nabla_{\mathbf{y}}\mathbf{v}) \\ &= g(\mathbf{v}, \nabla_{\mathbf{u}}\mathbf{v}) - a^{-2} \sum_{j=1}^{3} g(\mathbf{c}_j, \nabla_{\mathbf{c}_j}\mathbf{v}) \\ &= -a^{-2} \sum_{j=1}^{3} g(\mathbf{c}_j, \nabla_{\mathbf{v}}\mathbf{c}_j) = 3H > 0.\end{aligned}$$

Thus we reformulate our assumption of the presence of inflation throughout the entire "past portion" of the spacetime: We require the existence of a timelike geodesic congruence $\mathbf{v}$ covering the entire past, with everywhere positive divergence $\operatorname{div}\mathbf{v} > 0$. Here, the "past portion" of the spacetime is understood as the past of some Cauchy 3-surface. Note that the existence of an everywhere diverging congruence is a nontrivial condition. In any spacetime, one can always find geodesic congruences having arbitrary, positive or negative, divergence at any given point, but not necessarily *everywhere* to the past of a Cauchy surface.

Given an everywhere diverging geodesic congruence, we define the local Hubble expansion rate to be $H \equiv \frac{1}{3}\operatorname{div}\mathbf{v}$. We shall assume, for simplicity, that there exists a constant $H_0 > 0$ such that $H \ge H_0$ everywhere. (This condition can be significantly relaxed without affecting the conclusions.)

Now we imagine that some dust-like particles follow the geodesic congruence $\mathbf{v}$ everywhere, and consider a geodesic observer who moves through the universe and measures the velocities of the dust particles. This observer can conclude that the spacetime is inflating and measure the local value of H in the following way. Let the observer's worldline be $\gamma(\tau)$ with an affine tangent vector $\mathbf{u} \equiv \dot{\gamma}$, such that $g(\mathbf{u},\mathbf{u}) = 1$. In principle, the observer can measure the following two (3-dimensional) quantities: the relative velocity, $\vec{v}_{\mathrm{rel}}(\tau)$, of the particle that passes by the observer at a time τ, and the displacement vector $\delta\vec{x}(\tau;\tau')$, in the observer's rest frame, between nearby particles that passed the observer at times τ

[3] This section is based on the paper [3].

and τ'. After registering two nearby particles at times τ and $\tau' \equiv \tau + \delta\tau$, the observer can calculate the separation and the relative velocity of the two particles in the (approximate) local rest frame of the particles. Let $\mathbf{n}$ be a normalized, spacelike vector describing the separation between two particles in the particle rest frame. Then the observer can compute the quantity $g(\mathbf{n}, \nabla_{\mathbf{n}}\mathbf{v})$ and thus estimate the local Hubble rate from the rate of change of the separation in the direction $\mathbf{n}$,

$$H \equiv \frac{1}{3}\mathrm{div}\mathbf{v} \approx -g(\mathbf{n}, \nabla_{\mathbf{n}}\mathbf{v})$$

(the minus sign compensates for the spacelike character of the vectors $\mathbf{n}$ and $\nabla_{\mathbf{n}}\mathbf{v}$). Note that $\mathbf{n}$ must be normalized, $g(\mathbf{n}, \mathbf{n}) = -1$, so the observed Hubble rate depends only on the direction of the observer's motion but not on its speed. Since the spacetime is approximately isotropic, the estimate of the value of H from a single direction $\mathbf{n}$ is expected to be reasonably accurate.

The next calculations derive a formula for H in terms of the vectors $\mathbf{u}$ and $\mathbf{v}$. It will be convenient to denote $\alpha(\tau) \equiv g(\mathbf{u}, \mathbf{v})$. The quantity α is the "γ factor" of special relativity, describing the time dilation between the observer and the particle rest frames. Note that $g(\mathbf{u}, \mathbf{v}) > 1$ for two future-directed timelike vectors $\mathbf{u} \neq \mathbf{v}$ normalized to $g(\mathbf{u}, \mathbf{u}) = g(\mathbf{v}, \mathbf{v}) = 1$. Since we assume that the observer is not at rest with respect to any of the particles, we will always have $\alpha > 1$.

Calculation: Using the geodesic property of the fields $\mathbf{v}$ and $\mathbf{u}$, show that the vector $\mathbf{n}$ and the instantaneous Hubble rate H are expressed by

$$\mathbf{n} = \frac{\mathbf{u} - \alpha\mathbf{v}}{\sqrt{\alpha^2 - 1}}, \quad H = -\frac{\nabla_{\mathbf{u}}\alpha}{\alpha^2 - 1}.$$

Derive the relationship between the observed 3-dimensional quantities $\vec{v}_{\rm rel}(\tau)$, $\vec{v}_{\rm rel}(\tau + d\tau)$, $\delta\vec{x}(\tau; \tau + d\tau)$, and the 4-dimensional vectors $\mathbf{u}$, $\mathbf{v}$, as well as $\nabla_{\mathbf{u}}\alpha$, in the limit $d\tau \to 0$.

Solution: The vector $\mathbf{n}dt$ should represent the infinitesimal displacement of the observer in the particle rest frame, where the time interval is dt. Spatial vectors in the particle rest frame are (spacelike) vectors orthogonal to $\mathbf{v}$. Thus, the vector $\mathbf{n}$ is proportional to the projection of $\mathbf{u}$ onto the subspace $\mathbf{v}^\perp$:

$$\mathbf{n} = \frac{\mathbf{u} - \mathbf{v}g(\mathbf{u}, \mathbf{v})}{|\mathbf{u} - \mathbf{v}g(\mathbf{u}, \mathbf{v})|} = \frac{\mathbf{u} - \alpha\mathbf{v}}{\sqrt{\alpha^2 - 1}},$$

where we have used $g(\mathbf{u}, \mathbf{u}) = g(\mathbf{v}, \mathbf{v}) = 1$. Now we can directly compute $H \equiv -g(\mathbf{n}, \nabla_{\mathbf{n}}\mathbf{v})$. Since $\mathbf{v}$ is normalized and geodesic, we have $g(\mathbf{v}, \nabla_{\mathbf{x}}\mathbf{v}) = 0 = g(\mathbf{x}, \nabla_{\mathbf{v}}\mathbf{v})$ for all $\mathbf{x}$. So the only surviving term in $g(\mathbf{n}, \nabla_{\mathbf{n}}\mathbf{v})$ is proportional to $g(\mathbf{u}, \nabla_{\mathbf{u}}\mathbf{v})$,

$$H = -g(\mathbf{n}, \nabla_{\mathbf{n}}\mathbf{u}) = -\frac{g(\mathbf{u}, \nabla_{\mathbf{u}}\mathbf{v})}{\alpha^2 - 1} = -\frac{\nabla_{\mathbf{u}}\alpha}{\alpha^2 - 1}.$$

This is the desired formula. In the instantaneous inertial frame where the observer is at rest, we have $\mathbf{u} = \{1, 0, 0, 0\}$ and thus $\mathbf{v} = \{\alpha, \alpha\vec{v}_{\rm rel}\}$, where

$$\alpha(\tau) \equiv \frac{1}{\sqrt{1 - \vec{v}^2_{\rm rel}(\tau)}}$$

and τ is the observer's proper time. In this way, the observer can compute $\alpha(\tau)$ and $\nabla_{\mathbf{u}}\alpha \equiv \partial_\tau \alpha$.

Replacing $\nabla_{\mathbf{u}}$ by ∂_τ, we can simplify the formula for H to

$$H(\tau) = \frac{1}{2}\partial_\tau \ln\frac{\alpha(\tau) + 1}{\alpha(\tau) - 1}.$$

Since $\alpha > 1$, we have

$$\ln\frac{\alpha(\tau) + 1}{\alpha(\tau) - 1} > 0 \quad \text{for all } \tau.$$

Finally, our hypothetical observer can integrate the instantaneous value of $H(\tau)$ along the worldline $\gamma(\tau)$. By assumption, $H \geq H_0$, therefore

$$\int_{\tau_1}^{\tau_2} H(\tau)d\tau \geq H_0(\tau_2 - \tau_1)$$

for all τ_1, τ_2. On the other hand,

$$\begin{aligned}\int_{\tau_1}^{\tau_2} H(\tau)d\tau &= \frac{1}{2}\ln\frac{\alpha(\tau_2) + 1}{\alpha(\tau_2) - 1} - \frac{1}{2}\ln\frac{\alpha(\tau_1) + 1}{\alpha(\tau_1) - 1} \\ &< \frac{1}{2}\ln\frac{\alpha(\tau_2) + 1}{\alpha(\tau_2) - 1}.\end{aligned}$$

Hence, there is a lower bound on the admissible values τ_1 of the proper time,

$$\tau_1 > \tau_2 - \frac{1}{2H_0}\ln\frac{\alpha(\tau_2) + 1}{\alpha(\tau_2) - 1} \equiv \tau_{\rm min}(\tau_2).$$

This means that no timelike geodesic can be extended to the past further than to $\tau_{\rm min}(\tau_2)$. Thus, the spacetime is geodesically incomplete to the past. (A similar argument shows that null geodesics are also past-incomplete.) The only geodesic worldlines that might be complete to the past are the particle worldlines.

What could the observer see at earlier times $\tau \leq \tau_{\rm min}$? One possibility is that the observer did not exist at those times because the spacetime is singular in the distant past. In this case, we can say that a past-directed geodesic "hits a singularity" at $\tau = \tau_{\rm min}$ or sooner, but we cannot deduce any more information about the singularity. For instance, different observers might hit different, "separate" singularities. The only alternative to the presence of singularities is to conclude that the spacetime was not inflating at $\tau < \tau_{\rm min}$. In both cases, the inflationary epoch could not be past-eternal.

4.1.3 Conjugate points on geodesics

Proofs of several singularity theorems are based on the properties of geodesic curves. In this and the following sections we shall obtain some necessary geometric results about geodesics.

Let $\mathbf{u}$ be a tangent vector field to a congruence which consists of geodesics emitted from a single point p; thus p is a focal point for the congruence $\mathbf{u}$. There is a possibility that some of the lines of $\mathbf{u}$ will focus again at another focal point q. The condition for this possibility is easy to derive using the geodesic deviation equation (see Sec. 1.9.5).

Namely, consider a connecting vector $\mathbf{c}$ for $\mathbf{u}$. The vector $\mathbf{c}$ satisfies the geodesic deviation equation,

$$\nabla_{\mathbf{u}}\nabla_{\mathbf{u}}\mathbf{c} - R(\mathbf{u}, \mathbf{c})\mathbf{u} = 0. \tag{4.2}$$

We now pick a single geodesic $\gamma(\tau)$ such that $\tau = 0$ at p, and solve Eq. (4.2) only along that geodesic for $\tau > 0$. Since we are solving a second-order equation, we need to supply initial values $\mathbf{c}$ and $\nabla_{\mathbf{u}}\mathbf{c}$ at $\tau = 0$. At p, we must have $\mathbf{c} = 0$ since p is a focal point for $\mathbf{u}$. We may pick the value $\nabla_{\mathbf{u}}\mathbf{c}(\tau = 0)$ arbitrarily, and obtain $\mathbf{c}(\tau)$ from the deviation equation. Suppose

that there exists a particular solution $\mathbf{c}(\tau)$ such that $\mathbf{c}(\tau_1) = 0$ at some other point $q \equiv \gamma(\tau_1)$ on the curve. It means that an *infinitesimally close*, neighbor geodesic[4] from the congruence $\mathbf{u}$ intersects γ at the point q. Moreover, note that the equation (4.2) is linear in $\mathbf{c}$; if $\mathbf{c}(\tau)$ is a solution, so is $\lambda\mathbf{c}(\tau)$ for any *constant* λ. It follows that there exists a one-parametric set of infinitesimally close neighbor geodesics corresponding to connecting vectors $\lambda\mathbf{c}$, and each of those geodesics intersects γ at q. It is clear that q is a second focal point for the congruence $\mathbf{u}$. In that case, the point q is called **conjugate** to p along the line $\gamma(\tau)$.

Remark: A vector field $\mathbf{c}$ is called a **Jacobi field** for a curve γ if it satisfies Eq. (4.2) along γ with $\mathbf{u} \equiv \dot{\gamma}$. It is clear that Jacobi fields are essentially connecting vectors for a family of geodesics around γ. One uses the concept of Jacobi fields when one would like to examine the focusing properties of a single curve $\gamma(\tau)$ and to avoid the need to introduce a whole congruence of geodesics around γ and to talk about intersections of "infinitesimally close" geodesics. For instance, one can define conjugate points without involving a congruence, in the following way: Two points $p = \gamma(\tau_1)$, $q = \gamma(\tau_2)$ on a geodesic curve $\gamma(\tau)$ are conjugate if there exists a Jacobi field $\mathbf{c}(\tau)$, not identically zero, which vanishes at τ_1 and τ_2.

It is intuitively obvious that the divergence of the field $\mathbf{u}$ tends to infinity at a focal point. To show this more formally, consider the defining property of the divergence $\mathrm{div}\mathbf{u}$,

$$\mathcal{L}_\mathbf{u} V = \partial_\tau V = (\mathrm{div}\mathbf{u})V,$$

where V is a 3-volume spanned by three connecting vectors transverse to $\mathbf{u}$. As we cross a focal point, at least one of the connecting vectors $\mathbf{c}$ that span V will vanish. Thus V passes through 0 at a focal point; in the generic case where the function $V(\tau)$ has only simple zeros, V will also change sign after a focal point. Note that $\partial_\tau V \equiv \nabla_\mathbf{u} V$ is always finite since $V = \varepsilon(\mathbf{u}, \mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3)$ and every connecting vector $\mathbf{c}_j$ satisfies a linear equation and thus cannot become infinite. Therefore, any point where $\mathrm{div}\mathbf{u}$ becomes infinite must be a focal point where $V = 0$. Also, we can rewrite the above equation as $\mathrm{div}\mathbf{u} = \partial_\tau \ln V$. If $\mathrm{div}\mathbf{u}$ remains finite, $\ln V$ cannot go to $-\infty$ within a finite time τ. It follows that $\mathrm{div}\mathbf{u} \to \infty$ at a focal point. (In the generic case, we will have $\mathrm{div}\mathbf{u} \to -\infty$ before a focal point and $\mathrm{div}\mathbf{u} \to +\infty$ after a focal point.) Thus we have proved that the condition $\mathrm{div}\mathbf{u} \to \infty$ is necessary and sufficient for the existence of a focal point of a field $\mathbf{u}$.

Finally, let us develop a qualitative picture of why conjugate points occur. We have seen that gravity generally acts as an attractive force that makes geodesics converge. Therefore, a bunch of geodesics emitted from a point p will usually tend to refocus at a later point q. The refocusing may take a while, but once the divergence $\mathrm{div}\mathbf{u}$ becomes negative along one line, refocusing is inevitable within a finite time. (Note that the focusing theorem applies to hypersurface orthogonal geodesics, and the congruence emitted from p is hypersurface orthogonal by Statement 1 from Sec. 2.4.2. Moreover, this congruence is integrable by Statement 1 from Sec. 2.2.3.) Focal points can occur along a caustic surface or caustic line, so we select one geodesic γ on which focusing occurs at one point q. Then q is the point conjugate to p along γ; other geodesics γ' may possess different conjugate points to p, and also there may exist several conjugate points q, q', ..., to p along a single curve γ. Thus the property of q to be conjugate to p is a property of the local geometry around a selected geodesic γ. Note also that the focusing of geodesics does not yet indicate any singularities in the spacetime; after all, we can select congruences that focus at any point by simply emitting geodesics from that point.

The focusing theorem (see Sec. 2.4.2) shows that $\mathrm{div}\mathbf{u}$ will reach infinity within a finite proper time if the congruence initially has negative divergence and if the strong energy condition holds. Hence, conjugate points must exist along each curve of the congruence, under the same assumptions.

4.1.4 Second variation of proper length

In Sec. 1.9.3 we have seen that a geodesic curve extremizes the proper length among all curves connecting two fixed points. For instance, in the Minkowski space, timelike geodesics *maximize* the proper time; note that spacelike geodesics do not *minimize* the proper distance, since the metric g has the signature $(+---)$. However, in more general spacetimes the proper length of a geodesic does not always deliver the maximum proper time. We shall now see that the maximization of proper length is closely related to the existence of conjugate points for neighbor geodesics.

The infinitesimal change of the proper length of a curve under an infinitesimal variation is described by Eq. (1.79) of Sec. 1.9.3. For a timelike curve $\gamma(\tau)$ with a tangent vector $\mathbf{v} = \dot{\gamma}$ normalized by $g(\mathbf{v}, \mathbf{v}) = 1$, perturbed by the flow of a vector field $\mathbf{t}$ with a parameter σ along the flow, we have

$$\frac{d}{d\sigma} L[\gamma; \mathbf{t}] = -\int_{\tau_1}^{\tau_2} g(\mathbf{t}, \nabla_\mathbf{v}\mathbf{v}) d\tau.$$

The derivative $dL[\gamma;\mathbf{t}]/d\sigma$ vanishes when γ is a geodesic, but this is not sufficient to have an actual *maximum* of the functional L. The proper length is maximized if $d^2L/d\sigma^2 < 0$. To verify this condition, we again consider a vector field $\mathbf{v}$ created by transporting the tangent vector $\dot{\gamma}$ along the orbits of $\mathbf{t}$. Then

$$\begin{aligned}\frac{d^2}{d\sigma^2} L[\gamma; \mathbf{t}] &= -\int_{\tau_1}^{\tau_2} \nabla_\mathbf{t} g(\mathbf{t}, \nabla_\mathbf{v}\mathbf{v}) d\tau \\ &= -\int_{\tau_1}^{\tau_2} g(\nabla_\mathbf{t}\mathbf{t}, \nabla_\mathbf{v}\mathbf{v}) d\tau - \int_{\tau_1}^{\tau_2} g(\mathbf{t}, \nabla_\mathbf{t}\nabla_\mathbf{v}\mathbf{v}) d\tau.\end{aligned}$$

(Note that $\nabla_\mathbf{v}\mathbf{v} \neq 0$ away from the curve γ.) Since $[\mathbf{v}, \mathbf{t}] = 0$, the term $\nabla_\mathbf{t}\nabla_\mathbf{v}\mathbf{v}$ is expressed through the Riemann tensor as

$$g(\mathbf{t}, \nabla_\mathbf{t}\nabla_\mathbf{v}\mathbf{v}) = R(\mathbf{t}, \mathbf{v}, \mathbf{v}, \mathbf{t}) + g(\mathbf{t}, \nabla_\mathbf{v}\nabla_\mathbf{v}\mathbf{t}).$$

Finally, on the geodesic curve γ we have $\nabla_\mathbf{v}\mathbf{v} = 0$ and thus

$$\frac{d^2}{d\sigma^2} L[\gamma; \mathbf{t}] = \int_{\tau_1}^{\tau_2} \left(R(\mathbf{v}, \mathbf{t}, \mathbf{v}, \mathbf{t}) - g(\mathbf{t}, \nabla_\mathbf{v}\nabla_\mathbf{v}\mathbf{t}) \right) d\tau.$$

The integral expression in the right-hand side above must be nonpositive for all vector fields $\mathbf{t}$ (and for all τ) if the curve γ maximizes the proper length under infinitesimal variations. We cannot simplify $d^2L/d\sigma^2$ any further, but we note that the expression in the right-hand side is similar to the geodesic deviation equation (4.2). Denoting by $\hat{G}$ the linear operator involved in that equation, we can rewrite $d^2L/d\sigma^2$ as

$$\frac{d^2}{d\sigma^2} L[\gamma; \mathbf{t}] = \int_{\tau_1}^{\tau_2} g(\mathbf{t}, \hat{G}\mathbf{t} - \ddot{\mathbf{t}}) d\tau,$$

[4] Generally, the focal intersection exists *only* among infinitesimally close geodesics! This is so because the connecting vector $\mathbf{c}$ will generally fail to satisfy the geodesic deviation equation for a neighbor geodesic at a finite distance from γ.

where the precise definition of $\hat{G}$ is

$$\hat{G}\mathbf{t} \equiv R(\mathbf{v},\mathbf{t})\mathbf{v}, \quad \ddot{\mathbf{t}} \equiv \partial_\tau\partial_\tau\mathbf{t} \equiv \nabla_\mathbf{v}\nabla_\mathbf{v}\mathbf{t}.$$

The vector $\mathbf{t}$ will be a Jacobi field if $\ddot{\mathbf{t}} - \hat{G}\mathbf{t} = 0$. Thus, $d^2L/d\sigma^2$ will vanish if $\mathbf{t}$ is a Jacobi field. In the present case, the field $\mathbf{t}$ plays the role of a "perturbation field" and the curve $\gamma(\tau)$ is fixed at endpoints $\gamma(\tau_1)$ and $\gamma(\tau_2)$. Thus an admissible field $\mathbf{t}$ must vanish at endpoints. But a Jacobi field that vanishes at the endpoints can exist only if both endpoints are conjugate points with respect to the curve γ. The functional $d^2L[\gamma;\mathbf{t}]/d\sigma^2$ is a quadratic functional of the vector field $\mathbf{t}$. Heuristically, one expects that a negative-definite quadratic functional will not vanish at nonzero values of its argument. Hence, we expect that the proper length cannot be maximized if there exists a Jacobi field that vanishes at both the endpoints, or (as will be shown below) if one of the endpoints is conjugate to a point within the curve. Thus we found a crucial link between the existence of conjugate points and the maximization of proper length. This link is formalized in the following statement.

Statement 1: [5]A geodesic curve γ maximizes the proper length between points $\gamma(\tau_1)$ and $\gamma(\tau_2)$ with respect to infinitesimal variations iff γ has no conjugate points to $\gamma(\tau_1)$ within the interval $[\tau_1, \tau_2]$.

Sketch of proof: If there are no conjugate points then there is a one-to-one map between a Jacobi field $\mathbf{c}$ at a point $\gamma(\tau)$ and the initial value of the derivative $\nabla_\mathbf{u}\mathbf{c}$ at $\tau = 0$. This map is specified by a nondegenerate transformation $A(\tau)$. The functional $d^2L[\gamma;\mathbf{t}]/d\sigma^2$ evaluated on an arbitrary vector $\mathbf{t}$ can be expressed through the auxiliary vector field $\mathbf{z} \equiv A^{-1}\mathbf{t}$, and shown to be a negative-definite functional of $\mathbf{z}$ by an explicit calculation that uses the integrability of the geodesic congruence emitted at $\gamma(\tau_1)$.

Conversely, if there exists a focal point $\gamma(\tau_*)$, where τ_* is between τ_1 and τ_2, then there exists a Jacobi field $\mathbf{c}$, not identically zero, and such that $\mathbf{c}(\tau_1) = \mathbf{c}(\tau_*) = 0$. To show that the proper length is not maximized, it is sufficient to find even a single example of a vector field $\mathbf{t}$ such that $d^2L[\gamma;\mathbf{t}]/d\sigma^2 > 0$. If we define the field $\mathbf{t}_0$ along the curve $\gamma(\tau)$ by $\mathbf{t}_0(\tau) \equiv \mathbf{c}(\tau)$ for $\tau_1 < \tau < \tau_*$ and $\mathbf{t}_0(\tau) \equiv 0$ for $\tau_* < \tau < \tau_2$, then we will have $d^2L[\gamma;\mathbf{t}_0]/d\sigma^2 = 0$ except at $\tau = \tau_*$ where $\mathbf{t}_0(\tau)$ has a "corner." Now, one can deform this field $\mathbf{t}_0(\tau)$ by an infinitesimal amount and construct a smooth field $\mathbf{t}(\tau)$ such that still $d^2L[\gamma;\mathbf{t}]/d\sigma^2 > 0$. The construction of $\mathbf{t}(\tau)$ goes heuristically as follows: The field $\mathbf{t}(\tau)$ is almost equal to $\mathbf{c}(\tau)$ up to $\tau = \tau_*$, which produces an almost zero value of the integral, and then $\mathbf{t}(\tau)$ can be chosen to be almost zero for $\tau_* < \tau < \tau_2$ and such that the resulting integral is positive. (The curve corresponding to $\mathbf{t}_0(\tau)$ has a corner that can be "straightened out.")

More detailed proof: First we prove that there is a maximum of $L[\gamma]$ if no conjugate points are present on γ. For brevity, we denote (covariant) derivatives along the curve γ by an overdot. The time-dependent transformation $\hat{A}(\tau)$ that maps initial values of $\dot{\mathbf{c}}(\tau_1)$ to final values $\mathbf{c}(\tau)$ is the (transformation-valued) solution of the equation

$$\ddot{\hat{A}} - \hat{G}\hat{A} = 0, \quad \hat{A}(\tau_1) = 0, \ \dot{\hat{A}}(\tau_1) = \hat{1}.$$

Note that $\hat{G}\mathbf{v} = 0$ and $g(\hat{G}\mathbf{x},\mathbf{v}) = 0$ for all $\mathbf{x}$, in other words, $\hat{G}$ is transverse to the vector $\mathbf{v}$. Thus $\hat{G}$ and $\hat{A}$ are effectively three-dimensional transformations acting in the subspace $\mathbf{v}^\perp$.

[5][12], Proposition 4.5.8.

If there are no conjugate points, the transformation $\hat{A}$ is nondegenerate for all $\tau > \tau_1$. In that case, we will now show that $d^2L[\gamma;\mathbf{t}]/d\sigma^2 \le 0$ for all vector functions $\mathbf{t}(\tau)$ that satisfy $\mathbf{t}(\tau_1) = \mathbf{t}(\tau_2) = 0$. This will be sufficient to show that $L[\gamma]$ has a maximum. Since $\hat{A}^{-1}$ is well-defined, we can set $\mathbf{z}(\tau) \equiv \hat{A}^{-1}(\tau)\mathbf{t}(\tau)$; the vector $\mathbf{z}(\tau)$ will then belong to the subspace $\mathbf{v}^\perp$. We can express $d^2L[\gamma;\mathbf{t}]/d\sigma^2$ through $\mathbf{z}(\tau)$ as follows,

$$\begin{aligned}\frac{d^2}{d\sigma^2}L[\gamma;\mathbf{t}] &= \int_{\tau_1}^{\tau_2} g(\hat{A}\mathbf{z}, \hat{G}\hat{A}\mathbf{z} - \partial_\tau\partial_\tau(\hat{A}\mathbf{z}))\mathrm{d}\tau \\ &= \int_{\tau_1}^{\tau_2} \left[-g(\hat{A}\mathbf{z}, 2\dot{\hat{A}}\dot{\mathbf{z}}) - g(\hat{A}\mathbf{z}, \hat{A}\ddot{\mathbf{z}})\right]\mathrm{d}\tau \\ &= \int_{\tau_1}^{\tau_2} \left[g(\dot{\hat{A}}\mathbf{z}, \hat{A}\dot{\mathbf{z}}) - g(\hat{A}\mathbf{z}, \dot{\hat{A}}\dot{\mathbf{z}}) + g(\hat{A}\dot{\mathbf{z}}, \hat{A}\dot{\mathbf{z}}\right]\mathrm{d}\tau,\end{aligned} \tag{4.3}$$

where in the last line we have canceled the total derivative $\partial_\tau g(\hat{A}\mathbf{z}, \hat{A}\dot{\mathbf{z}})$. Now let us rewrite the first two terms in the last line in Eq. (4.3) through the adjoint operator A^T,

$$g(\dot{\hat{A}}\mathbf{z}, \hat{A}\dot{\mathbf{z}}) - g(\hat{A}\mathbf{z}, \dot{\hat{A}}\dot{\mathbf{z}}) = g(\mathbf{z}, (\dot{\hat{A}}^T\hat{A} - \hat{A}^T\dot{\hat{A}})\dot{\mathbf{z}}).$$

Note that $g(\hat{G}\mathbf{x},\mathbf{y}) = R(\mathbf{v},\mathbf{x},\mathbf{v},\mathbf{y})$, so the known symmetry of the Riemann tensor, $R(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d}) = R(\mathbf{c},\mathbf{d},\mathbf{a},\mathbf{b})$, means that $\hat{G}$ is self-adjoint, $\hat{G}^T = \hat{G}$. Thus, the operator $\hat{A}^T(\tau)$ satisfies the equation

$$\ddot{\hat{A}}^T - \hat{A}^T\hat{G} = 0, \quad \hat{A}^T(\tau_1) = 0, \ \dot{\hat{A}}^T(\tau_1) = \hat{1}.$$

The operator $\dot{\hat{A}}^T\hat{A} - \hat{A}^T\dot{\hat{A}}$ is analogous to a Wronskian and is time-independent because

$$\partial_\tau\left(\dot{\hat{A}}^T\hat{A} - \hat{A}^T\dot{\hat{A}}\right) = \hat{A}^T\hat{G}\hat{A} - \hat{A}^T\hat{G}\hat{A} = 0.$$

Since the operator $\dot{\hat{A}}^T\hat{A} - \hat{A}^T\dot{\hat{A}}$ is equal to zero at $\tau = \tau_1$ due to the initial condition $\hat{A}(\tau_1) = 0$, we have $\dot{\hat{A}}^T\hat{A} - \hat{A}^T\dot{\hat{A}} = 0$ for all τ. Thus

$$g(\dot{\hat{A}}\mathbf{z}, \hat{A}\dot{\mathbf{z}}) - g(\hat{A}\mathbf{z}, \dot{\hat{A}}\dot{\mathbf{z}}) \equiv 0$$

and the expression (4.3) is simplified to

$$\frac{d^2}{d\sigma^2}L[\gamma;\mathbf{t}] = \int_{\tau_1}^{\tau_2} g(\hat{A}\dot{\mathbf{z}}, \hat{A}\dot{\mathbf{z}})\mathrm{d}\tau.$$

Since the vector $\hat{A}\dot{\mathbf{z}}$ belongs to the subspace $\mathbf{v}^\perp$, it is spacelike, hence the above expression is always nonpositive.

Now we consider the second case where a conjugate point is present. We will then show that $L[\gamma]$ does not have a maximum. The presence of a conjugate point means that there exists a Jacobi field $\mathbf{c}(\tau)$, not everywhere zero, satisfying $\ddot{\mathbf{c}} = \hat{G}\mathbf{c}$, $\mathbf{c}(\tau_1) = 0$, $\mathbf{c}(\tau_*) = 0$ for $\tau_1 < \tau_* < \tau_2$. We shall now demonstrate that $d^2L[\gamma;\mathbf{t}]/d\sigma^2 > 0$ with a specific choice of the perturbation vector $\mathbf{t}$. The vector $\mathbf{t}$ will be defined by "smoothing the corner" of $\mathbf{c}(\tau)$ near $\tau = \tau_*$, with a small additional perturbation:

$$\mathbf{t}(\tau) \equiv s(\tau)\mathbf{c}(\tau) + \varepsilon q(\tau)\mathbf{a},$$

where $s(\tau)$ is a smooth step-like function such that $s(\tau < \tau_*) \approx 1$, while $s(\tau > \tau_*) \approx 0$; $\varepsilon > 0$ is a "small" constant, $q(\tau)$ is a smooth function that vanishes at $\tau = \tau_{1,2}$ and is positive for $\tau_1 < \tau < \tau_2$, for example, $q(\tau) = (\tau - \tau_1)(\tau_2 - \tau)$; and $\mathbf{a}$ is a *constant* spacelike vector (to be determined later).

The function $s(\tau)$ is chosen to be nonconstant only within a very narrow interval around $\tau = \tau_*$. The idea is to have ε very small and $s(\tau)$ very close to the Heaviside step function, so that $\mathbf{t}(\tau)$ is smooth but $\mathbf{t}(\tau) \approx \mathbf{c}(\tau)$ for $\tau_1 < \tau < \tau_*$ and $\mathbf{t}(\tau) \approx 0$ for $\tau_* < \tau < \tau_2$. By construction, $\mathbf{t}$ exactly vanishes at the endpoints $\tau = \tau_{1,2}$. With the above definition of $\mathbf{t}(\tau)$, we can directly compute the relevant quantities, neglecting unimportant terms of order ε^2:

$$\hat{G}\mathbf{t} - \ddot{\mathbf{t}} = \varepsilon q(\tau)\hat{G}\mathbf{a} - 2\dot{s}\dot{\mathbf{c}} - \ddot{s}\mathbf{c},$$
$$\frac{d^2}{d\sigma^2}L[\gamma;\mathbf{t}] = \int_{\tau_1}^{\tau_2} g(\mathbf{t}, \hat{G}\mathbf{t} - \ddot{\mathbf{t}})\mathrm{d}\tau = -\int_{\tau_1}^{\tau_2} g(s\mathbf{c}, 2\dot{s}\dot{\mathbf{c}} + \ddot{s}\mathbf{c})\mathrm{d}\tau$$
$$- 2\varepsilon\int_{\tau_1}^{\tau_2} q(\tau)g(\mathbf{a}, 2\dot{s}\dot{\mathbf{c}} + \ddot{s}\mathbf{c})\mathrm{d}\tau + O(\varepsilon^2),$$

where in the last line we have used the self-adjointness property (derived using integration by parts),

$$\int_{\tau_1}^{\tau_2} g(\mathbf{x}, \hat{G}\mathbf{y} - \ddot{\mathbf{y}})\mathrm{d}\tau = \int_{\tau_1}^{\tau_2} g(\hat{G}\mathbf{x} - \ddot{\mathbf{x}}, \mathbf{y})\mathrm{d}\tau,$$

which holds for vectors $\mathbf{x}, \mathbf{y}$ that vanish at endpoints $\tau = \tau_{1,2}$. Since $\dot{s}$ and $\ddot{s}$ are nonzero only in a narrow interval around $\tau = \tau_*$, it is sufficient to approximate

$$\mathbf{c}(\tau) \approx (\tau - \tau_*)\,\mathbf{b}, \quad \mathbf{b} \equiv \dot{\mathbf{c}}(\tau_*)$$
$$\mathbf{t}(\tau) \approx (\tau - \tau_*)\,s(\tau)\mathbf{b} + \varepsilon q(\tau_*)\mathbf{a},$$

where $\mathbf{b}$ is a known spacelike vector. (With a suitable choice of the function $s(\tau)$, the error of this approximation will be of order ε^2.) Therefore,

$$\frac{\mathrm{d}^2 L[\mathbf{t}]}{\mathrm{d}\sigma^2} = O(\varepsilon^2) - g(\mathbf{b},\mathbf{b})\int_{\tau_1}^{\tau_2}(\tau - \tau_*)\,(2s\dot{s} + (\tau - \tau_*)s\ddot{s})\,\mathrm{d}\tau$$
$$- 2\varepsilon q(\tau_*)g(\mathbf{a},\mathbf{b})\int_{\tau_1}^{\tau_2}(2\dot{s} + (\tau - \tau_*)\ddot{s})\,\mathrm{d}\tau.$$

The step-like function s can be chosen so that the first integral in the above equation is arbitrarily small, e.g. of order ε^2. Indeed, integration by parts yields

$$\int_{\tau_1}^{\tau_2}(\tau - \tau_*)\,(2s\dot{s} + (\tau - \tau_*)s\ddot{s})\,\mathrm{d}\tau = -\int_{\tau_1}^{\tau_2}(\tau - \tau_*)^2\dot{s}^2\mathrm{d}\tau;$$

heuristically, $\dot{s}(\tau) \approx \delta(\tau - \tau_*)$ and thus $(\tau - \tau_*)\dot{s} \approx 0$. These statements can be made mathematically precise, for instance, by using an explicit formula for $s(\tau)$. The remaining integral is easily evaluated,

$$\int_{\tau_1}^{\tau_2}(2\dot{s} + (\tau - \tau_*)\ddot{s})\,\mathrm{d}\tau = \int_{\tau_1}^{\tau_2}[\dot{s} + \partial_\tau\,((\tau - \tau_*)\dot{s})]\,\mathrm{d}\tau = -1.$$

Finally, we note that $\mathbf{b} \neq 0$, since by assumption $\mathbf{c}$ is not everywhere zero and $\mathbf{b} = \dot{\mathbf{c}} \neq 0$ at the point $\tau = \tau_*$ where $\mathbf{c} = 0$. Also, $q(\tau_*) > 0$. So we can choose the vector $\mathbf{a}$ such that

$$2q(\tau_*)g(\mathbf{a},\mathbf{b}) = 1.$$

With this choice, we obtain

$$\frac{d^2}{d\sigma^2}L[\gamma;\mathbf{t}] = \varepsilon + O(\varepsilon^2),$$

which will be positive for small enough ε. This concludes the proof of Statement 1.

Thus, in the presence of (at least one) conjugate point on a curve γ, the maximum proper length is not achieved by γ. For example, a body on a circular orbit in a central gravitational field does not maximize the proper time over one period after it comes back to an initial point p, since the elapsed proper time is smaller than that of a body at rest at p. In fact, the proper time is maximized by another geodesic, corresponding to a body thrown away from the center of gravity from p such that it comes back to p at the right time.

A related situation is a congruence of timelike geodesics normal to a given spacelike 3-surface S. Then this congruence is hypersurface orthogonal everywhere (not just at S), by an argument analogous to that of Statement 1 in Sec. 2.4.2. For a point p outside of S, there may or may not be a geodesic that maximizes the proper distance between p and S. If a maximizing geodesic exists, it must be one of the congruence normal to S. However, maximization is impossible if the congruence has focal points between p and S. As before, a definition can be given in terms of Jacobi fields: A point q on a geodesic γ normal to S is called **conjugate to** S **with respect to** γ if there exists a Jacobi field $\mathbf{c}$ along γ, such that $\mathbf{c} = 0$ at q but $\mathbf{c} \neq 0$ on S. Similarly to Statement 1 above, one can prove the following.

Statement 2: A timelike curve γ maximizes the proper distance between a point p and a spacelike 3-surface S if γ is a geodesic normal to S which contains no points conjugate to S before the point p.

The focusing theorem now shows that conjugate points must exist under certain conditions.

Statement 3: Suppose that S is a spacelike 3-surface, $\mathbf{u}$ is a normalized timelike vector field orthogonal to S, such that $\mathrm{div}\mathbf{u} \equiv -\theta_0 < 0$ at some point $p \in S$, and the strong energy condition holds, $\mathrm{Ric}(\mathbf{x},\mathbf{x}) \leq 0$ for all timelike $\mathbf{x}$. Then the geodesic γ normal to S at the point p contains a conjugate point to S within proper time $3/\theta_0$ from p.

Under the conditions of Statement 3, the geodesic γ normal to S at p that stretches up to a point q which is further than $3/\theta_0$ from p (as measured along γ), cannot maximize the proper length between q and S. Our conclusion is the following.

Statement 4: Suppose the conditions of Statement 3 hold, and additionally $\mathrm{div}\mathbf{u} \leq -\theta_0 < 0$ everywhere on the surface S. If a point q can be connected to S by a timelike curve of proper length $\geq 3/\theta_0$ then there is no timelike curve of maximum length connecting q and S.

Proof: If a timelike curve γ of maximal proper length exists and reaches S at a point p then γ must be a timelike geodesic that is orthogonal to S at p (or else we can deform γ or shift the point p and increase the proper length). However, if the proper length of γ is larger than $3/\theta_0$, there exists a point on γ conjugate to p, and thus γ does not maximize the proper length.

An example of this situation is a half-hyperboloid $t^2 - r^2 = 1$, $t < 0$, in the Minkowski spacetime. This spacelike 3-surface is orthogonal to every straight line (i.e. to every geodesic) passing through the origin. Thus, the origin is the focal point for all the geodesics emitted orthogonally from S. The half-lightcone, $t^2 - r^2 = 0$, $t < 0$, bounds the domain where conjugate points do not arise. For any point q such that $t^2 - r^2 > 0$, $t \leq 0$, and for any point in the upper half-spacetime, $t > 0$,

one can find timelike curves of arbitrarily large proper length leading from q to some point on S. Thus, the maximal-length curve connecting some point q to S must have proper length < 1 between q and S.

Calculation: Compute the divergence of the future-directed normal vector field $\mathbf{u}$ to the half-hyperboloid $t^2 - x^2 - y^2 - z^2 = 1$, $t < 0$, in the Minkowski spacetime, and show that this 3-surface satisfies the conditions of Statement 4 as well as its conclusion.

Solution: The hyperboloid is specified by the equation $g(\mathbf{x}, \mathbf{x}) = 1$, where $\mathbf{x} \equiv (t, x, y, z)$ is the position vector. The future-directed normal vector field is

$$\mathbf{u} = -\frac{\mathbf{x}}{\sqrt{g(\mathbf{x}, \mathbf{x})}},$$

and the Levi-Civita connection $\nabla = \partial$, so we have $\nabla_{\mathbf{a}}\mathbf{x} = \mathbf{a}$ and thus (on the surface)

$$\nabla_{\mathbf{a}}\mathbf{u} = -\partial_{\mathbf{a}}\frac{\mathbf{x}}{\sqrt{g(\mathbf{x}, \mathbf{x})}} = -\mathbf{a} + \mathbf{x}g(\mathbf{a}, \mathbf{x}).$$

The divergence is found as

$$\begin{aligned}\mathrm{div}\mathbf{u} &= \mathrm{Tr}_{(\mathbf{a},\mathbf{b})}g(\nabla_{\mathbf{a}}\mathbf{u}, \mathbf{b}) \\ &= \mathrm{Tr}_{(\mathbf{a},\mathbf{b})}\left[g(\mathbf{a}, \mathbf{x})g(\mathbf{b}, \mathbf{x}) - g(\mathbf{a}, \mathbf{b})\right] = 1 - 4 = -3.\end{aligned}$$

So we may set $\theta_0 = 3$ and apply Statement 4. The proper length to the focus is $3/\theta_0 = 1$, in agreement with the geometric result.

Finally, we state an analogous theorem for null geodesics. The differences between the timelike and the null case are purely technical: 1) Jacobi fields for null geodesics $\mathbf{n}$ are effectively two-dimensional because connecting vectors differing by a multiple of $\mathbf{n}$ are equivalent. 2) Null geodesics containing conjugate points do not maximize the proper length between endpoints, thus the maximal proper length is positive, and there is a timelike curve connecting the endpoints. 3) One considers a spacelike 2-surface to which null geodesics are orthogonal (note that there are two null directions orthogonal to a spacelike 2-surface, but no null vectors orthogonal to a spacelike 3-surface). 4) One uses the null energy condition instead of the strong energy condition. 5) The focusing theorem for null geodesics predicts focusing within parameter interval $2/\theta_0$, instead of $3/\theta_0$ in the timelike case.

Statement 5: Suppose that S is a spacelike 2-surface, $\mathbf{n}$ is a null vector field orthogonal to S, such that $\mathrm{div}\mathbf{n} \equiv -\theta_0 < 0$ at a point $p \in S$, and the null energy condition holds, $\mathrm{Ric}(\mathbf{x}, \mathbf{x}) \leq 0$ for all null $\mathbf{x}$. Then the null geodesic γ normal to S at p, with initial tangent vector $\mathbf{n}$, contains a conjugate point to S within the affine parameter interval $2/\theta_0$. If the geodesic γ can be extended beyond the conjugate point to some point q then γ does not maximize the proper length between p and q, and thus q can be connected to p by a timelike curve.

4.1.5 Singularity in collapsing or expanding universe

We shall now derive some basic singularity theorems. Both theorems below are due to S. W. Hawking. (Stronger theorems exist but require much more technical work.)

An intuitive picture of a gravitational collapse is that of a cloud of particles that keep moving closer to each other. The mathematical expression of this property is that the congruence of the particle worldlines has a negative divergence. According to the focusing theorem, a consequence of negative divergence is a focusing of the geodesics. As a result, the geodesics fail to maximize proper length. This is the technical basis of the singularity theorems. Of course, the geodesic focusing by itself does not signal any singularities; we need additional conditions on the initial velocities of the particles to conclude that singularities will develop. In the first singularity theorem, these conditions are formulated as requirements on a spacelike 3-surface to which the particle worldlines are orthogonal.

Recall that a **Cauchy surface** Σ for a domain is a 3-surface such that any timelike curve within the domain intersects Σ exactly once. The intuitive meaning of a Cauchy surface is that a partial differential equation has a unique solution throughout the domain if appropriate initial data are specified on Σ. The theorem states that a Cauchy surface having an everywhere negative divergence leads to a singularity.

Statement 1: If Σ is a spacelike Cauchy surface for the entire spacetime, with a normal vector field $\mathbf{u}$ whose divergence is everywhere negative-bounded, $\mathrm{div}\mathbf{u} \leq \theta_0 < 0$, and if the strong energy condition holds, then every timelike geodesic normal to Σ has finite proper length no greater than $3/\theta_0$.

Sketch of a proof: Consider a point p to the future of Σ. Every timelike geodesic leading from p to Σ must intersect Σ exactly once. The points of intersection form an open set $I(\Sigma; p)$ bounded by intersection points with null geodesics from p to Σ. Timelike geodesics intersecting Σ near the boundary of $I(\Sigma; p)$ have very small proper length, so we may consider the subset $\tilde{I}(\Sigma; p; L) \subset I(\Sigma; p)$ of intersection points for curves with proper length $\geq L$ for some lower bound L. We can select L such that $\tilde{I}(\Sigma; p; L) \neq \emptyset$, and then the subset $\tilde{I}(\Sigma; p; L)$ is compact. (These intuitively clear topological statements will be given a formal proof in Lemma 1 below.) Therefore, the proper time is maximized by *some* geodesic curve leading from p to Σ. However, by Statement 4 in the previous section, we know that there are *no* curves maximizing the proper length and stretching further than $3/\theta_0$. Therefore the spacetime cannot include points at a proper length greater than $3/\theta_0$ from Σ. In other words, every timelike geodesic emitted from Σ must be *future incomplete*.

Lemma 1: Let $\tilde{I}(\Sigma; p; L)$ be the set of intersection points of timelike geodesics leading from a point p to a Cauchy surface Σ, whose proper length is $\geq L$, where $L > 0$ is a given constant. Then $\tilde{I}(\Sigma; p; L)$ is a compact set.

Proof: Suppose that the set $\tilde{I}(\Sigma; p; L)$ is not compact. Then there exists a sequence $\{\gamma_j\}$ of geodesics, whose proper lengths are $\geq L$, such that the corresponding intersection points $q_1, q_2, ...$, on Σ constitute a sequence without accumulation points within $I(\Sigma; p; L)$. However, the initial tangent vectors $\dot{\boldsymbol{\gamma}}_j$ at p are within the past lightcone within $T_p\mathcal{M}$, which is a compact set if we also include the null directions. Hence, the sequence $\left\{\dot{\boldsymbol{\gamma}}_j\right\}$ of tangent vectors must have an accumulation point $\dot{\boldsymbol{\gamma}}_\infty$ within the past lightcone. We can emit a geodesic $\gamma_\infty(\tau)$ from p with the initial tangent vector $\dot{\boldsymbol{\gamma}}_\infty$. If $\gamma_\infty(\tau)$ intersects Σ at a point q_∞ then q_∞ must be an accumulation point for $\{q_j\}$. Hence the proper length of γ_∞ cannot be less than L and so $q_\infty \in I(\Sigma; p; L)$. This yields a contradiction with the assumption that $\{q_j\}$ has no accumulation points within $I(\Sigma; p; L)$. Thus $\gamma_\infty(\tau)$ cannot intersect Σ. Since every time-

like curve intersects Σ, we see that the vector $\dot{\gamma}_\infty$ must be null and $\gamma_\infty(\tau)$ is a null geodesic. Since $\gamma_\infty(\tau)$ does not intersect Σ, for any τ there exists a neighborhood of $\gamma_\infty(\tau)$ that does not intersect Σ. It is then possible to find a timelike curve (which is not necessarily a geodesic, not necessarily starting at p) that runs along $\gamma_\infty(\tau)$ for all τ and always remains away from Σ. The existence of an inextendible timelike curve that never crosses Σ contradicts the definition of a Cauchy surface. This concludes the proof.

An interpretation of the first theorem is the following: If we see a universe that is everywhere contracting (everywhere along a single Cauchy surface), we should expect to hit a singularity within a finite time. By inverting the direction of time, we also come to the conclusion that a Cauchy surface with everywhere positive divergence contains a singularity in its past. Thus, if our universe is everywhere expanding, and if the strong or the null energy condition is satisfied, then the universe must have begun at a singularity at a finite time in the past.

Self-test question: The half-hyperboloid $t^2 - x^2 - y^2 - z^2 = 1$, $t < 0$, in the Minkowski spacetime, is an infinite spacelike surface whose normal vector $\mathbf{x}$ has an everywhere negative divergence. Does it follow from Statement 1 that the Minkowski spacetime must contain a singularity to the future of the half-hyperboloid?

Answer: No. (The half-hyperboloid is not a Cauchy surface.)

4.1.6 Singularity in a closed universe

The assumption that a given 3-surface is a Cauchy surface for the spacetime is quite strong, since it is a global relationship between Σ and the entire spacetime, and not merely a local property of the surface Σ. One can replace this assumption by the requirement that Σ be *compact* (this will apply to closed universes). The result will be the incompleteness of *some* geodesics emitted from Σ, which may fill only a part of the entire spacetime. Here is how this conclusion can be derived.

We restrict our attention to the part of the spacetime consisting of points p for which there exists a timelike curve γ between Σ and p. This part of the spacetime is called the **domain of dependence** $\mathcal{D}(\Sigma)$ of the surface Σ. We would like to prove that $\mathcal{D}(\Sigma)$ contains some incomplete geodesics if Σ satisfies the conditions of Statement 1 above, except for being a Cauchy surface. It follows from the proof of Statement 1 that, within $\mathcal{D}(\Sigma)$, a geodesic γ connecting Σ to a point p at proper length $> 3/\theta_0$ does not maximize the proper length. Therefore, any geodesic that *does* maximize the proper length must be shorter than $3/\theta_0$.

Let us denote by $\mathcal{C}(\Sigma) \equiv \partial\mathcal{D}(\Sigma)$ the boundary of the domain of dependence; it is a 3-surface called the **Cauchy horizon**. It can be seen that the Cauchy horizon $\mathcal{C}(\Sigma)$ is locally a null surface (although it may also contain some "corners"). The fact that a boundary $\partial\mathcal{D}$ of a domain of dependence $\mathcal{D}$ is a null 3-surface will be derived below (see the proof of Statement 1 in Sec. 4.2). There are now two interesting possibilities: either $\mathcal{C}(\Sigma)$ is compact, or it is not compact. Suppose first that $\mathcal{C}(\Sigma)$ is a compact set. Note that a null surface is generated by null geodesics (see Sec. 2.2.8). Then the null generators of $\mathcal{C}(\Sigma)$ must wind around it, coming arbitrarily near to themselves infinitely many times; informally, we may call such curves "almost closed" curves. Since timelike geodesics can be found arbitrarily close to null geodesics, the presence of "almost closed" null geodesics is an undesirable feature of the spacetime, similar to a causality violation or the presence of a time machine. A spacetime is called **strongly causal** if it does not admit timelike curves that repeatedly come arbitrarily close to some points. Therefore, $\mathcal{C}(\Sigma)$ cannot be compact in a strongly causal spacetime. Now suppose $\mathcal{C}(\Sigma)$ is not compact; then $\mathcal{C}(\Sigma)$ contains a sequence of points q_1, q_2, ..., which does not have an accumulation point within $\mathcal{C}(\Sigma)$. We may choose a sequence of points $\{\tilde{q}_j\}$ arbitrarily close to $\{q_j\}$ but within $\mathcal{D}(\Sigma)$. Then, every point $\tilde{q}_j$ lies on at least one timelike curve intersecting Σ. Suppose that L is the length of that curve; then the set of all timelike curves with length $\geq L$ is compact, so one of these curves has the largest proper length. So, for each point $\tilde{q}_j$ we have a timelike geodesic $\gamma_j(\tau)$ that has the maximal proper length among all the timelike curves connecting $\tilde{q}_j$ with Σ. (We have already shown that each of these curves has proper length $\tau_{\max} < 3/\theta_0$, assuming that $\gamma_j(0)$ is on Σ.) Let p_j be the point where the curve γ_j intersects Σ, and let $\mathbf{u}_j \equiv \dot{\gamma}_j(0)$ be the initial tangent vector. The intersection points p_1, p_2, ..., and the corresponding initial tangent vectors $\mathbf{u}_1$, $\mathbf{u}_2$, ..., are sequences on a compact set Σ and thus must accumulate at some point p_∞ and vector $\mathbf{u}_\infty$. Now, we can show that the geodesic $\gamma_\infty(\tau)$ emitted from Σ at p_∞ with the initial tangent vector $\mathbf{u}_\infty$ cannot be extended past the proper time $3/\theta_0$. If $\gamma_\infty(\tau)$ were well-defined for $\tau > 3/\theta_0$ then we would have a narrow tube-shaped neighborhood stretching along $\gamma_\infty(\tau)$ for $0 < \tau \leq 3/\theta_0$, such that infinitely many geodesics γ_j are inside that tube. Thus the point sequences $\{\tilde{q}_j\}$ and therefore $\{q_j\}$ would accumulate somewhere within the (compact) neighborhood of γ_∞. This contradicts the assumption that the sequence $\{q_j\}$ has no accumulation points.

Thus we have proved the following theorem.

Statement 2: If a spacetime contains a compact spacelike surface Σ (without boundary) having a normal vector $\mathbf{u}$ such that $\mathrm{div}\mathbf{u} < 0$ everywhere on Σ (thus there exists $\theta_0 < 0$ such that $\mathrm{div}\mathbf{u} \leq \theta_0$), and if the strong energy condition holds, and if the spacetime is strongly causal[6] (contains no almost-closed null geodesics), then there is at least one incomplete timelike geodesic emitted from Σ.

A universe can contain a compact, "edgeless" spacelike 3-surface only if the universe is *closed*. In open universes, every spacelike 3-surface without boundary must stretch to infinity and so cannot be compact. The interpretation of the second singularity theorem is that a closed universe must contain a singularity if there exists an everywhere contracting spacelike section.

Note that this theorem demonstrates only the existence of a *single* incomplete timelike geodesic, whereas Statement 1 of the previous section showed (under stronger assumptions) that *every* timelike geodesic is incomplete. The presence of incomplete geodesics emitted from Σ means that the spacetime contains a singularity somewhere to the future of Σ.

Analogous statements can be derived for null geodesics using the null energy condition.

4.1.7 Singularity in gravitational collapse

Qualitatively, gravitational collapse occurs when a mass m is concentrated within a region smaller than the Schwarzschild

[6] S. W. Hawking also proved a version of this theorem without the strong causality condition. **Reference?**

radius $R_s \equiv 2m$. The collapse results in a singularity, regardless of the composition of the collapsing matter.

In this section we shall consider a singularity theorem due to Penrose, showing that the formation of a singularity is inevitable once the normally "outgoing" lightrays start to converge. Let us investigate this condition in more detail.

There exist precisely two null directions orthogonal to a 2-dimensional subspace spanned by two spacelike vectors. Thus, a spacelike 2-surface can emit two congruences of lightrays orthogonal to it. Normally, one congruence will be converging (the "ingoing" lightrays) and the other diverging (the "outgoing" lightrays). A 2-surface T is called a **trapped surface** if both congruences of lightrays emitted orthogonally to T into the future have a negative divergence.

Penrose's argument begins by considering a spherically symmetric distribution of matter. Suppose at first that a mass m is concentrated within a region of radius $r_0 < 2m$ in a spherically symmetric fashion. Then the metric outside of the matter distribution is Schwarzschild, and it is easy to see that at points r such that $r_0 < r < 2m$, both lightcones emitted into the future direction are converging (have negative divergence). Thus a sphere of radius r_1 is a compact, spacelike, trapped 2-surface for $r_0 < r_1 < 2m$. We shall shortly prove that the existence of such a trapped 2-surface indicates the presence of a singularity in the future. Now, if the matter distribution is slightly nonspherical, the sphere will remain a trapped surface because the divergence of the lightrays will be only slightly perturbed and will not become positive. Thus a singularity arises even for asymmetric configurations of collapsing matter.

The precise formulation of the theorem is the following.

Statement 1:[7] If a time-orientable spacetime manifold $\mathcal{M}$ has a non-compact Cauchy surface $\mathcal{C}$, if the null energy condition holds to the future of $\mathcal{C}$, and if there exists a compact, spacelike, trapped 2-surface T, then there exists at least one incomplete null geodesic to the future of T.

Proof: Consider the "future subdomain" $\mathcal{F}(T)$ of the surface T; this subdomain consists of all the points p that can be connected to T by some timelike curve. The proof of the theorem is based on studying the properties of the boundary $\partial\mathcal{F}(T)$ of the subdomain $\mathcal{F}(T)$. By considering geodesics near T, it is easy to see that the 3-surface $\partial\mathcal{F}$ starts from T as a null surface made by null geodesics emitted orthogonally from T. Since T is compact, the everywhere negative divergence of null geodesics emitted from T has an upper bound $-\theta_0 < 0$. Thus, by the focussing theorem, every null geodesic will have a conjugate point to T within a finite interval of the affine parameter (not larger than $2\theta_0^{-1}$). Let us assume that every null geodesic to the future of T is complete; in particular, extendable beyond the parameter interval $2\theta_0^{-1}$. We know that a geodesic does not maximize proper length beyond a conjugate point; therefore, points p on null geodesics beyond $2\theta_0^{-1}$ can be connected with T by a timelike geodesic with positive proper length. Thus, points on null geodesics beyond $2\theta_0^{-1}$ are inside $\mathcal{F}(T)$, and so the boundary $\partial\mathcal{F}$ consists of finite segments of null geodesics, each segment not longer than $2\theta_0^{-1}$. Since T is compact, the set $\partial\mathcal{F}$ can be viewed as a subset of a compact set $T \times [0, 2\theta_0^{-1}]$, therefore $\partial\mathcal{F}$ is itself compact. Since $\mathcal{M}$ is time-orientable, there exists a smooth vector field $\mathbf{v}$ such that $g(\mathbf{v},\mathbf{v}) = 1$. By definition of the Cauchy surface, every orbit of $\mathbf{v}$ intersects $\mathcal{C}$ exactly once. Also, an orbit of $\mathbf{v}$ can intersect the 3-surface $\partial\mathcal{F}$ at most once; this follows from the definition of $\partial\mathcal{F}$ as the boundary of the set $\mathcal{F}(T)$. Thus, the orbits of $\mathbf{v}$ define a map from $\partial\mathcal{F}$ to $\mathcal{C}$: a point $p \in \partial\mathcal{F}$ is mapped into the intersection of the corresponding timelike curve with $\mathcal{C}$. This map is continuous and one-to-one between $\partial\mathcal{F}$ and its image; however, $\partial\mathcal{F}$ is compact while $\mathcal{C}$ is not, which is an impossible situation. (The image of $\partial\mathcal{F}$ must be a compact subset of $\mathcal{C}$, so it must have a nonempty boundary within $\mathcal{C}$ which will be mapped to a boundary of $\partial\mathcal{F}$, and yet the boundary of $\partial\mathcal{F}$ is empty since $\partial\partial = 0$.) Therefore we must reject the assumption that all null geodesics can be extended beyond $2\theta_0^{-1}$; in other words, there must exist an incomplete null geodesic to the future of T.

[7] [12], §8.2, Theorem 1. See also the paper [24].

Calculation: Using the Schwarzschild metric (1.38), compute the affine tangent vector $\mathbf{n}$ for radial null geodesics emitted "outwards" from a sphere of radius r. Show that the divergence of $\mathbf{n}$ is $\sim r^{-1}$ for $r > 2m$, zero for $r = 2m$, and $\sim -r^{-1}$ for $r < 2m$.

Hint: In the Schwarzschild metric, a radial null geodesic has an affine tangent vector of the form $\mathbf{n} = \alpha_1\partial_t + \alpha_2\partial_r$, where $\alpha_{1,2}$ are some functions of r only. Lightrays emitted "outwards" at $r = 2m$ have the tangent vector in the direction ∂_t. For $r < 2m$, the radial coordinate has the meaning of "time," and the "future" is the direction of *decreasing* values of r (because, e.g., the proper time of a timelike observer falling into the black hole will increase with decreasing r).

Solution: We begin with the case $r = 2m$. The Schwarzschild metric is ill-defined at $r = 2m$, so we shall use a geometric argument instead of explicit calculations: The null vector ∂_t is divergence-free because the cross-section area of the null surface $r = 2m$ is $4\pi r^2$ and thus remains constant along ∂_t (see Sec. 2.1.2 and the example in Sec. 2.2.5). Now consider the case $r \neq 2m$. To compute the divergence of $\mathbf{n}$, it is convenient to use a basis $\{\mathbf{l}, \mathbf{n}, \mathbf{e}_\theta, \mathbf{e}_\phi\}$, where $\mathbf{l}$ and $\mathbf{n}$ are null and $g(\mathbf{l},\mathbf{n}) = 1$, while the spacelike basis vectors are $\mathbf{e}_\theta = r^{-1}\partial_\theta$, $\mathbf{e}_\phi = (r\sin\theta)^{-1}\partial_\phi$, in spherical coordinates. The vector $\mathbf{n}$ should be an "outward" radial geodesic while $\mathbf{l}$ points along the "inward" lightray (but is not necessarily an affine tangent vector to it). After we compute the vectors $\mathbf{n}$ and $\mathbf{l}$, it will follow that

$$\begin{aligned}\operatorname{div}\mathbf{n} &= \operatorname{Tr}_{(\mathbf{x},\mathbf{y})} g(\nabla_{\mathbf{x}}\mathbf{n}, \mathbf{y}) \\ &= g(\nabla_{\mathbf{n}}\mathbf{n}, \mathbf{l}) + g(\nabla_{\mathbf{l}}\mathbf{n}, \mathbf{n}) - g(\nabla_{\mathbf{e}_\theta}\mathbf{n}, \mathbf{e}_\theta) - g(\nabla_{\mathbf{e}_\phi}\mathbf{n}, \mathbf{e}_\phi) \\ &= g([\mathbf{n},\mathbf{e}_\theta], \mathbf{e}_\theta) + g([\mathbf{n},\mathbf{e}_\phi], \mathbf{e}_\phi).\end{aligned}$$

Now let us determine $\mathbf{n}$ and $\mathbf{l}$. Denote by

$$z(r) \equiv \left(1 - \frac{2m}{r}\right)^{1/2}$$

the redshift factor for the Schwarzschild spacetime (see Sec. 3.1.2). Assuming that

$$\mathbf{n} = \alpha_1\partial_t + \alpha_2\partial_r, \quad \mathbf{l} = \beta_1\partial_t + \beta_2\partial_r,$$

where $\alpha_{1,2}$ and $\beta_{1,2}$ are unknown scalar functions, we use the conditions

$$g(\mathbf{l},\mathbf{l}) = g(\mathbf{n},\mathbf{n}) = 0, \quad g(\mathbf{l},\mathbf{n}) = 1,$$

and find

$$z^2\alpha_1^2 = \frac{1}{z^2}\alpha_2^2, \quad , z^2\beta_1^2 = \frac{1}{z^2}\beta_2^2,$$
$$z^2\alpha_1\beta_1 - \frac{1}{z^2}\alpha_2\beta_2 = 1;$$

thus $\alpha_{1,2}$ and $\beta_{1,2}$ may be chosen as functions only of r. Further, we have

$$[\mathbf{l}, \mathbf{n}] = \beta_2 \left(\alpha_1' \partial_t + \alpha_2' \partial_r\right) - \alpha_2 \left(\beta_1' \partial_t + \beta_2' \partial_r\right),$$

where the prime $'$ denotes derivatives with respect to r. Using the condition

$$0 = g(\nabla_{\mathbf{n}}\mathbf{n}, \mathbf{l}) = -g(\mathbf{n}, \nabla_{\mathbf{n}}\mathbf{l}) = g(\mathbf{n}, [\mathbf{l}, \mathbf{n}]),$$

we find

$$z^2\alpha_1 \left(\alpha_1'\beta_2 - \alpha_2\beta_1'\right) - \frac{1}{z^2}\alpha_2 \left(\alpha_2'\beta_2 - \alpha_2\beta_2'\right) = 0.$$

Thus the four equations determine the unknown functions $\alpha_{1,2}, \beta_{1,2}$. A solution is (up to a constant)

$$\mathbf{n} = \frac{1}{z^2}\partial_t \pm \partial_r, \quad \mathbf{l} = \frac{1}{2}\left(\partial_t \mp z^2\partial_r\right),$$

where $\pm$ corresponds to the sign of z^2. Finally, we compute the commutators

$$[\mathbf{n}, \mathbf{e}_\theta] = \pm[\partial_r, \mathbf{e}_\theta] = \pm[\partial_r, \frac{1}{r}\partial_\theta] = \mp\frac{1}{r}\mathbf{e}_\theta,$$
$$[\mathbf{n}, \mathbf{e}_\phi] = \pm[\partial_r, \frac{1}{r\sin\theta}\partial_\phi] = \mp\frac{1}{r}\mathbf{e}_\phi,$$

and therefore

$$\mathrm{div}\mathbf{n} = \pm\frac{2}{r}.$$

This means that the divergence of an "outward"-pointing, future-directed null congruence is positive for $r > 2m$ and negative for $r < 2m$.

4.2 Hawking's area theorem

We have seen in Sec. 2.2.5 that the event horizon of a Schwarzschild black hole is a null surface. In more general situations, e.g. in the presence of several black holes that may move with respect to each other or even merge with each other, the event horizon is still a null surface. In this section we shall study the global properties of event horizons.

Consider an asymptotically flat spacetime that contains some points from which signals cannot escape to infinity.[8] Heuristically, the set $\mathcal{B}$ of all such points is interpreted as a "black hole-like" region. For a point $p \notin \mathcal{B}$, there exists a timelike curve connecting p to infinity; this curve may be perturbed while remaining timelike, thus there exists a neighborhood of p which is not in $\mathcal{B}$. Thus the set $\mathcal{B}$ is topologically closed. The **event horizon** $\mathcal{H}$ is defined as the boundary of $\mathcal{B}$; since $\mathcal{B}$ is closed, we have $\mathcal{H} \subset \mathcal{B}$. The horizon $\mathcal{H}$ is a 3-surface that separates $\mathcal{B}$ from the domain of spacetime from which timelike curves *can* escape to infinity (and hence, null curves can escape as well). Thus, the horizon $\mathcal{H}$ can also be thought of as the boundary of the past domain of dependence of the future null infinity $\mathscr{I}^+$. Let us study the properties of $\mathcal{H}$ in more detail.

The first important property of an event horizon is stated by a theorem due to R. Penrose.

Statement 1: An event horizon is a null surface whose generators are null geodesics without future endpoints, except if a generator hits a singularity. (Generators cannot leave $\mathcal{H}$ and enter the interior of $\mathcal{B}$.)

Proof: By construction, there cannot exist any future-directed timelike curves crossing $\mathcal{H}$ from its interior $\mathcal{B}$ outwards; also, for any point $p \notin \mathcal{B}$ there should exist at least one timelike curve not crossing $\mathcal{H}$. It follows from these requirements that the 3-surface $\mathcal{H}$ cannot be locally timelike (i.e. spanned by one timelike and two spacelike directions). Also, $\mathcal{H}$ cannot be spacelike, because a spacelike 3-surface separates a past subdomain from a future subdomain, so the future of $\mathcal{H}$ would have to be within $\mathcal{B}$, but then some points to the past of $\mathcal{H}$ will be unable to emit signals that do not enter $\mathcal{B}$. Therefore, $\mathcal{H}$ must be a null surface, except for points where $\mathcal{H}$ has a "corner" so the tangent 3-space to $\mathcal{H}$ is undefined. A null surface locally separates a past subdomain from a future subdomain, and it is clear that $\mathcal{B}$ must be on the future side from $\mathcal{H}$. We also know that null surface is generated by null geodesics (see Sec. 2.2.8). Suppose that a null generator γ had an endpoint p, and consider the causal structure of the tangent space at p. We are assuming that p is not a singularity, so that γ can be continued past p. It is clear that γ cannot exit $\mathcal{B}$; thus, since γ leaves $\mathcal{H}$, it must enter the interior of $\mathcal{B}$. There exists a point $p' \notin \mathcal{B}$ infinitesimally displaced to the past of γ, and there exists at least one null curve γ' starting from p' which escapes to infinity (or else we would have no timelike curves from p' escaping to infinity, contradicting the assumption $p' \notin \mathcal{B}$). Therefore, γ' can never cross $\mathcal{H}$. However, the curve γ' is arbitrarily close to γ, hence it enters $\mathcal{B}$. This contradiction proves that γ cannot have an endpoint.

The condition that singularities do not occur at the horizon is motivated as follows. We expect that the future history of an asymptotically flat spacetime is completely predictable from initial data on a Cauchy 3-surface $\mathcal{C}$ preceding the gravitational collapse. Thus, any null curve escaping to infinity must intersect $\mathcal{C}$ at one point. (This condition is called **asymptotic predictability** with respect to $\mathcal{C}$.) If a null generator γ of a horizon $\mathcal{H}$ stops at a singularity p, it means that the curve γ cannot be continued past p. Since there exist null curves $\gamma' \not\subset \mathcal{B}$ infinitely close to γ and reaching infinity, it follows that there exists a (null and/or spacelike) curve γ'' that "starts" at the singularity p and reaches infinity. To an observer at infinity, this curve appears to be a past-directed curve that stops at p. Thus, the curve γ'' does not intersect $\mathcal{C}$ and the asymptotic behavior at infinity in the direction of γ'' cannot be predicted from initial data on $\mathcal{C}$. Another way to express this is to say that the singularity is "directly visible" to an observer at infinity. There is a (so far, not proved but strongly supported) conjecture that any singularities that can occur in reasonable physical processes, such as gravitational collapse, must be "hidden" behind event horizons and are thus invisible to observers at infinity. This statement is called the **cosmic censorship** conjecture ("nature abhors naked singularities"). There are some artificial examples of spacetimes with "naked" singularities, but all such examples are patho-

[8] Note that the notion of "escaping to infinity" is well-defined if the spacetime is asymptotically flat (a "neighborhood of infinity" is the place where the spacetime becomes almost Minkowski, i.e. a neighborhood of $\mathscr{I}^+$ on a conformal diagram). Generalizations of this notion exist for non-asymptotically flat cases, such as the de Sitter spacetime.

logical in one or another way. Therefore, the assumption of asymptotic predictability appears reasonable.

Let us now examine the area of the event horizon. To be precise, we need to define what it means to consider an event horizon $\mathcal{H}$ "at a certain time." Let us select a global spacelike Cauchy 3-surface $\mathcal{C}$ in the entire spacetime. The surface $\mathcal{C}$ may be interpreted as the surface of constant "time." The intersection $\mathcal{B} \cap \mathcal{C}$ of the "black hole domain" $\mathcal{B}$ with $\mathcal{C}$ will be the "black hole interior at a given time." (The intersection $\mathcal{B} \cap \mathcal{C}$ may consist of several disconnected pieces, in which case we would say that there are several black holes at this time.) The boundary $\partial(\mathcal{B} \cap \mathcal{C})$ will be a spacelike 2-surface $\mathcal{H} \cap \mathcal{C}$; this is the "total event horizon at a given time." We have seen in Sec. 2.1.2 that the cross-section area A of a null 3-surface is defined independently of the observer's reference frame, and satisfies Eq. (2.6),

$$\nabla_{\mathbf{n}} A = (\mathrm{div}\mathbf{n})\, A,$$

if the null 3-surface has generators $\mathbf{n}$. In Sec. 2.2.5 we showed that the Schwarzschild horizon is a null surface whose generators are $\mathbf{n} = \partial_t$. Since the total area of the Schwarzschild horizon in the stationary reference frame is time-independent, $A = 4\pi r^2$, the field ∂_t has zero divergence. In a general, nonstationary spacetime containing black holes, the area of a horizon $\mathcal{H}$ may change with time (i.e. with the choice of the Cauchy surface $\mathcal{C}$), and thus the divergence of the null generators $\mathbf{n}$ of $\mathcal{H}$ may be nonzero. However, it turns out that the divergence of $\mathbf{n}$ cannot be negative, so the area cannot diminish with time. This is the essential content of the **area theorem** due to S. W. Hawking. Let us first derive the property $\mathrm{div}\mathbf{n} \geq 0$ and then prove the area theorem.

Statement 2: The divergence of null generators of an event horizon $\mathcal{H}$ is everywhere nonnegative, if the null energy condition is satisfied and the spacetime is asymptotically predictable.

Proof: We use the notation from the proof of Statement 1. If $\mathrm{div}\mathbf{n} < 0$ at a point p of $\mathcal{H}$ then $\mathrm{div}\mathbf{n} < 0$ also within a neighborhood of p. By the focusing theorem, the null generators of $\mathcal{H}$ emitted from the neighborhood of p will focus within a finite interval of the affine parameter. After focusing, some point q on a null geodesic γ emitted from p will be reachable from p by a timelike curve. A timelike curve starting at p must be completely within $\mathcal{B}$ (or else this curve will escape to infinity, but $p \in \mathcal{B}$). Thus, the point q will be inside $\mathcal{B}$ and not on the horizon surface $\mathcal{H}$ any more. Thus, a null geodesic γ emitted from p will have to enter $\mathcal{B}$ within a finite interval of the affine parameter; the part of γ within $\mathcal{H}$ will then have an endpoint where γ enters the interior of $\mathcal{B}$. This contradicts Statement 1.

Statement 3: The total area of all the event horizons cannot diminish with time, under the conditions of Statements 1 and 2. In other words: For a Cauchy surface $\mathcal{C}_1$ and another Cauchy surface $\mathcal{C}_2$ to the future of $\mathcal{C}_1$, the area of the intersection does not decrease, $A[\mathcal{H} \cap \mathcal{C}_2] \geq A[\mathcal{H} \cap \mathcal{C}_1]$.

Proof: Consider first the intersection $\mathcal{H} \cap \mathcal{C}_1$. A portion of $\mathcal{H}$ between $\mathcal{C}_1$ and $\mathcal{C}_2$ will be generated by null geodesics emitted from $\mathcal{H} \cap \mathcal{C}_1$ to the future. Since each generator has no endpoints on $\mathcal{H}$, it must intersect $\mathcal{C}_2$ by the Cauchy surface property of $\mathcal{C}_2$. The intersection of the generators with $\mathcal{C}_2$ will have an area not smaller than $A[\mathcal{H} \cap \mathcal{C}_1]$, since these geodesics have everywhere nonnegative divergence. There is also a possibility that other disconnected pieces of $\mathcal{H}$ will have formed between $\mathcal{C}_1$ and $\mathcal{C}_2$; this would only increase the total area $A[\mathcal{H} \cap \mathcal{C}_2]$. Therefore, $A[\mathcal{H} \cap \mathcal{C}_2] \geq A[\mathcal{H} \cap \mathcal{C}_1]$.

The area theorem has had a wide-reaching impact on all of fundamental physics. It has lead to the hypothesis of black hole entropy,

$$S = \frac{1}{4} A = 4\pi m^2,$$

and to the holographic principle. Let us consider a simple application of the area theorem. In an asymptotically flat spacetime containing two distant black holes with masses m_1 and m_2, there may be some process that lets the two black holes come closer together and eventually merge. Initially, the total area of the event horizon of this spacetime is $16\pi(m_1^2 + m_2^2)$. After merging, the spacetime contains a larger black hole of mass M, with the area of the event horizon $16\pi M^2$. The area theorem says that $16\pi M^2 \geq 16\pi(m_1^2 + m_2^2)$ regardless of the details of the merger, and regardless of whether any extra matter or radiation was absorbed or emitted in the process of merging. This inequality sets a fundamental bound on the amount of energy that can be extracted from a black hole merger.

4.3 Holographic principle

See the review [5].

In the lecture, I covered: general ideas behind the Generalized Second Law of thermodynamics, Bekenstein's black hole entropy, Susskind's collapse argument for spacelike bound, problems with spacelike bounds (closed universes, expanding universes, almost-null spacelike surfaces), Bousso's covariant entropy bound formulated using lightsheets, and Bousso's projection theorem.

A rough map of ideas: Black holes must contain entropy for the 2nd law to hold. Hawking's area theorem supports the identification of entropy and the horizon area. Assuming that black hole is the highest-entropy state, we obtain a bound on the entropy of matter within a (spacelike) 2-sphere. Since entropy counts the internal states, there is "enough space" on the boundary of a region to encode all the possible states within the region; hence the term "holography." Bousso's covariant entropy bound is a generalization of this bound that remains valid in more complicated situations, and yields the spacelike bound as a particular case if appropriate conditions hold. Presently, the holographic principle has the status of a strongly supported conjecture.

5 Variational principle

Additional literature: [34].

Under **variational principles** we understand considerations based on the variation of functionals. The commonly used variational principle states that the equation of motion for the fields are derived by extremizing a certain functional of the fields called the action or the Lagrangian. The Hamiltonian formulation of field dynamics is then derived from the Lagrangian formulation. In this chapter we shall examine Lagrangian and Hamiltonian formulations of general relativity and explore some related issues.

5.1 Lagrangian formulation

5.1.1 Classical field theory

Classical field theory describes **fields**, i.e. functions on spacetime $\mathcal{M}$ or sections of some vector bundle with the spacetime $\mathcal{M}$ as the base. The fields under consideration may be scalar fields ϕ_1, ϕ_2, ...; vector fields, spinor fields, and so on. Let us denote all the fields collectively by ϕ_j, where j is an index enumerating each individual field and each of their components. Field theory is based on the variational principle (or **Hamilton's principle of stationary of action**, or **action principle** for short): The field equations are the conditions for having a local extremum of the action functional,

$$S\left[\phi\right] = \int_{\mathcal{M}} \mathrm{d}^4x\, L\left(\phi_i, \nabla_\mu \phi_i, ...\right) \tag{5.1}$$

where the Lagrangian density L depends on the field ϕ_j and its (covariant) derivatives.

The action principle states that a physically realized configuration $\phi_j(x)$ of a field ϕ must be an extremum of the action functional. The variation of the action under a small change $\delta\phi^j(x)$ of fields $\phi^j(x)$ is

$$\delta S[\phi] = \int_{\mathcal{M}} \mathrm{d}^4x \sum_j \frac{\delta S}{\delta \phi^j(x)} \delta\phi^j(x) + O\left(\left[\delta\phi^j\right]^2\right).$$

This yields the **Euler-Lagrange equation** for the field,

$$\frac{\delta S}{\delta \phi^j(x)} = 0.$$

This method of deriving field equations is called a **Lagrangian formulation** of a field theory.

The currently established field theories (electrodynamics, gravitation, weak and strong interactions) are described by Lagrangian densities which depend only on the fields and their first derivatives, $L = L\left(\phi^j, \phi^j_{,\mu}\right)$. For such Lagrangians, the first-order variation of the action is given by the formula

$$\delta S = \int_{\mathcal{M}} \mathrm{d}^4x \sum_j \left(\frac{\partial L\left(\phi^i, \nabla\phi^i\right)}{\partial \phi^j} - \frac{\partial}{\partial x^\mu} \frac{\partial L\left(\phi^i, \nabla\phi^i\right)}{\partial \phi^j_{,\mu}} \right) \delta\phi^j(x),$$

where the summation over μ is implied. The boundary terms vanish if

$$\sum_j \frac{\partial L}{\partial \phi^j_{,\mu}} \delta\phi^j \to 0 \text{ sufficiently rapidly as } |\mathbf{x}| \to \infty,\ |t| \to \infty,$$

which is the usual assumption. Thus we obtain the following Euler-Lagrange equations for the fields ϕ^j,

$$\frac{\delta S\left[\phi\right]}{\delta \phi^j(x)} \equiv \frac{\partial L\left(\phi^i, \nabla\phi^i\right)}{\partial \phi^j} - \frac{\partial}{\partial x^\mu} \frac{\partial L\left(\phi^i, \nabla\phi^i\right)}{\partial \phi^j_{,\mu}} = 0. \tag{5.2}$$

The formula (5.2) holds for all Lagrangians that depend on fields and their first derivatives. If a Lagrangian for a field ϕ contains second-order derivatives such as $\phi_{;\mu\nu}$, the corresponding Euler-Lagrange equations will generally contain derivatives of third and fourth order.

If a classical field theory is to be compatible with general relativity, the Lagrangian must be generally covariant. In practice, this means that the Lagrangian should be an invariant combination of covariant derivatives of fields and the metric $g_{\mu\nu}$. The integration must be performed with the volume element $d^4x\sqrt{-g}$, where $g \equiv \det\left(g_{\mu\nu}\right)$ is the determinant of the covariant metric tensor. Normally, this factor is included into the Lagrangian L.

The simplest example of a field is a scalar field. The Lagrangian density for a real-valued scalar field $\phi(x)$ is

$$L\left(\phi, \partial_\mu\phi\right) = \left[\frac{1}{2} g^{\mu\nu}\phi_{,\mu}\phi_{,\nu} - V(\phi)\right]\sqrt{-g}, \tag{5.3}$$

where $V(\phi)$ is a potential that describes self-interaction of the field.

Covariant volume element

The expression d^4x does not give the correct volume element if the coordinates x are not Cartesian or if the spacetime is curved. It is not difficult to show that the volume element in any number of dimensions is given by the formula $\mathrm{d}V = \mathrm{d}^n x\sqrt{|g(x)|}$. If $\mathbf{u}_1, ..., \mathbf{u}_n$ are some vectors in a Euclidean space, let $G_{ij} \equiv \mathbf{u}_i \cdot \mathbf{u}_j$ be the $n \times n$ matrix of their pairwise scalar products. Then the volume of the n-dimensional parallelepiped spanned by the vectors $\mathbf{u}_i$ is $V = \sqrt{|\det G|}$. We can prove this standard statement by considering the matrix U^j_i of components of the vectors $\mathbf{u}_i$ in an orthonormal basis $\{\mathbf{e}_j\}$, i.e.

$$\mathbf{u}_i = \sum_j U^j_i \mathbf{e}_j.$$

The matrix U can be also understood as a linear transformation that maps the unit parallelepiped spanned by $\{\mathbf{e}_j\}$ to a parallelepiped spanned by $\{\mathbf{u}_i\}$. A standard definition of the determinant of a linear transformation is the volume of the image of a unit parallelepiped after the transformation. Therefore, the volume V is $V = \det U$. Then we observe that the matrix G_{ij} satisfies $G = U^T U$, therefore $\det G = (\det U)^2 = V^2$ and $V = \sqrt{|\det G|}$.

In general relativity, the spacetime has a metric with signature $(+,-,-,-)$ and the determinant $\det g_{\mu\nu}$ is always negative (except at singular points where it may be zero or infinite). Therefore we change the sign of g and write the volume element as $d^4x\sqrt{-g}$.

Minimal coupling to gravity

The action for a scalar field,

$$S[\phi] = \int_{\mathcal{M}} \mathrm{d}^4x\sqrt{-g}\left[\frac{1}{2}g^{\mu\nu}\phi_{,\mu}\phi_{,\nu} - V(\phi)\right], \tag{5.4}$$

explicitly depends on $g_{\mu\nu}$ and thus describes a field coupled to gravity. This form of coupling is called the **minimal coupling** to gravity; this is the minimal required interaction of a field with gravitation which necessarily follows from the requirement of compatibility with general relativity.

The Euler-Lagrange equation for a minimally coupled field ϕ follows from the action (5.4),

$$\partial_\alpha \frac{\partial L}{\partial \phi_{,\alpha}} - \frac{\partial L}{\partial \phi} = \left(\sqrt{-g}g^{\alpha\beta}\phi_{,\beta}\right)_{,\alpha} + \frac{\partial V}{\partial \phi} = 0.$$

This equation can be rewritten in a manifestly covariant form as

$$\phi^{;\alpha}_{;\alpha} + \frac{\partial V}{\partial \phi} = 0.$$

This is similar to the Klein-Gordon equation, $\phi^{,\alpha}_{,\alpha} + m^2\phi = 0$, except for the presence of covariant derivatives.

A **free** (i.e. noninteracting) **field** has the potential

$$V(\phi) = \frac{1}{2}m^2\phi^2.$$

This is the simplest nontrivial potential; an additional linear term $A\phi$ can be removed by a field redefinition $\phi(x) = \tilde{\phi}(x) + \phi_0$. The parameter m is the rest mass of the particles described by the field ϕ in quantum theory. The equation of motion for a free field ϕ is *linear* and thus describes "waves" that can cross without distorting each other. In other words, the field ϕ has **no self-interaction**. A field would have self-interaction if the potential $V(\phi)$ were such that $V''' \neq 0$, so that the equation of motion would be nonlinear.

Gauss's law with covariant derivatives

When computing the variation of a generally covariant action such as the action (5.4), one needs to integrate by parts. A useful shortcut in such calculations is an analog of Gauss's law with covariant derivatives.

The covariant divergence of a vector field A^μ can be written as

$$A^\mu_{;\mu} = \frac{1}{\sqrt{-g}}\partial_\mu\left(\sqrt{-g}A^\mu\right).$$

The formula analogous to the Gauss law is

$$\int_V A^\mu{}_{;\mu}\sqrt{-g}\mathrm{d}^nx = \oint_{\partial V}\sqrt{-h}A^\mu \mathrm{d}^{n-1}S_\mu,$$

where h is the partial metric on the hypersurface ∂V and $d^{n-1}S_\mu$ is the oriented element of the $(n-1)$-volume on the hypersurface. Using this formula, we may reduce volume integrals of total divergences to boundary integrals. Assuming that the contribution of the boundary terms vanishes, we obtain

$$\int \mathrm{d}^nx\sqrt{-g}A^\mu_{;\mu}B = -\int \mathrm{d}^nx\sqrt{-g}A^\mu B_{,\mu}. \tag{5.5}$$

This formula can be used to integrate by parts: We set $A^\mu \equiv \phi_{,\alpha}g^{\alpha\mu}$, $B \equiv \psi$, and find

$$\int \mathrm{d}^4x\sqrt{-g}\phi_{,\alpha}\psi_{,\beta}g^{\alpha\beta} = -\int \mathrm{d}^4x\sqrt{-g}\left(\phi_{,\alpha}g^{\alpha\beta}\right)_{;\beta}\psi.$$

Note that the covariant derivative of the metric is zero, $g^{;\alpha}_{\alpha\beta} = 0$, so we may lower or raise the indices under covariant derivatives at will; for example, $A^\mu_{;\mu} = A^{;\mu}_{\mu}$.

5.1.2 Einstein-Hilbert action

Given a metric g on a manifold $\mathcal{M}$, one can compute the Riemann tensor $R_{\kappa\lambda\mu\nu}$ and the Ricci tensor $R_{\mu\nu}$. Einstein postulated that the metric in a spacetime must be such that the Ricci tensor satisfies the **Einstein equation**

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = -8\pi G T_{\mu\nu},$$

where $R \equiv R_{\mu\nu}g^{\mu\nu}$ and $T_{\mu\nu}$ is the energy-momentum tensor of matter. This equation can be derived from a variational principle by extremizing an appropriately chosen action functional.

We shall begin with the **vacuum** Einstein equation (i.e. the Einstein equation in the absence of matter, $T_{\mu\nu} = 0$),

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = 0. \tag{5.6}$$

This equation can be derived from the **Einstein-Hilbert action**,

$$S_{EH}[g] \equiv \frac{1}{16\pi G}\int_{\mathcal{M}} R\sqrt{-g}\mathrm{d}^4x, \frac{1}{16\pi G} \tag{5.7}$$

where G is Newton's gravitational constant and $\sqrt{-g} \equiv \sqrt{-\det g_{\mu\nu}}$, so that $\sqrt{-g}d^4x$ is the covariant volume element. The integration in Eq. (5.7) is performed over the entire spacetime $\mathcal{M}$. Such an integral might diverge when the manifold $\mathcal{M}$ is noncompact and if the curvature R does not vanish at infinity, in which case one should limit the integration to a very large but finite subdomain $\tilde{\mathcal{M}} \subset \mathcal{M}$.

In the framework of the variational principle, one considers a small change in the metric,

$$g_{\mu\nu} \to g_{\mu\nu} + \lambda h_{\mu\nu},$$

where λ is a "small" number and $h_{\mu\nu}$ is an arbitrary symmetric tensor field, which is assumed to vanish outside of the subdomain $\tilde{\mathcal{M}}$. The first-order variation of the EH action $S_{EH}[g]$ is a linear functional of $h_{\mu\nu}(p)$ that can be defined as

$$\int_{\mathcal{M}} \mathrm{d}^np\, h_{\mu\nu}(p)\frac{\delta S_{EH}[g]}{\delta g_{\mu\nu}(p)} \equiv \left.\frac{d}{d\lambda}\right|_{\lambda=0} S_{EH}[g+\lambda h].$$

Note that the integration above is performed over all points $p \in \mathcal{M}$ and involves the coordinate-based volume element d^np rather than the invariant volume element $\sqrt{-g}\mathrm{d}^nx$. The point p plays the role of a label, much like the indices μ, ν in $h_{\mu\nu}(p)$, and thus a simple "summation" over all the points p is needed.

The expression

$$\left.\frac{d}{d\lambda}\right|_{\lambda=0} S_{EH}[g+\lambda h]$$

can be interpreted as the "directional derivative" of the action functional $S_{EH}[g]$ in the "direction" of $h_{\mu\nu}(p)$. (I put the word "direction" in quotes because $h_{\mu\nu}(p)$ is a tensor-valued function on the manifold that describes an arbitrary modification of the metric $g_{\mu\nu}(p)$ at every point.) The action S_{EH} is extremized if its derivative vanishes in every "direction," i.e. if $\delta S_{EH}/\delta g_{\mu\nu}(p) = 0$. This yields an equation for g. This method of deriving the Einstein equation is the Lagrangian formulation of general relativity.

Let us adopt the following, somewhat more abstract geometric picture of the variational principle. The set of all the possible metrics $g_{\mu\nu}(p)$ can be considered as an (infinite-dimensional) manifold $\mathcal{G}$, every individual metric $g_{\mu\nu}$ being a single "point" of $\mathcal{G}$. The Einstein-Hilbert functional $S_{EH}[g]$ is then a real-valued scalar "function" on this manifold, $S_{EH} : \mathcal{G} \to \mathbb{R}$. As usual, the tangent space $T_g\mathcal{G}$ to the manifold $\mathcal{G}$ is the set of all directional derivatives at a "point" $g_{\mu\nu}$. Thus, a tangent vector

$$\mathbf{h} \equiv \int_{\mathcal{M}} \mathrm{d}^n p\, h_{\mu\nu}(p) \frac{\delta}{\delta g_{\mu\nu}(p)}$$

defines the derivative in a particular "direction" $h_{\mu\nu}$. The tangent vector $\mathbf{h}$ acts on the "function" $S_{EH}[g]$, yielding

$$\mathbf{h} \circ S_{EH} \equiv \int_{\mathcal{M}} \mathrm{d}^n p\, h_{\mu\nu}(p) \frac{\delta S_{EH}[g]}{\delta g_{\mu\nu}(p)},$$

which is the first-order variation of S_{EH} in the direction $h_{\mu\nu}$. This can be also written as

$$\int_{\mathcal{M}} d^n p\, h_{\mu\nu}(p) \frac{\delta S_{EH}[g]}{\delta g_{\mu\nu}(p)} \equiv (dS_{EH}[g]) \circ \mathbf{h},$$

where $dS_{EH}[g]$ is the naturally defined 1-form on $\mathcal{G}$, the "gradient" of S_{EH} in the space of metrics. The "components" of the 1-form $dS_{EH}[g]$ are $\delta S_{EH}/\delta g_{\mu\nu}(p)$, where we interpret p as an "index" on par with μ, ν. The action S_{EH} has a (local) extremum at a particular "point" $g_{\mu\nu}$ if the 1-form $dS_{EH}[g]$ is equal to zero at $g_{\mu\nu}$.

In the rest of this section, we shall show, by a direct calculation, that the condition $\delta S_{EH}/\delta g_{\mu\nu}(p) = 0$ is equivalent to the vacuum Einstein equation.

It is convenient to compute the first-order variation of $S_{EH}[g]$ with respect to the *inverse* metric $g^{\mu\nu}$ rather than with respect to $g_{\mu\nu}$. We need to find the variation of $\sqrt{-g}$ and $R \equiv g^{\mu\nu}R_{\mu\nu}$, where $R_{\mu\nu}$ is the Ricci tensor, under an infinitesimal change $g^{\mu\nu} \to g^{\mu\nu} + \delta g^{\mu\nu}$. The calculation of $\delta\sqrt{-g}$ is easy if we use the standard formula for the derivative of the determinant of a matrix (note that $\sqrt{-g} = \sqrt{-\det(g_{\mu\nu})}$, while $\det(g^{\mu\nu}) = 1/\det(g_{\mu\nu}) = g^{-1}$):

$$\frac{\partial\sqrt{-g}}{\partial g^{\mu\nu}} = -\frac{1}{2} g_{\mu\nu}\sqrt{-g}.$$

Calculation: Derive the formula

$$\frac{\partial}{\partial A_{jk}} \det A = (\det A)\left(A^{-1}\right)_{jk},$$

where $A \equiv A_{jk}$ is a finite-dimensional square matrix and A^{-1} is the inverse matrix.

Hint: Show that

$$\det\left(\hat{1} + \lambda B\right) = 1 + \lambda \mathrm{Tr}\, B + O(\lambda^2),$$

where $\hat{1}$ is the identity matrix, λ is "small enough," and B is an arbitrary matrix.

The variation of $S_{EH}[g]$ can be split into three terms,

$$\begin{aligned}\delta S_{EH}[g] &= \delta \int_{\mathcal{M}} \mathrm{d}^4 p\, R_{\mu\nu} g^{\mu\nu}\sqrt{-g} \\ &= \int_{\mathcal{M}} \mathrm{d}^4 p \sqrt{-g}\left[g^{\mu\nu}\delta R_{\mu\nu} + R_{\mu\nu}\delta g^{\mu\nu} - \frac{1}{2} R g_{\mu\nu}\delta g^{\mu\nu}\right].\end{aligned}$$

Shortly, we shall show that the first term in parentheses above,

$$\int_{\mathcal{M}} \mathrm{d}^4 p \sqrt{-g} g^{\mu\nu}\delta R_{\mu\nu}, \tag{5.8}$$

vanishes (under suitable boundary conditions) because it is a total divergence. Thus, we shall have

$$\delta S_{EH}[g] = \int_{\mathcal{M}} \mathrm{d}^4 p \sqrt{-g}\left(R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu}\right)\delta g^{\mu\nu},$$

which is equivalent to

$$\frac{\delta S_{EH}[g]}{\delta g^{\mu\nu}(p)} = R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu}. \tag{5.9}$$

So the condition $\delta S_{EH}/\delta g^{\mu\nu} = 0$ is equivalent to the vacuum Einstein equation (5.6).

It remains to analyze the term (5.8). We shall now use the index-free notation for clarity. The term of interest is

$$g^{\mu\nu}\delta R_{\mu\nu} = \mathrm{Tr}_{(\mathbf{x},\mathbf{y})}\delta\mathrm{Ric}(\mathbf{x},\mathbf{y}). \tag{5.10}$$

The Ricci tensor is a contraction of the Riemann tensor,

$$\mathrm{Ric}(\mathbf{x},\mathbf{y}) = \mathrm{Tr}_{(\mathbf{a},\mathbf{b})} R(\mathbf{x},\mathbf{a},\mathbf{y},\mathbf{b}).$$

To determine the variation of the Riemann tensor, we need to consider the change in the Levi-Civita connection, $\nabla \to \tilde{\nabla}$, under $g \to \tilde{g} = g + \delta g$. It is easy to show that the difference $\tilde{\nabla} - \nabla$ between any two connections does not contain derivatives,

$$(\tilde{\nabla} - \nabla)(f\mathbf{x}) = f \cdot (\tilde{\nabla} - \nabla)\mathbf{x},$$

and hence acts as a transformation-valued 1-form $\delta\Gamma$,

$$\tilde{\nabla}_{\mathbf{u}}\mathbf{x} - \nabla_{\mathbf{u}}\mathbf{x} \equiv \delta\Gamma(\mathbf{u})\mathbf{x}.$$

An explicit formula for $\delta\Gamma$ can be derived (see the Calculation below) but is not required for the present derivation.

Now we use the definition (1.64) to express the variation of the Riemann tensor through $\delta\Gamma$,

$$\begin{aligned}\delta R(\mathbf{u},\mathbf{v})\mathbf{w} &= [\delta\Gamma(\mathbf{u}), \nabla_{\mathbf{v}}]\,\mathbf{w} + [\nabla_{\mathbf{u}}, \delta\Gamma(\mathbf{v})]\,\mathbf{w} - \delta\Gamma([\mathbf{u},\mathbf{v}])\mathbf{w} \\ &= (\nabla_{\mathbf{u}}\delta\Gamma)(\mathbf{v})\mathbf{w} - (\nabla_{\mathbf{v}}\delta\Gamma)(\mathbf{u})\mathbf{w}.\end{aligned} \tag{5.11}$$

Calculation: Verify Eq. (5.11). **Details: omitted.**

The variation of the Ricci tensor is found[1] as a trace of the variation of the Riemann tensor,

$$\begin{aligned}\delta\mathrm{Ric}(\mathbf{x},\mathbf{y}) &= \mathrm{Tr}_{(\mathbf{a},\mathbf{b})} g(\delta R(\mathbf{x},\mathbf{a})\mathbf{y},\mathbf{b}) \\ &= \mathrm{Tr}_{(\mathbf{a},\mathbf{b})} g\left(\mathbf{b}, (\nabla_{\mathbf{x}}\delta\Gamma)(\mathbf{a})\mathbf{y} - (\nabla_{\mathbf{a}}\delta\Gamma)(\mathbf{x})\mathbf{y}\right).\end{aligned}$$

[1] From this point and until the end of the section, index-free computations involve traces and are more cumbersome than computations in the index notation. However, I would like to show how one could manage a complete calculation without indices. Below, an equivalent computation using indices is shown as well.

We cannot simplify the last expression any further, until we consider the trace (5.10),

$$\mathrm{Tr}_{(\mathbf{x},\mathbf{y})(\mathbf{a},\mathbf{b})} g\left(\mathbf{b}, (\nabla_{\mathbf{x}} \delta\Gamma)(\mathbf{a})\mathbf{y}\right) - \mathrm{Tr}_{(\mathbf{x},\mathbf{y})(\mathbf{a},\mathbf{b})} g\left(\mathbf{b}, (\nabla_{\mathbf{a}} \delta\Gamma)(\mathbf{x})\mathbf{y}\right). \tag{5.12}$$

The tensor $\delta\Gamma$ has rank $(1,2)$ and can be contracted in two different ways, yielding two vector fields $\mathbf{q}_{1,2}$ defined by

$$\begin{aligned} \mathbf{q}_1 &\equiv \mathrm{Tr}_{(\mathbf{x},\mathbf{y})} \delta\Gamma(\mathbf{x})\mathbf{y}, \\ g(\mathbf{q}_2, \mathbf{y}) &\equiv \mathrm{Tr}_{(\mathbf{a},\mathbf{b})} g(\mathbf{b}, \delta\Gamma(\mathbf{a})\mathbf{y}), \text{ for all } \mathbf{y}. \end{aligned}$$

The vector fields $\mathbf{q}_{1,2}$ involve *derivatives* of the metric perturbation δg. In the index notation, we have

$$(\delta\Gamma(\mathbf{x})\,\mathbf{y})^{\mu} \equiv \delta\Gamma^{\mu}_{\alpha\beta} x^{\alpha} y^{\beta}; \quad q_1^{\mu} \equiv \delta\Gamma^{\mu}_{\alpha\beta} g^{\alpha\beta}, \quad q_2^{\mu} \equiv g^{\mu\nu}\delta\Gamma^{\alpha}_{\alpha\nu}.$$

Now we use the properties of the trace (see Sec. 1.7.3) and the fact that ∇ can be interchanged with g and with the "mute vectors" $\mathbf{a}, \mathbf{b}, \mathbf{x}, \mathbf{y}$ appearing under the trace operation. The first term in Eq. (5.12) is expressed as

$$\mathrm{Tr}_{(\mathbf{x},\mathbf{y})} \nabla_{\mathbf{x}} \mathrm{Tr}_{(\mathbf{a},\mathbf{b})} g(\mathbf{b}, \delta\Gamma(\mathbf{a})\mathbf{y}) = \mathrm{Tr}_{(\mathbf{x},\mathbf{y})} g(\nabla_{\mathbf{x}}\mathbf{q}_2, \mathbf{y}) \equiv \mathrm{div}\ \mathbf{q}_2,$$

and the second term as

$$\begin{aligned} \mathrm{Tr}_{(\mathbf{x},\mathbf{y})(\mathbf{a},\mathbf{b})} g\left(\mathbf{b}, (\nabla_{\mathbf{a}}\delta\Gamma)(\mathbf{x})\mathbf{y}\right) &= \mathrm{Tr}_{(\mathbf{a},\mathbf{b})} g\left(\mathbf{b}, \nabla_{\mathbf{a}} \mathrm{Tr}_{(\mathbf{x},\mathbf{y})} \delta\Gamma(\mathbf{x})\mathbf{y}\right) \\ &= \mathrm{Tr}_{(\mathbf{a},\mathbf{b})} g\left(\mathbf{b}, \nabla_{\mathbf{a}}\mathbf{q}_1\right) \equiv \mathrm{div}\ \mathbf{q}_1. \end{aligned}$$

Hence,

$$g^{\mu\nu}\delta R_{\mu\nu} = \mathrm{div}\,(\mathbf{q}_2 - \mathbf{q}_1) \equiv \mathrm{div}\ \mathbf{q}, \tag{5.13}$$

where $\mathbf{q} \equiv \mathbf{q}_2 - \mathbf{q}_1$ is an auxiliary vector field defined through $\delta g^{\mu\nu}$. Therefore, the term (5.8) is an integral of a total divergence and thus vanishes,

$$\int_{\tilde{\mathcal{M}}} \mathrm{d}^4 p \sqrt{-g} g^{\mu\nu}\delta R_{\mu\nu} = \int_{\tilde{\mathcal{M}}} \mathrm{d}^4 p \sqrt{-g}\,\mathrm{div}\,(\mathbf{q}) = 0,$$

as long as $\mathbf{q}$ vanishes at the boundary $\partial\tilde{\mathcal{M}}$ of the domain of integration $\tilde{\mathcal{M}}$. (Since $\mathbf{q}$ depends on the derivatives of $\delta g^{\mu\nu}$, this condition is satisfied if $\delta g^{\mu\nu}$ identically vanishes outside the domain $\tilde{\mathcal{M}}$, or alternatively if both $\delta g^{\mu\nu}$ and $\delta g^{\mu\nu}_{;\alpha}$ vanish at the boundary of $\tilde{\mathcal{M}}$. The Einstein-Hilbert action $S_{EH}[g]$ depends on second derivatives of the metric. It is a fortunate fact that the equations of motion (the Einstein equations) are only second-order in g; but, in any case, it is no surprise that the variation $\delta S_{EH}[g]$ contains boundary terms linear in $\nabla\delta g$.) This result completes the derivation of the vacuum Einstein equation from the Einstein-Hilbert action.

Remark: The index-free computations above are cumbersome, mainly because different traces of a high-rank tensor $\delta\Gamma$ are required. The index-free notation is best for low-rank tensor computations that do not involve complicated traces. After arriving at Eq. (5.11), it is definitely easier to proceed using the index notation.

Calculation: Using the index notation, reproduce the derivation of Eq. (5.13), starting with Eq. (5.11).

Solution: The index notation for the Riemann tensor is

$$a^{\alpha}b^{b}x^{\lambda}y^{\mu}R_{\alpha\beta\lambda}{}^{\mu} \equiv R(\mathbf{a},\mathbf{b},\mathbf{x},\mathbf{y}) \equiv g(R(\mathbf{a},\mathbf{b})\mathbf{x},\mathbf{y}),$$

hence Eq. (5.11) gives

$$\begin{aligned} \delta R_{\alpha\beta\lambda}{}^{\mu} &= \delta\Gamma^{\mu}_{\beta\lambda;\alpha} - \delta\Gamma^{\mu}_{\alpha\lambda;\beta}, \\ \delta R_{\alpha\lambda} &= \delta R_{\alpha\beta\lambda}{}^{\beta} = \delta\Gamma^{\beta}_{\beta\lambda;\alpha} - \delta\Gamma^{\beta}_{\alpha\lambda;\beta}, \\ g^{\alpha\lambda}\delta R_{\alpha\lambda} &= g^{\alpha\lambda}\left(\delta\Gamma^{\beta}_{\beta\lambda;\alpha} - \delta\Gamma^{\beta}_{\alpha\lambda;\beta}\right) \\ &= \left(g^{\alpha\lambda}\delta\Gamma^{\beta}_{\beta\lambda} - g^{\beta\lambda}\delta\Gamma^{\alpha}_{\beta\lambda}\right)_{;\alpha} \equiv q^{\alpha}{}_{;\alpha}. \end{aligned} \tag{5.14}$$

The last term is a total divergence, as required.

Calculation: Consider the first-order variation of the Levi-Civita connection,

$$\Gamma(\mathbf{x})\mathbf{y} \equiv \lim_{\lambda\to 0} \frac{1}{\lambda}\left(\tilde{\nabla}_{\mathbf{x}}\mathbf{y} - \nabla_{\mathbf{x}}\mathbf{y}\right),$$

where ∇ and $\tilde{\nabla}$ are Levi-Civita connections corresponding to the metrics g and $g + \lambda h$, and h is an arbitrary symmetric tensor of rank $(0,2)$. An explicit formula for the tensor Γ in terms of h can be derived,

$$2g(\Gamma(\mathbf{x})\mathbf{y},\mathbf{z}) = (\nabla_{\mathbf{x}}h)\circ(\mathbf{y},\mathbf{z}) + (\nabla_{\mathbf{y}}h)\circ(\mathbf{x},\mathbf{z}) - (\nabla_{\mathbf{z}}h)\circ(\mathbf{x},\mathbf{y}).$$

In the index notation,

$$\Gamma^{\lambda}_{\alpha\beta} = \frac{1}{2}g^{\lambda\mu}\left(h_{\alpha\mu;\beta} + h_{\beta\mu;\alpha} - h_{\alpha\beta;\mu}\right). \tag{5.15}$$

Derivation: We use Eq. (1.45) and choose vectors $\mathbf{x}, \mathbf{y}, \mathbf{z}$ with *all* vanishing derivatives, i.e. vector fields $\mathbf{x}, \mathbf{y}, \mathbf{z}$ such that

$$\nabla_{(\mathbf{x}\text{ or }\mathbf{y}\text{ or }\mathbf{z})}(\mathbf{x}\text{ or }\mathbf{y}\text{ or }\mathbf{z}) = 0.$$

This choice is certainly possible at one point p and does not influence the result since $\Gamma(\mathbf{x})\mathbf{y}$ involves no derivatives of $\mathbf{x}$ or $\mathbf{y}$ but only values of $\mathbf{x}$ and $\mathbf{y}$ at p. Note that with this choice, we still have $\tilde{\nabla}_{\mathbf{x}}\mathbf{y} \neq 0$ since $\tilde{\nabla}$ is a different connection from ∇, however

$$\tilde{\nabla}_{\mathbf{x}}\mathbf{y} - \nabla_{\mathbf{x}}\mathbf{y} = O(\lambda).$$

Now Eq. (1.45) yields (up to terms of order λ^2)

$$\begin{aligned} &2\,(g + \lambda h)\,(\tilde{\nabla}_{\mathbf{x}}\mathbf{y}, \mathbf{z}) - 2g(\nabla_{\mathbf{x}}\mathbf{y}, \mathbf{z}) \\ &= 2\lambda h\left(\tilde{\nabla}_{\mathbf{x}}\mathbf{y}, \mathbf{z}\right) \\ &\quad + \lambda\big(\mathbf{x}\circ h(\mathbf{y},\mathbf{z}) + \mathbf{y}\circ h(\mathbf{x},\mathbf{z}) - \mathbf{z}\circ h(\mathbf{x},\mathbf{y})\big) \\ &= \lambda\big((\nabla_{\mathbf{x}}h)\circ(\mathbf{y},\mathbf{z}) + (\nabla_{\mathbf{y}}h)\circ(\mathbf{x},\mathbf{z}) - (\nabla_{\mathbf{z}}h)\circ(\mathbf{x},\mathbf{y})\big) \end{aligned}$$

because the term

$$2\lambda h\left(\tilde{\nabla}_{\mathbf{x}}\mathbf{y}, \mathbf{z}\right) = O(\lambda^2)$$

can be disregarded.

5.1.3 Nonlinear $f(R)$ gravity

In the Lagrangian formulation of field theory, every field ϕ is described by an action functional,

$$\int_{\mathcal{M}} \sqrt{-g}\mathrm{d}^n p\, L[\phi],$$

where L is an appropriate Lagrangian. Thus, Einstein's theory of gravity can be viewed as a field theory where the "field" is the point-dependent metric $g_{\mu\nu}(p)$ and the Lagrangian is

the Einstein-Hilbert one, Eq. (5.7). If we choose a different Lagrangian, we may obtain a different theory of gravity. Many alternative theories of gravity have been considered (and most of them have been rejected). Motivated by the requirement of general covariance, one usually constrains the choice of gravitational Lagrangians to combinations of invariant quantities, such as R, $R_{\mu\nu}R^{\mu\nu}$, $R_{\kappa\lambda\mu\nu}R^{\kappa\lambda\mu\nu}$, $\Box R$, etc. These modified theories of gravity give rise to equations with high-order derivatives of $g_{\mu\nu}$ because the Ricci scalar $R[g_{\mu\nu}]$ contains second derivatives of $g_{\mu\nu}$.

We will consider only a relatively simple modification of the Einstein-Hilbert action consists of replacing the Ricci scalar R by a nonlinear function of R, so that the action is

$$S_{NL}[g] = \int_{\mathcal{M}} \sqrt{-g}\mathrm{d}^n p\, f(R). \tag{5.16}$$

(The function $f(R)$ is nonlinear iff $f''(R) \neq 0$). Such theories are called theories of **nonlinear gravity**, $f(R)$ **gravity**, or **scalar-tensor gravity** for the reason explained below.[2]

Let us now derive the equation of motion for $g_{\mu\nu}$ from the $f(R)$ Lagrangian. We shall follow the derivation in Sec. 5.1.2 with appropriate modifications, using the index notation where convenient.

As before, the variation of $S_{NL}[g]$ can be split into three terms,

$$\delta S_{NL}[g] = \delta \int_{\mathcal{M}} \mathrm{d}^4 p\, f\left(R_{\mu\nu}g^{\mu\nu}\right)\sqrt{-g}$$
$$= \int_{\mathcal{M}} \mathrm{d}^4 p \sqrt{-g}\left[f'(R)\left(g^{\mu\nu}\delta R_{\mu\nu} + R_{\mu\nu}\delta g^{\mu\nu}\right) - \frac{f(R)}{2}g_{\mu\nu}\delta g^{\mu\nu}\right].$$

The term $g^{\mu\nu}\delta R_{\mu\nu} = q^{\alpha}{}_{;\alpha}$ is a total divergence, as shown by Eq. (5.14), but this term is now multiplied by $f'(R)$ which is not a constant since, by assumption, $f(R)$ is nonlinear. Therefore the integral

$$\int_{\mathcal{M}} \mathrm{d}^4 p \sqrt{-g}\, f'(R) g^{\mu\nu}\delta R_{\mu\nu} = \int_{\mathcal{M}} d^4 p \sqrt{-g} f'(R) q^{\alpha}{}_{;\alpha}$$

does not identically vanish any more, but instead contributes to the equation for $g_{\mu\nu}$. Integrating by parts and omitting boundary terms, we have

$$\int_{\mathcal{M}} \mathrm{d}^4 p \sqrt{-g}\, f'(R) q^{\alpha}{}_{;\alpha} = -\int_{\mathcal{M}} d^4 p \sqrt{-g}\left[f'(R)\right]_{;\alpha} q^{\alpha}.$$

However, q^{α} still contains derivatives of $\delta g^{\mu\nu}$, and we now need to use Eq. (5.15) to find

$$q^{\beta} = \delta g^{\alpha\beta}{}_{;\alpha} - g_{\mu\nu}\delta g^{\mu\nu;\beta}. \tag{5.17}$$

Calculation: Derive Eq. (5.17) from Eqs. (5.14) and (5.15).

Hint: Note that $\delta g^{\mu\nu}$ is not equal to $\delta g_{\mu\nu}$ with "indices raised," but

$$\delta g_{\alpha\beta} = -g_{\alpha\lambda}g_{\beta\mu}\delta g^{\mu\nu},$$

where $\delta g_{\alpha\beta}$ is the perturbation in $g_{\mu\nu}$ entering Eq. (5.15). Express q^{α} through $\delta g^{\mu\nu}$ rather than through $\delta g_{\mu\nu}$.

Finally, we obtain

$$-\int_{\mathcal{M}} \mathrm{d}^4 p \sqrt{-g}\left[f'(R)\right]_{;\alpha}\delta q^{\alpha}$$
$$= \int_{\mathcal{M}} \mathrm{d}^4 p \sqrt{-g}\left(\left[f'(R)\right]^{;\alpha}_{;\alpha} g_{\mu\nu} - \left[f'(R)\right]_{;\mu\nu}\right)\delta g^{\mu\nu}.$$

[2] See, for instance, G. Magnano and L. M. Sokolowski, arxiv:gr-qc/9312008.

Therefore, the equation of motion for $g_{\mu\nu}$ is

$$f'(R)R_{\mu\nu} - \frac{1}{2}f(R)g_{\mu\nu} + g_{\mu\nu}\Box f'(R) - \left[f'(R)\right]_{;\mu\nu} = 0.$$

This equation has fourth-order derivatives of the metric and is quite complicated. In order to simplify this theory, the standard practice is to reduce the above equation to a system of second-order equations by introducing an auxiliary scalar field and changing the metric by a suitable conformal transformation.

Rather than guessing the necessary change of variables, one can perform the simplification in a systematic way by using the method of Lagrange multipliers. I will first describe the application of this method to a trivial example borrowed from classical mechanics. Then I will explain the application of that method to the theory (5.16).

In classical mechanics, the dynamics of a system may be described by the action

$$S[\mathbf{q}] = \int L(\mathbf{q}, \dot{\mathbf{q}})dt,$$

where $\mathbf{q}(t)$ is the trajectory of the system, which we will assume to be a vector-valued function of time. If the Lagrangian $L[\mathbf{q}, \dot{\mathbf{q}}]$ depends on first derivatives of $\mathbf{q}(t)$ nonlinearly (which is usually the case), the equations of motion for $\mathbf{q}(t)$ are second-order (they contain $\ddot{\mathbf{q}}$). Suppose we would like to reduce the Lagrangian to a simpler form that generates only first-order equations of motion. One way is to introduce a new independent variable $\mathbf{v}(t)$ which is equal to the velocity $\dot{\mathbf{q}}(t)$; then one expects to obtain a system of first-order equations for $\{\mathbf{q}(t), \mathbf{v}(t)\}$. A straightforward way to replace velocities with an independent variable is to add a constraint $\dot{\mathbf{q}} - \mathbf{v} = 0$ to the Lagrangian with a Lagrange multiplier. Since we have a vector-valued constraint at every moment of time t, the Lagrange multiplier must also be vector-valued function of time. Let us denote this vector-valued Lagrange multiplier by $\mathbf{s}(t)$. Thus the modified action can be written as

$$\tilde{S}[\mathbf{q}, \mathbf{v}, \mathbf{s}] = \int \left[L(\mathbf{q}, \mathbf{v}) + \mathbf{s}\cdot(\dot{\mathbf{q}} - \mathbf{v})\right]dt. \tag{5.18}$$

The variation of the new action $\tilde{S}$ with respect to three independent vector-valued functions $\mathbf{q}, \mathbf{v}, \mathbf{s}$ yields equations of motion that are equivalent to the original ones.

We note that the variables $\mathbf{s}$ and $\mathbf{v}$ enter the action (5.18) as Lagrange multipliers; the action involves derivatives of $\mathbf{q}$ but no derivatives of $\mathbf{s}$ or $\mathbf{v}$. As we suspect that two Lagrange multipliers constraining a single function $\mathbf{q}$ are too many, we would like to eliminate one of these Lagrange multipliers. To this end, let us compute the variation of $\tilde{S}$ with respect to $\mathbf{v}$:

$$\frac{\delta\tilde{S}[\mathbf{q}, \mathbf{v}, \mathbf{s}]}{\delta\mathbf{v}(t)} = \frac{\partial L(\mathbf{q}(t), \mathbf{v}(t))}{\partial \mathbf{v}} - \mathbf{s}(t).$$

This is an algebraic equation that can be solved with respect to $\mathbf{v}$ (if L is a nonlinear and nondegenerate function of $\mathbf{v}$). Let us denote by $\mathbf{V}(\mathbf{q}, \mathbf{s})$ the functions that express $\mathbf{v}$ through $\mathbf{q}$ and $\mathbf{s}$,

$$\mathbf{v} = \mathbf{V}(\mathbf{q}, \mathbf{s}).$$

Since the action (5.18) contains $\mathbf{v}$ merely as a Lagrange multiplier (no derivatives of $\mathbf{v}$ are involved), we may now substitute $\mathbf{v} = \mathbf{V}(\mathbf{q}, \mathbf{s})$ into the action (5.18) and obtain an equivalent action that depends only on $\mathbf{q}$ and $\mathbf{s}$:

$$\tilde{S}[\mathbf{q}, \mathbf{s}] = \int \left[L(\mathbf{q}, \mathbf{V}(\mathbf{q}, \mathbf{s})) - \mathbf{s}\cdot\mathbf{V}(\mathbf{q}, \mathbf{s}) + \mathbf{s}\cdot\dot{\mathbf{q}}\right]dt.$$

This action is linear in time derivatives, and hence the equations of motion for $\mathbf{q}(t), \mathbf{s}(t)$ are first-order in the time derivatives. One may recognize this as the Hamiltonian action, where $\mathbf{s}$ is the canonical momentum corresponding to $\mathbf{q}$, and

$$L - \mathbf{s} \cdot \mathbf{V}(\mathbf{s}, \mathbf{q}) \equiv H(\mathbf{s}, \mathbf{q})$$

is the Hamiltonian.

Remark: As we have seen, the Lagrange multiplier $\mathbf{s}$ has the significance of the canonical momentum. Usually, the canonical momentum corresponding to $\mathbf{q}$ would be denoted by the letter $\mathbf{p}$. In the present derivation, I chose a different letter, $\mathbf{s}$, because I wanted to emphasize that we do not know the significance of the new variables in advance. It will not be always the case that the new variables introduced through the Lagrange multiplier method have the significance of canonical momenta.

The method of Lagrange multipliers is more general than the familiar passage from the Lagrangian to the Hamiltonian description. For instance, one could treat Lagrangians containing higher derivatives in the same systematic manner. One could also replace only *some* of the higher derivatives but not all, if this proves to be convenient in a particular situation. ■

Let us now apply the method of Lagrange multipliers to the action (5.16). At this point, we do not have the purpose of reducing the system to first-order equations, but rather to remove the nonlinearity in the $f(R)$ term. Therefore, we are motivated to introduce a new field r equal to the Ricci scalar $R[g_{\mu\nu}]$. Hence, an action equivalent to the action (5.16) is

$$\int_{\mathcal{M}} \sqrt{-g} \mathrm{d}^n p \left[f(r) + \left(R[g_{\mu\nu}] - r \right) \lambda \right], \tag{5.19}$$

where λ is a Lagrange multiplier field. Variation of the above action with respect to r yields the algebraic equation

$$f'(r) = \lambda,$$

which can be solved to express r through λ (since by assumption $f''(r) \neq 0$). Let us denote by $r(\lambda)$ the function obtained by solving the algebraic equation $f'(r) = \lambda$. We can then substitute r into the action (5.19) and obtain an equivalent action,

$$S[g_{\mu\nu}, \lambda] = \int_{\mathcal{M}} \sqrt{-g} \mathrm{d}^n p \left[f(r(\lambda)) - r(\lambda)\lambda + \lambda R[g_{\mu\nu}] \right], \tag{5.20}$$

which now depends only on the metric $g_{\mu\nu}$ and the new field λ. Note that the field λ must be varied independently of $g_{\mu\nu}$; the constraint $R[g_{\mu\nu}] = r(\lambda)$ will be a consequence of the equations of motion.

At this point, the action (5.20) depends *linearly* on R, and we can use another trick to simplify the equations further. Namely, we perform a suitable conformal transformation of the metric. According to Eq. (3.21) with $N = 4$, a conformal transformation $g_{\mu\nu} \to \tilde{g}_{\mu\nu}$ with a conformal factor $e^{2\Omega} \neq 0$ will change the variables as follows,

$$g_{\mu\nu} = e^{2\Omega} \tilde{g}_{\mu\nu}, \quad \sqrt{-g} = e^{4\Omega} \sqrt{-\tilde{g}},$$
$$R[g_{\mu\nu}] = e^{-2\Omega} \left(R[\tilde{g}_{\mu\nu}] + 6\tilde{\Box}\Omega + 6\tilde{g}^{\mu\nu}\Omega_{,\mu}\Omega_{,\nu} \right),$$

where the D'Alembert operator $\tilde{\Box}$ is defined with respect to the new metric $\tilde{g}_{\mu\nu}$. We can now express the action (5.20) through the new metric. An appropriate choice of Ω, namely

$$\Omega = -\frac{1}{2} \ln \lambda, \quad e^{2\Omega} = \frac{1}{\lambda}, \tag{5.21}$$

cancels the factor λ in the action (5.20). The action in the new variables is

$$S[g_{\mu\nu}, \lambda] \equiv \tilde{S}[\tilde{g}_{\mu\nu}, \Omega] = \int_{\mathcal{M}} \sqrt{-\tilde{g}} \mathrm{d}^n p \left[\frac{f(r(\lambda)) - r(\lambda)\lambda}{\lambda^2} + R[\tilde{g}_{\mu\nu}] + 6\tilde{\Box}\Omega + 6\tilde{g}^{\mu\nu}\Omega_{,\mu}\Omega_{,\nu} \right]. \tag{5.22}$$

The term $6\tilde{\Box}\Omega$ is a total covariant derivative and can be omitted from the action. Also, we may express λ through Ω using Eq. (5.21) and regard Ω as a new scalar field. Introducing the auxiliary function

$$V(\Omega) \equiv \left. \frac{r(\lambda)\lambda - f(r(\lambda))}{\lambda^2} \right|_{\lambda = \lambda(\Omega)},$$

we rewrite the action (5.22) as

$$\tilde{S}[\tilde{g}_{\mu\nu}, \Omega] = \int_{\mathcal{M}} \sqrt{-\tilde{g}} \mathrm{d}^n p \left[R[\tilde{g}_{\mu\nu}] + 6\tilde{g}^{\mu\nu}\Omega_{,\mu}\Omega_{,\nu} - V(\Omega) \right].$$

This is the action of usual, unmodified Einstein gravity, minimally coupled to a scalar field Ω with a self-interaction potential $V(\Omega)$. The form of the potential $V(\Omega)$ is ultimately determined by the function $f(R)$ in the original action (5.16).

The additional scalar field Ω can be heuristically viewed as a "scalar component" of the original gravitational field $g_{\mu\nu}$. The metric tensor $\tilde{g}_{\mu\nu}$ alone is insufficient to describe the effects of gravity in these nonlinear theories. Therefore, these models are also called **scalar-tensor theories** of gravity. The original field variable $g_{\mu\nu}$ is called the **Jordan frame** variables, while the transformed variables $\{\Omega, \tilde{g}_{\mu\nu}\}$ are called the **Einstein frame** variables. The name "Einstein frame" means that in that frame the theory looks like the ordinary Einstein gravity coupled to some additional matter fields. Calculations are often easier in the Einstein frame. However, the metric $\tilde{g}_{\mu\nu}$ is an auxiliary variable that does not describe the physically observed metric (i.e. the action for other matter fields, such as the electromagnetic field, contains $g_{\mu\nu}$ and not $\tilde{g}_{\mu\nu}$). So one needs to perform the conformal transformation back to the Jordan frame to recover the physical metric $g_{\mu\nu}$.

After this brief excursion into alternative theories of gravitation, I continue to consider only the currently standard theory (Einstein's General Relativity).

5.1.4 Energy-momentum tensor

Consider a field theory that contains some matter fields ϕ_j interacting with gravity. In the Lagrangian formulation, the total system is described by an action of the form

$$S[\phi_j; g_{\mu\nu}] = \int_{\mathcal{M}} \left(L_{EH}[g] + L[\phi_j; g_{\mu\nu}] \right) \sqrt{-g} \mathrm{d}^4 x,$$

where $L[\phi_j; g_{\mu\nu}]$ is a Lagrangian for the matter fields and their interactions among themselves, while

$$L_{EH}[g] \equiv \frac{1}{16\pi G} R$$

is the Einstein-Hilbert Lagrangian for gravity. The Euler-Lagrange equation for a matter field ϕ_j is then

$$\frac{\delta L[\phi; g_{\mu\nu}]}{\delta \phi_j} = 0,$$

while the equation for the gravitational field is

$$\frac{\delta L_{EH}[g]}{\delta g^{\alpha\beta}} + \frac{\delta L[\phi_j; g_{\mu\nu}]}{\delta g^{\alpha\beta}} = 0.$$

We have already computed $\delta L_{EH}/\delta g^{\alpha\beta}$, and so we can rewrite this equation as the **Einstein equation**,

$$R_{\alpha\beta} - \frac{1}{2} R g_{\alpha\beta} = 16\pi G \frac{\delta L[\phi_j; g_{\mu\nu}]}{\delta g^{\alpha\beta}} \equiv -8\pi G T_{\alpha\beta},$$

where we have defined the tensor $T_{\alpha\beta}$ called the **energy-momentum tensor** (EMT) of matter,

$$T_{\alpha\beta} \equiv \frac{2}{\sqrt{-g}} \frac{\delta L[\phi_j; g_{\mu\nu}]}{\delta g^{\alpha\beta}}. \tag{5.23}$$

This tensor plays the role of a "source" for the gravitational field. In most field theories, this tensor also describes the distribution of local energy and momentum density of the field, defined in the conventional sense.

For example, a minimally coupled scalar field with the Lagrangian (5.3) has the EMT

$$T_{\alpha\beta} = \phi_{,\alpha}\phi_{,\beta} - \frac{1}{2} g_{\alpha\beta} \phi_{,\mu}\phi^{,\mu} + V(\phi) g_{\alpha\beta}. \tag{5.24}$$

The electromagnetic field is described by the Maxwell tensor $F_{\mu\nu}$, which satisfies the Maxwell equations. These equations can be derived from the Lagrangian

$$L_{EM}[F; g] = -\frac{\sqrt{-g}}{16\pi} F_{\alpha\beta} F^{\alpha\beta} = -\frac{\sqrt{-g}}{16\pi} F_{\alpha\beta} F_{\mu\nu} g^{\alpha\mu} g^{\beta\nu}, \tag{5.25}$$

from which also follows the EMT

$$T_{\mu\nu} = \frac{1}{4\pi} \left(\frac{1}{4} g_{\mu\nu} F_{\alpha\beta} F^{\alpha\beta} - F_{\alpha\mu} F_{\beta\nu} g^{\alpha\beta} \right).$$

Calculation: Derive the above expressions for the EMT of a scalar field and of the Maxwell field.

Solution: If a Lagrangian is of the form $L = \sqrt{-g}L$ then

$$\frac{2}{\sqrt{-g}} \frac{\delta L}{\delta g^{\alpha\beta}} = -L g_{\alpha\beta} + 2 \frac{\delta L}{\delta g^{\alpha\beta}}.$$

In the scalar field Lagrangian, the only term that depends on $g^{\mu\nu}$ is $\frac{1}{2} g^{\mu\nu} \phi_{,\mu}\phi_{,\nu}$. Hence,

$$\frac{\delta L[g, \phi]}{\delta g^{\alpha\beta}} = \frac{1}{2} \phi_{,\alpha}\phi_{,\beta},$$
$$T_{\alpha\beta} = \phi_{,\alpha}\phi_{,\beta} - g_{\alpha\beta} L$$

as required. The same procedure for the Lagrangian L_{EM} yields the required answer since

$$\frac{\delta L}{\delta g^{\mu\nu}} = -\frac{1}{16\pi} \left(F_{\alpha\mu} F_{\beta\nu} g^{\alpha\beta} + F_{\mu\alpha} F_{\nu\beta} g^{\alpha\beta} \right) = -\frac{g^{\alpha\beta}}{8\pi} F_{\alpha\mu} F_{\beta\nu}.$$

Calculation: Compute the 0-0 component of the electromagnetic EMT and show that it coincides with the familiar expression for the energy density of the electromagnetic field,

$$T_{00} = \frac{1}{8\pi} \left(|\vec{E}|^2 + |\vec{B}|^2 \right).$$

5.1.5 General covariance

Consider a field theory where every matter field ϕ_i, $i = 1, ..., N$, is coupled to gravity through the metric tensor $g_{\mu\nu}$ and the covariant derivatives ∇. Such a field theory is invariant under general coordinate transformations; this property is called the **general covariance** of the theory. General coordinate transformations can be parametrized by four functions $\tilde{\mathbf{x}}(\mathbf{x})$. According to the Noether theorem, the invariance with respect to gauge transformations leads to mathematical identities that hold regardless of the equations of motion. However, we *can* obtain useful conservation laws out of a gauge symmetry if we *do* use the equations of motion. We shall now show that the requirement of general covariance for a set of fields ϕ_i with a Lagrangian $L[\phi_i; g]$, together with the Euler-Lagrange equation of motion,

$$\frac{\delta L}{\delta \phi_i(x)} = 0, \tag{5.26}$$

yields a conservation law for the field's energy-momentum tensor $T_{\mu\nu}$ defined by Eq. (5.23).

To derive the conservation law, we consider an infinitesimal coordinate transformation

$$\mathbf{x} \to \tilde{\mathbf{x}} = \mathbf{x} + \varepsilon \mathbf{f}(\mathbf{x}),$$

where $\mathbf{f}$ is a vector field, which is the "generator" of the transformation (points are shifted along the orbits of $\mathbf{f}$). The matter fields ϕ_i are transformed by the flow of $\mathbf{f}$ according to

$$\phi_i(\mathbf{x}) \to \tilde{\phi}_i(\mathbf{x}) = \phi_i(\tilde{\mathbf{x}}) + \delta q_i(\tilde{\mathbf{x}}),$$

where $\delta q_i(\mathbf{x})$ is defined appropriately for each field, such that the local variation of the field ϕ_i is $\delta\phi_i = \varepsilon \mathcal{L}_{\mathbf{f}} \phi_i$. For instance, the metric $g^{\alpha\beta}$ is transformed as

$$g^{\alpha\beta}(\mathbf{x}) \to \tilde{g}^{\alpha\beta}(\mathbf{x}) = g^{\alpha\beta}(\mathbf{x}) + \varepsilon \mathcal{L}_{\mathbf{f}} g^{\alpha\beta}(\mathbf{x}) = g^{\alpha\beta} + \varepsilon \left(f^{\alpha;\beta} + f^{\beta;\alpha} \right).$$

We know that the action is invariant under the transformation, therefore the variation $\delta \int L d^4x$ must vanish:

$$\begin{aligned} 0 = \delta \int L d^4x &= \int \left[\frac{\delta L}{\delta g^{\alpha\beta}(\mathbf{x})} \delta g^{\alpha\beta}(\mathbf{x}) + \frac{\delta L}{\delta \phi_i(\mathbf{x})} \delta\phi_i(\mathbf{x}) \right] \mathrm{d}^4x \\ &= \int \frac{\delta L}{\delta g^{\alpha\beta}(\mathbf{x})} \varepsilon \left(f^{\alpha;\beta} + f^{\beta;\alpha} \right) \mathrm{d}^4x + \int \frac{\delta L}{\delta \phi_i(\mathbf{x})} \delta\phi_i(\mathbf{x}) \mathrm{d}^4x. \end{aligned} \tag{5.27}$$

Since the fields ϕ_i satisfy Eq. (5.26), the second term vanishes. Expressing the first term through the tensor $T_{\alpha\beta}$, we get

$$\begin{aligned} \int T_{\alpha\beta} f^{\beta;\alpha} \sqrt{-g} \mathrm{d}^4x &= \int \left[\left(T_{\alpha\beta} f^{\beta} \right)^{;\alpha} - T^{;\alpha}_{\alpha\beta} f^{\beta} \right] \sqrt{-g} \mathrm{d}^4x \\ &= -\int T^{;\alpha}_{\alpha\beta} f^{\beta} \sqrt{-g} \mathrm{d}^4x = 0. \end{aligned} \tag{5.28}$$

Here we assumed that f^α vanishes at infinity sufficiently quickly. Since Eq. (5.28) must be satisfied for arbitrary $f^\alpha(\mathbf{x})$, we conclude that the conservation law $T^{;\alpha}_{\alpha\beta} = 0$ holds.

Remark: The absence of the covariant volume factor $\sqrt{-g}$ in Eq. (5.27) is not a mistake; the result is nevertheless a covariant quantity. The derivative with respect to f^α is calculated using the chain rule, e.g.

$$\delta \int L \mathrm{d}^4x = \int \left[\frac{\delta L}{\delta g^{\alpha\beta}(\mathbf{x})} \delta g^{\alpha\beta}(\mathbf{x}) + \frac{\delta L}{\delta \phi_i} \delta\phi_i(\mathbf{x}) \right] \mathrm{d}^4x,$$

and the rule requires a simple integration over d^4x. The correct covariant behavior is supplied by the Lagrangian L that itself contains a factor $\sqrt{-g}$.

In a flat spacetime, the laws of energy and momentum conservation follow from the invariance of the action under spacetime translations. In the presence of gravitation the spacetime is curved, so in general the spacetime translations are not a physical symmetry any more. However, the action is covariant with respect to arbitrary coordinate transformations. The corresponding conservation law is the "covariant conservation" of the EMT, $T^{;\alpha}_{\alpha\beta} = 0$. However, this law does not actually express the conservation of energy or momentum of the matter field ϕ_i because of the presence of the covariant derivative. That equation would be a conservation law if it had the form $\partial_\mu \left(\sqrt{-g}T^{\mu\nu}\right) = 0$, but instead it can be shown that

$$\partial_\mu \left(\sqrt{-g}T^{\mu\nu}\right) = -\sqrt{-g}\Gamma^\nu_{\mu\lambda}T^{\mu\lambda} \neq 0,$$

where $\Gamma^\nu_{\mu\lambda}$ is the Christoffel symbol. The energy of the matter fields alone, described by the energy-momentum tensor $T_{\alpha\beta}$, is not necessarily conserved; the gravitational field can change the energy and the momentum of matter.

Remark: For one scalar field ϕ with the Lagrangian (5.3), the conservation of energy-momentum tensor, $T_{\mu\nu}{}^{;\mu} = 0$, where $T_{\mu\nu}$ is given by Eq. (5.24), and the equation of motion, $g^{\mu\nu}\phi_{;\mu\nu} + V'(\phi) = 0$, are equivalent:

$$\begin{aligned} 0 = T_{\mu\nu}{}^{;\mu} &= \left(\phi_{,\mu}\phi_{,\nu}\right)^{;\mu} - \frac{1}{2}\left(\phi_{,\alpha}\phi^{,\alpha}\right)_{;\nu} + V(\phi)\phi_{,\nu} \\ &= \left(\phi^{;\mu}_{;\mu} + V(\phi)\right)\phi_{,\nu}. \end{aligned}$$

In a theory of a single scalar field ϕ coupled to gravity, the conservation law $T_{\mu\nu}{}^{;\mu} = 0$ is essentially a consequence of the Einstein equation. Therefore, one may say that the Einstein equation contains the equation of motion for the scalar field. However, the same statement will not hold for theories with more than one scalar field. A single conservation law cannot yield several independent equations of motion for all the fields.

5.1.6 Symmetries and Noether theorems

The fundamental theorems of E. Noether explain the relationship between symmetries and conservation laws. We show simple examples of theories with symmetries.

Translational symmetry

A field theory is described by a Lagragian, and a **symmetry** of a field theory means a transformation that leaves the Lagrangian invariant. For example, consider the theory of a scalar field with the Lagrangian $L[\phi]$, for instance, that of Eq. (5.3) in a flat, Minkowski spacetime, where the metric is $g = \eta = \mathrm{diag}(1, -1, -1, -1)$. Since this Lagrangian in this spacetime does not depend explicitly on the coordinates, it is invariant under a translation of fields, $\phi(\mathbf{x}) \to \phi(\mathbf{x} + \mathbf{a})$, where $\mathbf{a}$ is an arbitrary constant 4-vector. We shall now study the consequences of this symmetry for any Lagrangian $L[\phi]$ that depends only on ϕ and $\nabla\phi$.

Remark: In this section only, we denote spacetime points by boldface letters $\mathbf{x} \equiv (t, x, y, z)$ for brevity. This notation is different from the notation in the rest of the text, where boldface letters are *vectors* or vector fields. In the Minkowski spacetime, one can identify points (t, x, y, z) with vectors from $\mathbb{R}^4$, but in a curved spacetime points are not vectors.

Let us first examine the transformation $\phi(\mathbf{x}) \to \phi(\mathbf{x} + \mathbf{a})$ in more detail. The transformation is understood as the replacement of the function $\phi(\mathbf{x})$ by a different function $\tilde{\phi}(\mathbf{x})$,

$$\phi(\mathbf{x}) \to \tilde{\phi}(\mathbf{x}) = \phi(\mathbf{x} + \mathbf{a}).$$

Note that the *value* of the new function $\tilde{\phi}$ at a location $\mathbf{x}$ is *equal* to the value of the old function ϕ at the location $\mathbf{x} + \mathbf{a}$; i.e. the value of the function is not modified beyond the necessary change due to the replacement of the argument, $\mathbf{x} \to \mathbf{x} + \mathbf{a}$.

The Lagrangian $L[\tilde{\phi}]$ is invariant under the replacement $\mathbf{x} \to \mathbf{x} + \mathbf{a}$ in the following sense: the *value* of $L[\tilde{\phi}](\mathbf{x})$, computed at a fixed location $\mathbf{x}$, is *equal* to the value of $L[\phi](\mathbf{x} + \mathbf{a})$, i.e., to the value of the same function $L[\phi]$, computed for the old field ϕ, at the location $\mathbf{x} + \mathbf{a}$. Let us now express this invariance as a mathematical identity.

It is convenient to consider an infinitesimal displacement $\delta\mathbf{a}$. At a given location $\mathbf{x}$, the value of any scalar function $f(\mathbf{x})$ changes under $\mathbf{x} \to \mathbf{x} + \delta\mathbf{a}$, to first order, as

$$\delta f(\mathbf{x}) \equiv \tilde{f}(\mathbf{x}) - f(\mathbf{x}) = f(\mathbf{x} + \delta\mathbf{a}) - f(\mathbf{x}) = \nabla_{\delta\mathbf{a}} f + O(\delta\mathbf{a}^2).$$

This is the **local variation** of the value of the function f. (From now on, we shall drop the terms $O(\delta\mathbf{a}^2)$ since we need to consider only first-order variations.) Because of the assumption of translational invariance, the same argument applies also to the Lagrangian $L[\phi](\mathbf{x})$, considered as a function of $\mathbf{x}$,

$$\begin{aligned} \delta L[\phi(\mathbf{x})] &\equiv L[\tilde{\phi}(\mathbf{x})] - L[\phi(\mathbf{x})] \\ &= L[\phi](\mathbf{x} + \delta\mathbf{a}) - L[\phi](\mathbf{x}) = \nabla_{\delta\mathbf{a}} L[\phi](\mathbf{x}). \end{aligned}$$

The above relation is, essentially, the requirement of translational invariance of the Lagrangian L. On the other hand, the same variation δL can be computed directly, expressing it through the local variation of the field,

$$\delta\phi(\mathbf{x}) \equiv \tilde{\phi}(\mathbf{x}) - \phi(\mathbf{x}) = \nabla_{\delta\mathbf{a}}\phi(\mathbf{x}).$$

We have

$$\delta L[\phi(\mathbf{x})] = \frac{\partial L}{\partial \phi}\delta\phi + \frac{\partial L}{\partial \phi_{,\mu}}\delta\phi_{,\mu}.$$

Thus the same quantity δL can be expressed in two different ways:

$$\delta L[\phi(\mathbf{x})] = \frac{\partial L}{\partial \phi}\delta\phi + \frac{\partial L}{\partial \phi_{,\mu}}\delta\phi_{,\mu} = \nabla_{\delta\mathbf{a}} L.$$

This identity expresses the invariance of the Lagrangian L under translations. A conservation law can be now derived, assuming that the Euler-Lagrange equation holds for a specific field configuration $\phi(\mathbf{x})$,

$$\frac{\partial L[\phi]}{\partial \phi} - \nabla_\mu \frac{\partial L[\phi]}{\partial \phi_{,\mu}} = 0.$$

Namely, since $\delta\mathbf{a}$ is a constant vector, we can express $\delta\phi$ and δL as total divergences,

$$\delta\phi = \nabla_{\delta\mathbf{a}}\phi = \phi_{,\mu}\delta a^\mu = \nabla_\mu\left(\phi\delta a^\mu\right), \quad \nabla_{\delta\mathbf{a}} L = \nabla_\mu\left(L\delta a^\mu\right),$$

and then rewrite the above identity as

$$\begin{aligned} 0 &= \frac{\partial L}{\partial \phi}\delta\phi + \frac{\partial L}{\partial \phi_{,\mu}}\delta\phi_{,\mu} - \nabla_{\delta\mathbf{a}} L \\ &= \left(\frac{\partial L}{\partial \phi} - \nabla_\mu \frac{\partial L}{\partial \phi_{,\mu}}\right)\delta\phi + \nabla_\mu\left(\frac{\partial L}{\partial \phi_{,\mu}}\phi_{,\nu} - L\delta^\mu_\nu\right)\delta a^\nu. \end{aligned}$$

Since the Euler-Lagrange equation is assumed to hold and $\delta\mathbf{a}$ is arbitrary, it follows that the following conservation law holds,

$$0 = \nabla_\mu \left(\frac{\partial L}{\partial \phi_{,\mu}} \phi_{,\nu} - L g_{\mu\nu} \right) \equiv \nabla_\mu T^\mu_\nu,$$

where

$$T^\mu_\nu = \frac{\partial L}{\partial \phi_{,\mu}} \phi_{,\nu} - L g_{\mu\nu} \tag{5.29}$$

is the energy-momentum tensor of the field ϕ. Indeed, the tensor T^μ_ν coincides (after lowering the index) with the expression (5.24).

A relation of the form $\nabla_\mu j^\mu \equiv \mathrm{div}\mathbf{j} = 0$ in flat space is called a **conservation law**, and the vector field $\mathbf{j}$ is called a **conserved current**. A simple interpretation can be given in Minkowski coordinates $(t, \vec{x})$: The 4-vector j^μ is decomposed into the time component j^0 and a spatial part $\vec{j}$. Then j^0 is a density of some "substance" at a point, while $\vec{j}$ is the 3-velocity. The conservation law $0 = j^\mu{}_{,\mu} = \partial_t j^0 + \mathrm{div}\vec{j}$ shows that the change in the density j^0 is always due to the transport of the "substance" from neighbor points, which means that the amount of the "substance" is conserved. Alternatively, we may consider a spacelike 3-surface $t = t_1$ and compute the total amount $Q(t_1)$ of the "substance" on the surface,

$$Q(t_1) \equiv \int_{t=t_1} j^0 \mathrm{d}^3\vec{x}.$$

The conservation law then says that $Q(t)$ is constant,

$$\frac{dQ(t)}{dt} = \int \frac{dj^0}{dt} \mathrm{d}^3\vec{x} = -\int \left(\mathrm{div}\vec{j}\right) \mathrm{d}^3\vec{x} = 0.$$

The quantity Q is called the **conserved charge** corresponding to the current j^μ with respect to a 3-surface. The equations of motion for the field are such that the charge Q is time-independent. This is the meaning of the conservation law.

More generally, we may consider a one-parametric group of coordinate transformations

$$\mathbf{x} \to \tilde{\mathbf{x}}(\mathbf{x}; \varepsilon), \quad \phi(\mathbf{x}) \to \tilde{\phi}(\mathbf{x}) \equiv \phi(\tilde{\mathbf{x}}),$$

where ε is a real parameter and $\varepsilon = 0$ corresponds to the identical transformation. If a Lagrangian $L[\phi; \mathbf{x}]$, possibly depending explicitly on $\mathbf{x}$, is invariant under this transformation group, in the sense explained above, then a similar calculation leads to the conservation law $j^\mu{}_{,\mu} = 0$ for a certain 4-vector field j^μ, called the **Noether current** corresponding to the symmetry transformation $\mathbf{x} \to \tilde{\mathbf{x}}$. The Noether current j^μ is then defined by

$$j^\mu = \frac{\partial L}{\partial \phi_{,\mu}} \phi_{,\nu} f^\nu - f^\mu L,$$

where

$$f^\mu \equiv \left. \frac{\partial \tilde{\mathbf{x}}}{\partial \varepsilon} \right|_{\varepsilon=0}.$$

Calculation: Derive the above expression for the Noether current j^μ corresponding to the transformation $\mathbf{x} \to \tilde{\mathbf{x}}$, $\phi \to \tilde{\phi}(\mathbf{x}) \equiv \phi(\tilde{\mathbf{x}})$, assuming that

$$\int d^4\mathbf{x}\, L[\tilde{\phi}(\mathbf{x}); \mathbf{x}] = \int d^4\tilde{\mathbf{x}}\, L[\phi(\tilde{\mathbf{x}}); \tilde{\mathbf{x}}].$$

Solution: Consider a transformation with an "infinitesimal" value of ε,

$$\mathbf{x} \to \tilde{\mathbf{x}} = \mathbf{x} + \mathbf{f}(\mathbf{x})\varepsilon, \quad \mathbf{f}(\mathbf{x}) \equiv \left. \frac{\partial \tilde{\mathbf{x}}(\mathbf{x}; \varepsilon)}{\partial \varepsilon} \right|_{\varepsilon=0}.$$

The vector field $\mathbf{f}$ describes the "flow" of the infinitesimal coordinate transformation. The local variation of the field ϕ is

$$\delta\phi \equiv \tilde{\phi}(\mathbf{x}) - \phi(\mathbf{x}) = \phi(\tilde{\mathbf{x}}) - \phi(\mathbf{x}) = \varepsilon f^\alpha \phi_{,\alpha} = \varepsilon \mathbf{f} \circ \phi.$$

The local variation of the derivative $\phi_{,\mu}$ is slightly more complicated due to the dependence of $\mathbf{f}$ on $\mathbf{x}$,

$$\delta(\phi_{,\mu}) = \frac{\partial \tilde{\phi}(\mathbf{x})}{\partial x^\mu} - \frac{\partial \phi(\mathbf{x})}{\partial x^\mu} = \frac{\partial \phi(\mathbf{x} + \varepsilon \mathbf{f})}{\partial x^\mu} - \frac{\partial \phi(\mathbf{x})}{\partial x^\mu} = (\varepsilon f^\alpha \phi_{,\alpha})_{,\mu}.$$

The local variation of the Lagrangian (at fixed $\mathbf{x}$) due to the local variation of the field is

$$\begin{aligned} \delta L[\phi(\mathbf{x}); \mathbf{x}] &= \frac{\partial L}{\partial \phi} \delta\phi + \frac{\partial L}{\partial \phi_{,\mu}} \delta(\phi_{,\mu}) \\ &= \left(\frac{\partial L}{\partial \phi_{,\mu}} \delta\phi \right)_{,\mu} = \left(\frac{\partial L}{\partial \phi_{,\mu}} \phi_{,\nu} f^\nu \varepsilon \right)_{,\mu}, \end{aligned}$$

where we have used the Euler-Lagrange equation. On the other hand, the invariance of the action under the transformation means that

$$\int \mathrm{d}^4\mathbf{x}\, \delta L[\phi(\mathbf{x}); \mathbf{x}] = \int \mathrm{d}^4\tilde{\mathbf{x}}\, L[\phi(\tilde{\mathbf{x}}); \tilde{\mathbf{x}}] - \int \mathrm{d}^4\mathbf{x}\, L[\phi(\mathbf{x}); \mathbf{x}].$$

The volume element is transformed as

$$\begin{aligned} \mathrm{d}^4\tilde{\mathbf{x}} &= \left(\mathrm{d}^4\mathbf{x}\right) \det \frac{\partial \tilde{x}^\mu}{\partial x^\nu} = \left(\mathrm{d}^4\mathbf{x}\right) \det \left(1 + \varepsilon \frac{\partial f^\mu}{\partial x^\nu}\right) \\ &= \left(\mathrm{d}^4\mathbf{x}\right) \left(1 + \varepsilon \frac{\partial f^\mu}{\partial x^\mu}\right) = \left(\mathrm{d}^4\mathbf{x}\right) (1 + \varepsilon \mathrm{div}\mathbf{f}). \end{aligned}$$

(Recall that $\det(1 + \varepsilon A) = 1 + \varepsilon \mathrm{Tr} A + O(\varepsilon^2)$ for a matrix A.) Since

$$L[\phi(\tilde{\mathbf{x}}); \tilde{\mathbf{x}}] - L[\phi; \mathbf{x}] = \varepsilon f^\mu \frac{\partial L}{\partial x^\mu} = \varepsilon \mathbf{f} \circ L,$$

we obtain the identity (up to terms of order ε^2)

$$\begin{aligned} 0 &= \left(\frac{\partial L}{\partial \phi_{,\mu}} \phi_{,\nu} f^\nu \varepsilon \right)_{,\mu} - \varepsilon f^\mu \frac{\partial L}{\partial x^\mu} - \varepsilon \frac{\partial f^\mu}{\partial x^\mu} L \\ &= \varepsilon \left(\frac{\partial L}{\partial \phi_{,\mu}} \phi_{,\nu} f^\nu - f^\mu L \right)_{,\mu} \equiv \varepsilon j^\mu{}_{,\mu}. \end{aligned}$$

Since ε is arbitrary, it follows that j^μ is a conserved current.

If the symmetry group has $n > 1$ parameters then n conservation laws can be found. Let us study this case in some more detail. Suppose we are given an n-parametric group of smooth transformations of a spacetime manifold $\mathcal{M}$. Continuous groups of transformations, i.e. groups that are themselves smooth multidimensional manifolds, are called **Lie groups**. If the symmetry group G is an n-dimensional Lie group then each element $\gamma \in G$ acts as a transformation $\mathbf{x} \to \tilde{\mathbf{x}}(\mathbf{x}; \gamma)$. An *infinitesimal* transformation $\mathbf{x} \to \mathbf{x} + \delta\mathbf{x}$ corresponds to an element of the group G which is "infinitesimally close" to the identity transformation $\mathbf{1} \in G$. Such elements can be (heuristically) parametrized as $\gamma = \exp(\varepsilon \mathbf{v})$ or "$\gamma = \mathbf{1} + \varepsilon\mathbf{v}$", where $\mathbf{v}$ is a tangent vector from the (n-dimensional) tangent space $T_\mathbf{1} G$ and $\exp(\mathbf{v})$ is the exponentiation map that produces points along the orbits of a vector $\mathbf{v}$. By definition, a tangent vector $\mathbf{v}$ is a derivation operator acting on functions on G. The transformation of points $\mathbf{x}$ corresponding to the vector $\mathbf{v}$ can be written as

$$\mathbf{x} \to \tilde{\mathbf{x}} = \mathbf{x} + \varepsilon \mathbf{f}(\mathbf{x}, \mathbf{v}),$$

where

$$\mathbf{f}(\mathbf{x},\mathbf{v}) \equiv \left.\frac{\partial \tilde{\mathbf{x}}(\mathbf{x})}{\partial \varepsilon}\right|_{\varepsilon=0} = \mathbf{v}\circ\tilde{\mathbf{x}}(\mathbf{x};\gamma),$$

and $\tilde{\mathbf{x}}(\mathbf{x};\gamma)$ is a function on G. It is clear that $\mathbf{f}(\mathbf{x},\mathbf{v})$ is linear in $\mathbf{v}$, thus $\mathbf{f}$ is a 1-form on $T_{\mathbf{1}}G$ with values in $T_{\mathbf{x}}\mathcal{M}$. In the index notation, $\mathbf{f}$ may be written as f_s^μ, where the index s is n-dimensional and labels the tangent space to the Lie group. The calculation leading to the definition of the Noether current yields

$$\left(\frac{\partial L}{\partial \phi_{,\mu}}\phi_{,\nu}f_s^\nu - f_s^\mu L\right)_{;\mu} v^s \equiv j^\mu_{s\ ;\mu} = 0,$$

thus the Noether current j_s^μ is also a 1-form on $T_{\mathbf{1}}G$ with values in $T_{\mathbf{x}}\mathcal{M}$. We have seen an example of such a Noether current (5.29): the group of transformations is $\mathbb{R}^4$ and thus the index s in j_s^μ can be identified as a four-dimensional Minkowski index, yielding a (1,1)-tensor T^μ_ν. The corresponding Noether charge

$$Q_\nu(t_0) = \int_{t=t_0} T^0_\nu \mathrm{d}^3x$$

represents the total 4-momentum on a spacelike 3-surface $t = t_0$. The total energy and the total momentum are conserved, $Q_\nu(t) = \text{const}$.

Remark: The EMT (5.29) obtained as a Noether current with respect to translation symmetry is called the **canonical EMT**, while the tensor defined by Eq. (5.23) is called the **metric EMT**. The canonical and the metric EMT coincide for the scalar field ϕ but not necessarily for other fields; however, this mismatch is merely a technical problem with the calculations. Both these tensors are conserved in flat space, and therefore the difference between them is a divergence-free tensor. Indeed, we can always add a divergence of an arbitrary antisymmetric tensor to a conserved current without changing the conservation law,

$$j^\mu \to \tilde{j}^\mu = j^\mu + (B^{\mu\nu} - B^{\nu\mu})_{,\nu}\,; \quad j^\mu{}_{,\mu} = \tilde{j}^\mu{}_{,\mu} = 0.$$

Note that the calculation leading to the definition of the Noether current j^μ only shows that $j^\mu{}_{,\mu} = 0$ and thus does not necessarily determine the physically correct expression for j^μ. It is the metric EMT, not the canonical EMT, that contributes to the Einstein equations and plays the role of the source of gravity.

Internal symmetry

The essense of Noether's theorem is that *every* symmetry of a Lagrangian leads to a conservation law. Our first example was a transformation that did not modify the value of the field ϕ (beyond the necessary change due to the coordinate shift). The second example is a field theory with an **internal symmetry**, i.e. a symmetry transformation that changes the *value* of the field ϕ but does not involve the coordinates.

Consider a complex-valued scalar field ψ with the Lagrangian

$$L[\psi] = \frac{1}{2}\left(g^{\mu\nu}\partial_\mu\psi\partial_\nu\psi^* - V(|\psi|)\right)\sqrt{-g}.$$

This Lagrangian is manifestly invariant under the group of transformations $\psi \to \tilde{\psi} = e^{\mathrm{i}\alpha}\psi$, where α is an arbitrary real number; this is an example of an internal symmetry of the theory. The identity $L[\tilde{\psi}] = L[\psi]$ leads to a conservation law. Again, it is convenient to apply an infinitesimal transformation,

$$\psi \to \tilde{\psi} = \psi\,(1 + \mathrm{i}\alpha) + O(\alpha^2),$$

and to suppress terms of order α^2 or higher. The local variation of the field ψ is

$$\delta\psi(\mathbf{x}) = \mathrm{i}\alpha\psi(\mathbf{x}), \quad \delta\psi^* = -\mathrm{i}\alpha\psi^*,$$

hence the local variation of the Lagrangian is

$$\begin{aligned} 0 = \delta L[\psi(\mathbf{x}),\mathbf{x}] &= L[\tilde{\psi}(\mathbf{x}),\mathbf{x}] - L[\psi(\mathbf{x}),\mathbf{x}] \\ &= \frac{\partial L}{\partial\psi}\delta\psi + \frac{\partial L}{\partial\psi^*}\delta\psi^* + \frac{\partial L}{\partial\psi_{,\mu}}\delta\psi_{,\mu} + \frac{\partial L}{\partial\psi^*_{,\mu}}\delta\psi^*_{,\mu} \\ &= \left(\frac{\partial L}{\partial\psi_{,\mu}}\delta\psi + \frac{\partial L}{\partial\psi^*_{,\mu}}\delta\psi^*\right)_{,\mu} = \left(\frac{\partial L}{\partial\psi^*_{,\mu}}\psi - \frac{\partial L}{\partial\psi_{,\mu}}\psi^*\right)_{,\mu}\mathrm{i}\alpha, \end{aligned}$$

where we have used the Euler-Lagrange equations. For the above Lagrangian, the Noether current is

$$j^\mu = g^{\mu\nu}\frac{\psi_{,\nu}\psi^* - \psi^*_{,\nu}\psi}{2i},$$

and the Noether charge is interpreted as the charge density of the field ψ.

Infinite-dimensional (gauge) symmetry

The Noether theorem applies to a more general case: namely, an n-dimensional Lie group of transformations that changes both the points $\mathbf{x}$ and the values of fields ϕ_a, $a = 1,\ldots,A$, according to

$$\mathbf{x} \to \tilde{\mathbf{x}} = \mathbf{x} + \varepsilon\mathbf{f}(\mathbf{x}), \quad \phi_a(\mathbf{x}) \to \tilde{\phi}_a(\mathbf{x}) = \phi_a(\tilde{\mathbf{x}}) + \varepsilon q_a(\tilde{\mathbf{x}}),$$

and adds a total divergence to the Lagrangian L,

$$\int \mathrm{d}^4\mathbf{x}\, L[\tilde{\phi}_a(\mathbf{x}),\mathbf{x}] = \int \mathrm{d}^4\tilde{\mathbf{x}}\, L[\phi_a(\tilde{\mathbf{x}}),\tilde{\mathbf{x}}] + \int \mathrm{d}^4\mathbf{x}\left(\partial_\mu C^\mu\right),$$

where C^μ is a suitable auxiliary vector field. A corresponding Noether current j_s^μ exists also in this case, indicating a conserved quantity Q that remains constant by virtue of the equations of motion.

Remark: The conservation of Q can be viewed as simply a consequence of the equations of motion, and derived by suitable algebraic manipulations from the Euler-Lagrange equations. However, in that case one may wonder what other conservation laws may be found and how the necessary manipulations are to be guessed. On the other hand, the Noether theorem explains that the existence of conserved quantities is necessary, as long as the Lagrangian has a continuous symmetry. (And, conversely, every conservation law is a consequence of a symmetry.) Thus, symmetries and the Noether theorem provide a more natural way to obtain conservation laws in a given field theory.

A qualitatively different situation arises if the theory is invariant under symmetry transformations parametrized by an *arbitrary function* $\alpha(\mathbf{x})$ of the spacetime. (This makes the group of transformations an *infinite*-dimensional group.) Such symmetry transformation is called a **gauge symmetry**. A familiar example of a gauge symmetry is the Maxwell theory with the Lagrangian

$$L_{EM} = -\frac{1}{16\pi}F_{\mu\nu}F^{\mu\nu}, \quad F_{\mu\nu} \equiv \partial_\mu A_\nu - \partial_\nu A_\mu,$$

which is invariant under the gauge transformation

$$A_\mu(\mathbf{x}) \to \tilde{A}_\mu(\mathbf{x}) = A_\mu(\mathbf{x}) + \partial_\mu \alpha(\mathbf{x}),$$

where $\alpha(\mathbf{x})$ is an arbitrary function.

We shall now examine the consequences of a gauge symmetry. Suppose that a field theory is described by a Lagrangian $L[\phi]$ that is invariant under the (infinitesimal) internal symmetry transformations

$$\phi \to \tilde{\phi} = \phi + \delta\phi(\phi;\alpha),$$

where $\delta\phi(\phi;\alpha)$ is linear in α, and α is an arbitrary function of $\mathbf{x}$. For simplicity, let us assume that $\delta\phi(\phi,\alpha)$ depends only on α and the first derivatives $\alpha_{,\mu}$, so that

$$\delta\phi(\phi,\alpha) = p(\phi)\alpha + q^\mu(\phi)\alpha_{,\mu},$$

where $p(\phi)$ and $q^\mu(\phi)$ are some known coefficients. Then the local variation of the Lagrangian is

$$\begin{aligned} 0 = \delta L[\phi(\mathbf{x}),\mathbf{x}] &= \frac{\partial L}{\partial\phi}\delta\phi + \frac{\partial L}{\partial\phi_{,\mu}}\delta\phi_{,\mu} \\ &= \frac{\partial L}{\partial\phi}p\alpha + \frac{\partial L}{\partial\phi}q^\mu\alpha_{,\mu} + \frac{\partial L}{\partial\phi_{,\mu}}(p\alpha)_{,\mu} + \frac{\partial L}{\partial\phi_{,\mu}}(q^\nu\alpha_{,\nu})_{,\mu} \\ &= \left(\frac{\partial L}{\partial\phi}p + \frac{\partial L}{\partial\phi_{,\mu}}p_{,\mu}\right)\alpha + \left(\frac{\partial L}{\partial\phi}q^\mu + \frac{\partial L}{\partial\phi_{,\nu}}\left(p\delta^\mu_\nu + q^\mu{}_{,\nu}\right)\right)\alpha_{,\mu} \\ &\quad + \frac{\partial L}{\partial\phi_{,\mu}}q^\nu\alpha_{,\mu\nu}. \end{aligned}$$

Since $\alpha(\mathbf{x})$ is an arbitrary function, the terms involving α, $\nabla\alpha$, $\nabla\nabla\alpha$, etc., must separately vanish. Thus we obtain the following three identities,

$$\begin{aligned} \frac{\partial L}{\partial\phi}p + \frac{\partial L}{\partial\phi_{,\mu}}p_{,\mu} &= 0, \\ \frac{\partial L}{\partial\phi}q^\mu + \frac{\partial L}{\partial\phi_{,\nu}}\left(p\delta^\mu_\nu + q^\mu{}_{,\nu}\right) &= 0, \\ \frac{\partial L}{\partial\phi_{,\mu}}q^\nu + \frac{\partial L}{\partial\phi_{,\nu}}q^\mu &= 0. \end{aligned}$$

(The last line follows because $\alpha_{,\mu\nu}$ is an arbitrary *symmetric* tensor.) Note that we did not assume that the field $\phi(\mathbf{x})$ satisfies the Euler-Lagrange equation. Hence, the above identities hold simply because of the mathematical properties of the Lagrangian and the fields.

As an illustration, let us evaluate the three identities in the Maxwell theory. The field ϕ is now the 1-form A_μ, and we have

$$\begin{aligned} &\frac{\partial L_{EM}}{\partial A_\mu} = 0, \quad \frac{\partial L_{EM}}{\partial A_{\mu,\nu}} = -\frac{1}{4\pi}F^{\mu\nu}, \\ &\delta A_\mu = \alpha_{,\mu} = \delta^\nu_\mu\alpha_{,\nu} \equiv q^\nu_\mu\alpha_{,\nu}; \quad q^\nu_{\mu;\rho} = 0; \quad p = 0. \end{aligned}$$

Of the three identities, the first two yield $0 = 0$ and the last one is

$$\frac{\partial L}{\partial A_{\lambda,\mu}}q^\nu_\lambda + \frac{\partial L}{\partial A_{\lambda,\nu}}q^\mu_\lambda = F^{\nu\mu} + F^{\mu\nu} = 0.$$

This is a simple mathematical consequence of the definition of $F^{\mu\nu}$.

The invariance of a field theory under gauge transformations $\phi \to \tilde{\phi}(\phi;\alpha)$ does not lead to useful conservation laws but has another important consequence for the theory. Consider a Cauchy problem for the field ϕ, which consists of specifying the initial values of the field $\phi_{in}(\mathbf{x})$ at an initial spacelike 3-surface Σ. Suppose that $\phi(\mathbf{x})$ is a solution of the Euler-Lagrange equation for the initial data ϕ_{in} on Σ. The transformation $\phi \to \tilde{\phi}$ is a symmetry of the equations of motion, so $\tilde{\phi}(\phi;\alpha)$ is another solution, for an arbitrary function $\alpha(\mathbf{x})$. The function $\alpha(\mathbf{x})$ can be chosen so that $\alpha = 0$ on Σ but $\alpha \neq 0$ to the future of Σ. The solution $\tilde{\phi}(\phi(\mathbf{x});\alpha)$ will then have the same initial data ϕ_{in} on Σ but will differ from $\phi(\mathbf{x})$ to the future of Σ. Thus, the solution of the Cauchy problem is not unique.

In fact, this lack of uniqueness was a problem initially encountered by A. Einstein who came up with the following "hole argument" when trying to derive the equations for the metric g. Suppose $g(\mathbf{x})$ is a solution for the metric in the spacetime, subject to some boundary conditions. If the boundary conditions are imposed on some 3-surface Σ that encircles a domain of spacetime, we can find a small region (a "hole") inside the domain. Due to the general covariance of the theory, we are free to transform the coordinates, $\mathbf{x} \to \tilde{\mathbf{x}}$, in such a way that the coordinates are changed only within the hole but remain the same everywhere else. This would change the metric $g(\mathbf{x})$ inside the hole but not elsewhere. The new metric $\tilde{g}$ still satisfies the same boundary conditions as $\tilde{g}$. Thus, the metric at any given point $\mathbf{x}$ is not uniquely specified by the equations of motion and boundary conditions.

The resolution of this paradox is that coordinates x^μ are arbitrary labels assigned to events in spacetime. A gauge transformation $\mathbf{x} \to \tilde{\mathbf{x}}$ merely relabels the events but does not influence the observable effects of gravitation (e.g. the distance between some physical events). Thus, solutions g and $\tilde{g}$ are *physically equivalent*. In any theory with a gauge symmetry, two solutions that differ by a gauge transformation are physically equivalent.

Moreover, the freedom to choose an arbitrary function $\alpha(\mathbf{x})$ in a gauge transformation means that any one of the fields ϕ_j may be set to zero (or to any other function). This will reduce the number of unknown functions to solve for. Such a reduction is called **gauge fixing**. For example, the gauge symmetry $A_\mu \to A_\mu + \partial_\mu\alpha$ in electrodynamics allows one to fix $A_0 = 0$ and to consider the remaining three components A_1, A_2, A_3 as the only physically relevant fields. However, the Maxwell equations in terms of A_1, A_2, A_3 appear much less symmetric and more difficult to solve. Alternatively, one may choose the **Lorentz gauge** condition $A_\mu{}^{,\mu} = 0$, or other suitable condition.

To conclude: The presence of a gauge symmetry indicates that the description of the theory contains "too many" fields. The freedom can be eliminated by fixing the gauge. However, a description in a fixed gauge may be much less convenient than the original description.

5.2 Hamiltonian formulation

Literature sources: [28], chapter 4; [36], Appendix.

The purpose of this section is to introduce the Hamiltonian formulation of general relativity.

The starting point of a classical theory is the Lagrangian $L(q_i, \dot{q}_i)$, where $q_i(t)$, $i = 1, ..., N_q$, are the generalized coor-

dinates, and the action functional is

$$S[q_i(t)] = \int L(q_i, \dot{q}_i)\mathrm{d}t.$$

A Hamiltonian formulation of the theory is then obtained through the following steps:

- Define the **canonical momenta** $p_i \equiv \partial L/\partial \dot{q}_i$.
- Express $\dot{q}_i$ through p_i from these equations.
- Define the Hamiltonian $H(p_i, q_i) \equiv \sum_i p_i \dot{q}_i - L(q_i, \dot{q}_i)$, where $\dot{q}_i$ are expressed as functions of p_i in the right-hand side.
- Use the **Hamilton equations** of motion,

$$\frac{d}{dt}q_i = \frac{\partial H}{\partial p_i}, \quad \frac{d}{dt}p_i = -\frac{\partial H}{\partial q_i}.$$

  These equations can be also considered as Euler-Lagrange equations following from the Hamiltonian action principle,

$$\delta \int_{t_1}^{t_2} \mathrm{d}t \left(\sum_i p_i \dot{q}_i - H(p, q) \right) = 0, \tag{5.30}$$

  where $p_i(t)$ and $q_i(t)$ are varied *independently*, subject only to the boundary conditions $\delta q_i(t_{1,2}) = 0$.

In field theory, the role of the generalized coordinates $q_i(t)$ is played by the fields $\phi_j(t, x, y, z)$, $j = 1, ..., N_\phi$. The index i in q_i now becomes a "condensed index" $i \equiv \{j, x, y, z\}$ in $\phi_j(t, x, y, z)$. Therefore, in the Hamiltonian formalism the dependence on time t must be separated from the dependence on spatial coordinates $(x, y, z) \equiv x_a$. (We shall use Latin indices for three-dimensional components.) In a curved spacetime, there is no natural time variable, which means that we need to *arbitrarily select* a coordinate t and the corresponding family of spacelike 3-surfaces $t = \text{const}$. Such a family of nonintersecting spacelike 3-surfaces, covering the entire spacetime, is called a **time foliation** of the spacetime. The traditional Hamiltonian approach is *noncovariant*: it requires an explicit time foliation of the spacetime and treats the time dependence of every variable differently from the spacial dependence.

The Lagrangian $L(q_i, \dot{q}_i)$ is replaced by a Lagrangian $\mathcal{L}$, which is an integral over a 3-surface,

$$\mathcal{L}[\phi_i, \dot{\phi}_i] = \int_{t=\text{const}} \mathrm{d}^3 x_a \, L(\phi_i, \partial_a \phi_i, \dot{\phi}_i),$$

where we separated the time derivatives $\dot{\phi}_i$ from the spatial derivatives $\partial_a \phi_i$. The Lagrangian $\mathcal{L}$ is now a *functional* and is denoted by a script letter, in distinction from the Lagrangian density L which is a *function*.

The canonical momentum $p_i(t, x_a)$ is defined as the functional derivative of $\mathcal{L}$ with respect to $\dot{\phi}_i(t, x_a)$,

$$p_i(t, x_a) = \frac{\delta \mathcal{L}[\phi_i, \dot{\phi}_i]}{\delta \dot{\phi}_i(t, x_a)}.$$

Finally, the summation over i in the definition of the Hamiltonian must be replaced by a summation over j and an integration over the 3-surface,

$$\mathcal{H}[p_i, \phi_i] = \int_{t=\text{const}} \sum_j p_j(t, x_a)\dot{\phi}_j(t, x_a)\mathrm{d}^3 x_a - \mathcal{L}[\phi_i, \dot{\phi}_i].$$

We shall now derive Hamiltonian formulations for the Maxwell theory where the "field ϕ" is the 4-potential $A_\mu(x)$, and for Einstein's General Relativity where the "field" is the metric $g_{\mu\nu}(x)$.

5.2.1 Electrodynamics in Hamiltonian formulation

The Lagrangian (5.25) of the Maxwell theory is

$$-\frac{\sqrt{-g}}{16\pi} F_{\mu\nu}F^{\mu\nu} = \frac{\sqrt{-g}}{8\pi}\left(|\vec{E}|^2 - |\vec{B}|^2\right),$$

which (in flat space) is the familiar expression for the pressure of the electromagnetic field.[3] The spacetime metric $g_{\mu\nu}$ is now considered as a known function, so only the electromagnetic potential A_μ is to be determined. (To express this, one says that $g_{\mu\nu}$ is a **background field**.) To compute the canonical momenta corresponding to A_μ, we need to separate the dependence of the Lagrangian on the *time* derivatives of the field A_μ.

Suppose a time foliation of the spacetime is fixed and some coordinates x_a, $a = 1, 2, 3$ are chosen on equal-time surfaces $t = \text{const}$. Let $\mathbf{n}$ be a normalized, timelike vector field orthogonal to the equal-time surfaces. For each point p, the subspace $\mathbf{n}^\perp(p) \subset T_p\mathcal{M}$ is the tangent space to the equal-time surface intersecting the point p. Then we may decompose the vector field A^μ as

$$\mathbf{A} = \mathbf{n}A_0 + P\mathbf{A},$$

where $A_0 \equiv g(\mathbf{n}, \mathbf{A})$ and $P\mathbf{x} = \mathbf{x} - \mathbf{n}g(\mathbf{n}, \mathbf{x})$ is the orthogonal projector onto $\mathbf{n}^\perp$. The 4-vector $P\mathbf{A}$ is spacelike and tangent to the equal-time surfaces. Hence, in local coordinates $\{x_a\}$ on equal-time surfaces, the vector $P\mathbf{A}$ is described by a 3-vector with components $\{A_1, A_2, A_3\} \equiv \vec{A}$. Similarly, we can define the 3-vectors $\vec{E}$ and $\vec{B}$. Time derivatives appear in the electric field $\vec{E}$,

$$\vec{E} \equiv E_a = \partial_t A_a - \partial_a A_0 \equiv \partial_t \vec{A} - \text{grad } A_0,$$

while the magnetic field

$$\vec{B} \equiv B_a = \sum_{b,c=1}^{3} \varepsilon_{abc}\partial_b A_c = \text{rot } \vec{A}$$

depends only on spatial derivatives. It is clear that the Lagrangian does not contain the time derivative of A_0. Therefore, we cannot define a canonical momentum p_0 for A_0, and there are no Hamilton equations of motion for the pair $\{A_0, p_0\}$; in this case, one says that A_0 is *not a dynamical variable*. This conclusion agrees with the existence of a gauge symmetry in electrodynamics. Namely, by performing a gauge transformation $A_\mu \to A_\mu + \partial_\mu \alpha$ one could set A_0 to an arbitrary fixed function (this is called "fixing the gauge"). For simplicity, let us set $A_0 = 0$. Note that the variation of the Lagrangian with respect to A_0 gives

$$\sum_{a=1}^{3} \partial_a \left(\sqrt{-g}E_a\right) \equiv \sqrt{-g}\,\text{div } \vec{E} = 0, \tag{5.31}$$

[3] Several field theories, e.g. the scalar field and the electromagnetic field, have the energy-momentum tensor of a **perfect fluid** form, $T_{\mu\nu} = \rho u_\mu u_\nu - (p + \rho) g_{\mu\nu}$. In that case, the Lagrangian density $\mathcal{L}$ is equal to the pressure p and the Hamiltonian density $\mathcal{H}$ to the energy density ρ.

which does not contain time derivatives and is therefore a **constraint** of the theory.

There might be a sign error here: $p = -E$? See Wald [36], p. 461-462.

The canonical momenta p_a ($a = 1, 2, 3$) for the spatial components of A_μ are easily found,

$$p_a(x) = -\frac{\sqrt{-g}}{4\pi} E_a = \frac{\sqrt{-g}}{4\pi} \dot{A}_a.$$

The time derivatives $\dot{A}_a$ are expressed through the momenta as

$$\dot{A}_a = \frac{4\pi}{\sqrt{-g}} p_a,$$

hence the Hamiltonian is

$$\begin{aligned} \mathcal{H}[p_a, A_a, A_0] &= \int_{t=\text{const}} d^3x_a \left(\frac{2\pi}{\sqrt{-g}} \sum_{b=1}^{3} p_b^2(x) + \frac{\sqrt{-g}}{8\pi} \left|\vec{B}\right|^2 \right) \\ &= \int_{t=\text{const}} d^3x_a \frac{\sqrt{-g}}{8\pi} \left(|\vec{E}|^2 - |\vec{B}|^2 \right), \end{aligned}$$

which (in flat space) is the familiar expression for the total energy of the electromagnetic field at a fixed time.

The Hamilton equations of motion are

$$\begin{aligned} \frac{\partial A_a}{\partial t} &= \frac{\delta \mathcal{H}}{\delta p_a} = \frac{4\pi}{\sqrt{-g}} p_a, \\ \frac{\partial p_a}{\partial t} &= -\frac{\delta \mathcal{H}}{\delta A_a} = \sum_{b,c} \partial_c (\sqrt{-g} \varepsilon_{abc} B_b) \equiv -\sqrt{-g}\, \text{rot}\, \vec{B}. \end{aligned}$$

It is straightforward to see that the above equations, together with the constraint (5.31), are equivalent to the Maxwell equations in vacuum. The constraint is merely a restriction on possible initial conditions: Once the constraint holds at an initial time, it will hold at any future time since $\partial_t \text{div}\, \vec{E} = -\text{div}\left(\text{rot}\, \vec{B}\right) = 0$.

Remark: The gauge fixing condition $A_0 = 0$ does not entirely remove the gauge freedom because one can still perform a transformation $A_a \to A_a + \partial_a \alpha$ with a function $\alpha(x_a)$ depending only on spatial coordinates. An additional gauge fixing condition, such as

$$\sum_a \partial_a \left(\sqrt{-g} A_a\right) = \text{div}\, \vec{A} = 0,$$

will remove the remaining gauge freedom. Note that the constraint $\text{div}\, \vec{E} = 0$ will be an automatic consequence of this gauge condition. The resulting gauge is called the **radiation gauge** because $\text{div}\, \vec{E} = 0$ is the Maxwell equation in vacuum, describing the propagation of pure electromagnetic field, and $A_0 = 0$ means, heuristically, the absence of an electrostatic field component.

Since we have fixed the gauge, we had to add the constraint (5.31) to the Hamilton equations of motion. It is possible instead to keep the nondynamical field A_0 in the Hamiltonian (without introducing the momentum p_0). The constraint will then be obtained as the equation $\delta \mathcal{H} / \delta A_0 = 0$. To get a better idea of the Hamiltonian treatment of constrained systems, in the next subsection we shall consider some simple examples from classical mechanics.

5.2.2 Hamiltonian mechanics of constrained systems

See also [18, 14] for a more detailed and wide-ranging developments.

The Hamiltonian formalism involves replacing velocities $\dot{q}_i$ by momenta p_i through the relations $p_i = \delta L / \delta \dot{q}_i$. However, in many cases these relations cannot be solved for some $\dot{q}_i$, and then the standard rule for finding the Hamiltonian cannot be used. For instance, the Lagrangian for electrodynamics is independent of $\dot{A}_0$ and thus the relation $p_0 = \delta L / \delta \dot{A}_0$ cannot be solved for A_0. Another example is when L is *linear* in, say, the velocity $\dot{q}_1$; in this case, $\delta L / \delta \dot{q}_1$ is independent of $\dot{q}_1$ and again the same problem arises: namely, we cannot express $\dot{q}_1$ through the momentum p_1. In these cases, one can use a slightly different but equivalent approach to develop a Hamiltonian formulation. This approach is known as the **Faddeev-Jackiw formalism**.

The main idea of the Faddeev-Jackiw approach is to recognize that the Hamilton equations of motion are first-order in time derivatives and follow, as Euler-Lagrange equations, from the action (5.30) that is *linear* in all the velocities $\dot{q}_i$. Given the Lagrangian (5.30), we do not attempt to introduce new canonical momenta for q_i, because we recognize that these momenta are already present in the Lagrangian (they are the variables p_i). Thus, the transition from a Lagrangian formulation to a Hamiltonian formulation can be seen as the replacement of a Lagrangian action $L(q, \dot{q})$, which is generally nonlinear in the velocities $\dot{q}_i$, by *another Lagrangian*, namely $\tilde{L} \equiv \sum_i p_i \dot{q}_i - H(p, q)$, which is *linear* in the velocities and thus yields first-order Euler-Lagrange equations of motion. The replacement $L \to \tilde{L}$ comes at the cost of extending the Lagrangian by introducing extra momentum variables p_i. Therefore, we need to add only as many momentum variables as needed for the Lagrangian to become linear in all the velocities. If we notice that the Lagrangian is linear in velocities, we conclude that all the necessary momentum variables are already introduced, and the Hamilton equations will be found as the ordinary Euler-Lagrange equations for an ordinary (Lagrangian) system. All these Euler-Lagrange equations will be (at most) first-order in time derivatives, as expected for Hamilton equations. After deriving these equations, one can easily determine whether constraints are present in the system. Namely, the constraints will be any equations not containing time derivatives.

As a first example, consider the Lagrangian

$$L_0(x, \dot{x}, y, \dot{y}, z, \dot{z}) = \dot{x} y + \frac{1}{2} m \dot{z}^2 - \frac{1}{2m} y^2 - V(x, z).$$

This Lagrangian is linear in $\dot{x}$, independent of $\dot{y}$, and quadratic in $\dot{z}$. We would like to obtain a Hamiltonian formulation for this system. Since the Lagrangian is nonlinear in $\dot{z}$, we introduce the canonical momentum for the variable z,

$$p_z \equiv \frac{\partial L_0}{\partial \dot{z}} = m \dot{z},$$

express $\dot{z} = \frac{1}{m} p_z$, and build the "temporary Hamiltonian"

$$H_0 \equiv p_z \dot{z} - L_0|_{\dot{z} = p_z/m} = -\dot{x} y + \frac{1}{2m} p_z^2 + \frac{1}{2m} y^2 + V(x, z).$$

Now we formulate the "temporary Hamiltonian action prin-

ciple" as the variation of the new, extended Lagrangian,

$$\begin{aligned}\tilde{L}_0(x,\dot{x},y,z,\dot{z},p_z) &\equiv p_z\dot{z} - H_0 \\ &= p_z\dot{z} + \dot{x}y - \frac{1}{2m}p_z^2 - \frac{1}{2m}y^2 - V(x,z).\end{aligned}$$

The Lagrangian $\tilde{L}_0$ is linear in all the velocities. Therefore, we do not need to introduce any more "momentum" variables; instead, we simply need to relabel some of the existing variables as "momenta." Presently, it is clear that y plays the role of the "momentum" for x. Therefore, relabeling $p_x \equiv y$, we obtain the Hamiltonian action,

$$\tilde{L}_0(x,\dot{x},p_x,z,\dot{z},p_z) = p_x\dot{x} + p_z\dot{z} - \frac{1}{2m}\left(p_x^2 + p_z^2\right) - V(x,z),$$

which describes a particle of mass m in a two-dimensional potential $V(x,z)$. The Euler-Lagrange equations of motion are

$$\begin{aligned}\dot{x} &= \frac{1}{m}p_x, \quad \dot{p}_x = -\partial V/\partial x, \\ \dot{z} &= \frac{1}{m}p_z, \quad \dot{p}_z = -\partial V/\partial z.\end{aligned}$$

Clearly, there are no constraints in this model. We conclude that the Lagrangian L_0 is actually *not constrained*, despite the fact that we could not introduce the momenta p_x and p_y in the conventional manner. The reason for the difficulty was purely technical: the initial Lagrangian L_0 was already "halfway done" becoming a Hamiltonian. However, the unusual Lagrangian L_0 is completely equivalent to the more conventional $\tilde{L}_0$.

Some systems, however, *are* constrained. As a further example, consider the Lagrangian

$$L_1(q_1,\dot{q}_1,q_2,\dot{q}_2) = \frac{1}{2}(\dot{q}_1 - q_2)^2.$$

This Lagrangian is quadratic in $\dot{q}_1$, so we introduce the momentum

$$p_1 \equiv \frac{\partial L_1}{\partial \dot{q}_1} = \dot{q}_1 - q_2,$$

solve for $\dot{q}_1 = p_1 + q_2$, and obtain the "temporary Hamiltonian"

$$H_1(q_1,p_1,q_2,p_2) = p_1\dot{q}_1 - L_1|_{\dot{q}_1 = p_1 + q_2} = \frac{1}{2}p_1^2 + p_1 q_2$$

and the extended Lagrangian

$$\tilde{L}_1(q_1,\dot{q}_1,q_2,\dot{q}_2,p_1) = p_1\dot{q}_1 - H_1 = p_1\dot{q}_1 - \frac{1}{2}p_1^2 - p_1 q_2.$$

Since $\tilde{L}_1$ is linear in all the velocities, we now find the Hamilton equations as the Euler-Lagrange equations for $\tilde{L}_1$:

$$\dot{q}_1 - p_1 - q_2 = 0, \quad p_1 = 0, \quad \dot{p}_1 = 0.$$

Clearly, we have one constraint, $p_1 = 0$, and one dynamical equation, $\dot{q}_1 = q_2$.

Practice problem: The system with the Lagrangian L_1 is invariant under the gauge symmetry

$$q_1 \to q_1 + \alpha(t), \quad q_2 \to q_2 + \partial_t\alpha(t),$$

where $\alpha(t)$ is an arbitrary function. Determine the identity that this symmetry generates via Noether's theorem.

Practice problem: Derive the Hamilton equations of motion for the following Lagrangians:

$$L_3 = \frac{\dot{x}^2}{2} + xy\dot{z}; \quad L_4 = \frac{\dot{x}^2}{2} + \frac{(x+y)^2}{2}; \quad L_5 = \dot{x}y + \frac{\dot{y}^2}{2} + xy.$$

Determine which of these Lagrangians are constrained.

Of course, after determining the Hamilton equations and the constraints, one needs to solve these equations. Sometimes it is easy to "solve the constraints," i.e. to express some canonical variables as functions of others, and thus to reduce the number of dynamical degrees of freedom. In other cases, solving the constraints is a nontrivial task, and instead one attempts to solve the dynamical equations and impose the constraints afterwards. A full consideration of this topic is beyond the scope of these lectures; there exists an extensive literature about constrained systems, both classical and quantized.

5.2.3 Gauss-Codazzi equation

In the Hamiltonian formalism, time derivatives play a completely different role from spatial derivatives. To compute the canonical momenta in general relativity, we will need to separate the spatial and the temporal derivatives in the Einstein-Hilbert Lagrangian $\sqrt{-g}R$. Under "spatial derivatives" it is natural to understand the intrinsic covariant derivatives within the 3-surfaces $t = \text{const}$. (We know from Sec. 1.10.1-1.10.2 that a 3-surface embedded in a manifold receives an induced metric and an induced Levi-Civita connection.) Therefore, the first step toward a Hamiltonian formulation of GR is to express the curvature scalar R as a sum of a term depending only on spatial derivatives of the metric and remaining terms.

Consider a 3-surface Σ embedded into a spacetime manifold $\mathcal{M}$ with a metric g. The surface Σ is spacelike if the normal vector is timelike. The surface Σ is **timelike** if the tangent space to Σ at each point, $T\Sigma(p)$, contains a timelike vector (and thus is spanned by one timelike and two spacelike directions). A timelike 3-surface can be considered as a "sub-spacetime" where two-dimensional creatures might live. (To emphasize this interpretation, a timelike 3-surface is frequently called "2+1-dimensional".) A surface Σ receives the induced metric $h = g|_{T\Sigma}$, so the two-dimensional creatures may define the induced Levi-Civita connection ${}^{(3)}\nabla$ and compute the corresponding (3-dimensional) Riemann tensor ${}^{(3)}R(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d})$ for tangent vectors $\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d}$. There is a relationship between the reduced curvature ${}^{(3)}R$ and the curvature of the full 4-dimensional spacetime, which we denote ${}^{(4)}R$ for clarity. This relationship is called the **Gauss-Codazzi equation**, which we shall now derive.

Let $\mathbf{n}$ be a (spacelike) vector field which is everywhere normal to the surface Σ and normalized, $g(\mathbf{n},\mathbf{n}) = -1$. (The case of a timelike $\mathbf{n}$ is fully analogous.) We know from Sec. 1.10.1 and 1.10.2 that the induced Levi-Civita connection ${}^{(3)}\nabla$ is found using the projector $P = \hat{1} - \mathbf{n}\otimes\hat{g}\mathbf{n}$ onto the tangent bundle $T\Sigma$,

$${}^{(3)}\nabla_{\mathbf{x}}\mathbf{t} = \nabla_{\mathbf{x}}\mathbf{t} - \Gamma(\mathbf{x})\mathbf{t},$$

where we have defined Γ as a transformation-valued 1-form,

$$\Gamma(\mathbf{x})\mathbf{t} \equiv \mathbf{n}g(\mathbf{t},\nabla_{\mathbf{x}}\mathbf{n}). \tag{5.32}$$

Although the calculations in Sec. 1.10.1-1.10.3 were performed for the case of flat spacetime $\mathcal{M}$, the assumption ${}^{(4)}R = 0$ can

be easily dropped. In the derivation leading to Eq. (1.88), we only need to retain the underlined term $\left[\partial_{\mathbf{x}}, \partial_{\mathbf{y}}\right] \mathbf{z} - \partial_{[\mathbf{x},\mathbf{y}]}\mathbf{z}$ and to replace it by ${}^{(4)}R(\mathbf{x},\mathbf{y})\mathbf{z}$. This yields

$$\begin{aligned}{}^{(3)}R(\mathbf{x},\mathbf{y})\mathbf{z} =& {}^{(4)}R(\mathbf{x},\mathbf{y})\mathbf{z} + [\Gamma(\mathbf{x}),\Gamma(\mathbf{y})]\,\mathbf{z}\\ &+ (\nabla_{\mathbf{x}}\Gamma)(\mathbf{y})\mathbf{z} - (\nabla_{\mathbf{y}}\Gamma)(\mathbf{x})\mathbf{z}.\end{aligned}$$

Substituting the definition (5.32) of Γ, we find (for tangent vectors $\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{t}$)

$$\begin{aligned}\Gamma(\mathbf{x})\Gamma(\mathbf{y})\mathbf{z} &= \mathbf{n}g(\Gamma(\mathbf{y})\mathbf{z},\nabla_{\mathbf{x}}\mathbf{n}) = \mathbf{n}g(\mathbf{n}(...),\nabla_{\mathbf{x}}\mathbf{n}) = 0,\\ (\nabla_{\mathbf{x}}\Gamma)(\mathbf{y})\mathbf{z} &= (\nabla_{\mathbf{x}}\mathbf{n})g(\mathbf{z},\nabla_{\mathbf{y}}\mathbf{n}) + \mathbf{n}(...),\\ g((\nabla_{\mathbf{x}}\Gamma)(\mathbf{y})\mathbf{z},\mathbf{t}) &= g(\mathbf{t},\nabla_{\mathbf{x}}\mathbf{n})g(\mathbf{z},\nabla_{\mathbf{y}}\mathbf{n}),\end{aligned}$$

where we denoted by $(...)$ some uninteresting terms that cancel under the scalar product. Hence,

$$\begin{aligned}{}^{(3)}R(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{t}) =& {}^{(4)}R(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{t})\\ &+ g(\mathbf{t},\nabla_{\mathbf{x}}\mathbf{n})g(\mathbf{z},\nabla_{\mathbf{y}}\mathbf{n}) - g(\mathbf{t},\nabla_{\mathbf{y}}\mathbf{n})g(\mathbf{z},\nabla_{\mathbf{x}}\mathbf{n}).\end{aligned}$$

The above expression contains the bilinear form

$$B_{(\mathbf{n})}(\mathbf{x},\mathbf{y}) \equiv g(\nabla_{\mathbf{x}}\mathbf{n},\mathbf{y}),$$

which is the **distortion tensor** of the field $\mathbf{n}$ (see Sec. 2.3.1). Since the vector field $\mathbf{n}$ is, by construction, orthogonal to the 3-surface Σ, we may extend $\mathbf{n}$ to a normalized geodesic vector field in a neighborhood of Σ, and thus the tensor $B_{(\mathbf{n})}$ is symmetric and transverse to $\mathbf{n}$. Effectively, $B_{(\mathbf{n})}$ is a 3-dimensional symmetric tensor in the tangent bundle $T\Sigma$. Therefore, the metric g in the definition of $B_{(\mathbf{n})}$ can be replaced by the partial metric

$$\begin{aligned}B_{(\mathbf{n})}(\mathbf{x},\mathbf{y}) &= h(\nabla_{\mathbf{x}}\mathbf{n},\mathbf{y}),\\ h(\mathbf{x},\mathbf{y}) &\equiv g(\mathbf{x},\mathbf{y}) + g(\mathbf{x},\mathbf{n})g(\mathbf{y},\mathbf{n}).\end{aligned}$$

In this context, the distortion tensor $B_{(\mathbf{n})}$ is called the **extrinsic curvature** tensor of the surface Σ and denoted K_Σ,

$$K_\Sigma(\mathbf{x},\mathbf{y}) = h(\nabla_{\mathbf{x}}\mathbf{n},\mathbf{y}); \quad K_{\mu\nu} = h_{\alpha\nu}n^{\alpha}_{;\mu}.$$

The extrinsic curvature K_Σ depends only on the derivatives of $\mathbf{n}$ in tangential directions and is thus independent of the completion of $\mathbf{n}$ to a vector field outside Σ.

Note that the same tensor K_Σ can be defined by using an arbitrary, unnormalized vector $\tilde{\mathbf{n}}$, in the following way:

$$\begin{aligned}K_\Sigma(\mathbf{x},\mathbf{y}) &= \tilde{h}(\nabla_{\mathbf{x}}\mathbf{n},\mathbf{y}),\\ \tilde{h}(\mathbf{x},\mathbf{y}) &\equiv g(\mathbf{x},\mathbf{y}) - \frac{1}{g(\tilde{\mathbf{n}},\tilde{\mathbf{n}})}g(\mathbf{x},\mathbf{n})g(\mathbf{y},\mathbf{n}).\end{aligned}$$

The resulting tensor is identical to $B_{(\mathbf{n})}$ for $\mathbf{n} = |g(\tilde{\mathbf{n}},\tilde{\mathbf{n}})|^{-1/2}\,\tilde{\mathbf{n}}$.

Using the extrinsic curvature K_Σ, the 3-dimensional Riemann tensor (defined only for *tangent* vectors $\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{t}$) can be rewritten as

$$\begin{aligned}{}^{(3)}R(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{t}) =& {}^{(4)}R(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{t}) + K_\Sigma(\mathbf{x},\mathbf{t})K_\Sigma(\mathbf{y},\mathbf{z})\\ &- K_\Sigma(\mathbf{y},\mathbf{t})K_\Sigma(\mathbf{x},\mathbf{z}).\end{aligned} \tag{5.33}$$

This is called the **Gauss-Codazzi equation**.

The induced connection ${}^{(3)}\nabla$ can be expressed through the extrinsic curvature K_Σ as

$${}^{(3)}\nabla_{\mathbf{x}}\mathbf{y} = \nabla_{\mathbf{x}}\mathbf{y} - \mathbf{n}g(\nabla_{\mathbf{x}}\mathbf{n},\mathbf{y}) = \nabla_{\mathbf{x}}\mathbf{y} - \mathbf{n}K_\Sigma(\mathbf{x},\mathbf{y}).$$

where $\mathbf{x},\mathbf{y}$ are tangent to Σ. In the index notation, the conventional way to denote induced derivatives is by a vertical bar,

$$\left({}^{(3)}\nabla_{\mathbf{x}}\mathbf{y}\right)^{\mu} \equiv x^{\nu}y^{\mu}{}_{|\nu}.$$

Thus we can write (for a tangent vector $\mathbf{y} \equiv y^\mu$)

$$y^{\mu}{}_{|\nu} = y^{\mu}{}_{;\nu} - n^{\mu}y^{\alpha}K_{\alpha\nu}.$$

Remark: The reason for calling K_Σ the "extrinsic" curvature is the following. The "intrinsic" curvature ${}^{(3)}R$ of the 3-surface Σ will differ from the full curvature ${}^{(4)}R$ of $\mathcal{M}$ if the 3-surface Σ is embedded into the spacetime $\mathcal{M}$ in a "curved" way. The tensor K_Σ contains the complete information needed to compute the intrinsic curvature on Σ if the full curvature ${}^{(4)}R$ is known. The covariant derivative ${}^{(3)}\nabla$ is also expressed through K_Σ and ${}^{(4)}\nabla$. Therefore, the tensor K_Σ completely characterizes the additional curvature on Σ due to the embedding of Σ into $\mathcal{M}$. Note, however, that not all the components of ${}^{(4)}R$ can be restored from the knowledge of three-dimensional tensors ${}^{(3)}R$ and K_Σ: one also needs some information about the 4-vector field $\mathbf{n}$ and its derivatives.

Statement: The relation (5.33) was derived only for tangent vectors; the tensor ${}^{(3)}R$ vanishes if any of its arguments is parallel to $\mathbf{n}$. It is possible, however, to express ${}^{(4)}R(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{n})$ in terms of K_Σ for tangent $\mathbf{x},\mathbf{y},\mathbf{z}$:

$${}^{(4)}R(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{n}) = \left({}^{(3)}\nabla_{\mathbf{y}}K_\Sigma\right)\circ(\mathbf{x},\mathbf{z}) - \left({}^{(3)}\nabla_{\mathbf{x}}K_\Sigma\right)\circ(\mathbf{y},\mathbf{z}). \tag{5.34}$$

We can derive Eq. (5.34), starting from Eq. (5.32).

Derivation: We start from

$$\begin{aligned}{}^{(3)}R(\mathbf{x},\mathbf{y})\mathbf{z} =& {}^{(4)}R(\mathbf{x},\mathbf{y})\mathbf{z} + [\Gamma(\mathbf{x}),\Gamma(\mathbf{y})]\,\mathbf{z}\\ &+ (\nabla_{\mathbf{x}}\Gamma)(\mathbf{y})\mathbf{z} - (\nabla_{\mathbf{y}}\Gamma)(\mathbf{x})\mathbf{z}.\end{aligned}$$

We write K instead of K_Σ for brevity. Substituting $\Gamma(\mathbf{x})\mathbf{y} = \mathbf{n}K(\mathbf{x},\mathbf{y})$ and using the fact that K is symmetric and transverse to $\mathbf{n}$, we find

$$\begin{aligned}\Gamma(\mathbf{x})\Gamma(\mathbf{y})\mathbf{z} &= \mathbf{n}K(\Gamma(\mathbf{y})\mathbf{z},\mathbf{x}) = \mathbf{n}K(\mathbf{n}(...),...) = 0,\\ (\nabla_{\mathbf{x}}\Gamma)(\mathbf{y})\mathbf{z} &= (\nabla_{\mathbf{x}}\mathbf{n})K(\mathbf{y},\mathbf{z}) + \mathbf{n}\,(\nabla_{\mathbf{x}}K)\circ(\mathbf{y},\mathbf{z}),\\ g((\nabla_{\mathbf{x}}\Gamma)(\mathbf{y})\mathbf{z},\mathbf{n}) &= (\nabla_{\mathbf{x}}K)\circ(\mathbf{y},\mathbf{z}).\end{aligned}$$

Since ${}^{(3)}R(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{n}) = 0$, we have

$${}^{(4)}R(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{n}) = \left(\nabla_{\mathbf{y}}K\right)\circ(\mathbf{x},\mathbf{z}) - (\nabla_{\mathbf{x}}K)\circ(\mathbf{y},\mathbf{z}).$$

The derivatives ∇K can be replaced by ${}^{(3)}\nabla K$ because K is transverse to $\mathbf{n}$:

$$\begin{aligned}K(\nabla_{\mathbf{x}}\mathbf{y},\mathbf{z}) &= K({}^{(3)}\nabla_{\mathbf{x}}\mathbf{y},\mathbf{z}) + K(\mathbf{n}(...),...) = K({}^{(3)}\nabla_{\mathbf{x}}\mathbf{y},\mathbf{z});\\ (\nabla_{\mathbf{x}}K)\circ(\mathbf{y},\mathbf{z}) &\equiv \mathcal{L}_{\mathbf{x}}\left(K(\mathbf{y},\mathbf{z})\right) - K(\nabla_{\mathbf{x}}\mathbf{y},\mathbf{z}) - K(\mathbf{y},\nabla_{\mathbf{x}}\mathbf{z})\\ &=^{(3)} \nabla_{\mathbf{x}}\left(K(\mathbf{y},\mathbf{z})\right) - K({}^{(3)}\nabla_{\mathbf{x}}\mathbf{y},\mathbf{z}) - K(\mathbf{y},{}^{(3)}\nabla_{\mathbf{x}}\mathbf{z})\\ &=^{(3)} (\nabla_{\mathbf{x}}K)\circ(\mathbf{y},\mathbf{z}).\end{aligned}$$

Thus we obtain the desired formula.

5.2.4 Boundary term in Einstein-Hilbert action

In Sec. 5.1.2 we derived the Einstein equation from a variational principle by imposing the condition that both δg and

$\nabla\delta g$ must vanish on a boundary surface $\partial\tilde{\mathcal{M}}$. This is different from the usual condition where only the variation of a field is constrained to vanish (but not its derivative). The formal reason for imposing the extra condition was the presence of a total divergence term, $\int_{\tilde{\mathcal{M}}}\sqrt{-g}d^4p\,\mathrm{div}\mathbf{q}$, where $\mathbf{q}$ is a vector field depending on the *derivatives* $\nabla\delta g$, as defined by Eq. (5.14). However, one can cancel this divergence term by subtracting an extra boundary term from the Einstein-Hilbert action, thus defining the "corrected" action $S_G[g]$,

$$S_G \equiv S_{EH} - \oint_{\partial\tilde{\mathcal{M}}} 2K\sqrt{h}\mathrm{d}^3p,$$

where K is a suitable function (the factor 2 is for later convenience), and h is the induced metric on the 3-surface $\partial\tilde{\mathcal{M}}$. In this section, we show how to choose the required function K so that the variation of the action $S_G[g]$ yields simply

$$\delta S_G[g] = \int_{\tilde{\mathcal{M}}}\sqrt{-g}\mathrm{d}^4p\left(R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu}\right)\delta g^{\mu\nu}(p),$$

without any total divergence terms. The only boundary condition will be that $\delta g^{\mu\nu}$ should vanish on $\partial\tilde{\mathcal{M}}$.

An analogy with ordinary mechanics may be helpful. Consider the familiar nonrelativistic Lagrangian

$$L(q,\dot{q}) = \frac{1}{2}m\dot{q}^2 - V(q),$$

describing a particle of mass m in a potential $V(q)$. This Lagrangian yields the same Euler-Lagrange equations as

$$\tilde{L}(q,\dot{q},\ddot{q}) = -\frac{1}{2}mq\ddot{q} - V(q).$$

However, the Lagrangian $\tilde{L}$ depends on the second derivative of q. The variation of $\tilde{L}$ with respect to $\delta q(t)$ contains boundary terms with $\delta\dot{q}(t)$ as well as $\delta q(t)$,

$$\delta\int_{t_1}^{t_2}\tilde{L}\mathrm{d}t = -\frac{1}{2}mq\delta\dot{q}\Big|_{t_1}^{t_2} + \frac{1}{2}m\dot{q}\delta q\Big|_{t_1}^{t_2} - \int_{t_1}^{t_2}\left(m\ddot{q} + V'(q)\right)\delta q(t)\mathrm{d}t.$$

At first glance, we are required to impose the boundary conditions $\delta\dot{q}(t) = \delta q(t) = 0$ for $t = t_{1,2}$. However, the problem with the Lagrangian $\tilde{L}$ can be fixed by adding a "boundary term," so that the "corrected" action becomes

$$S_{\mathrm{corr}}[q] = \int_{t_1}^{t_2}\tilde{L}(q,\dot{q},\ddot{q})dt + \frac{1}{2}mq\dot{q}\Big|_{t_1}^{t_2}.$$

The variation of the boundary term is

$$\delta\,\frac{1}{2}mq\dot{q}\Big|_{t_1}^{t_2} = \frac{1}{2}mq\delta\dot{q}\Big|_{t_1}^{t_2} + \frac{1}{2}m\dot{q}\delta q\Big|_{t_1}^{t_2},$$

which precisely cancels the unwanted term with $\delta\dot{q}$ in the variation $\delta\tilde{L}$.

Instead of adding a boundary term, we may add its derivative to the Lagrangian $\tilde{L}$, so that the "corrected" action is written as

$$S_{\mathrm{corr}}[q] = \int_{t_1}^{t_2}\left[\tilde{L}(q,\dot{q},\ddot{q}) + \frac{d}{dt}\left(\frac{1}{2}mq\dot{q}\right)\right]\mathrm{d}t.$$

Of course, this is identical to the usual action $\int Ldt$ for the particle in a potential. The additional term cancels the second derivative present in $\tilde{L}$.

In General Relativity, it is more convenient to keep the boundary term as such. The Einstein-Hilbert Lagrangian $\sqrt{-g}R$ is equivalent to other Lagrangians depending only on first derivatives of the metric, but none of the "corrected" Lagrangians cannot be written in a covariant form.

Presently, we are looking for a function K such that

$$\delta\left(2K\right) = g(\mathbf{n},\mathbf{q}), \tag{5.35}$$

where $\mathbf{n}$ is a normalized vector field orthogonal to the 3-surface $\partial\tilde{\mathcal{M}}$. If we find such K, then the boundary term will be equal to the volume integral of $\mathrm{div}\left(2K\mathbf{n}\right)$.

The induced partial metric h is defined using the orthogonal projector onto the tangent bundle $T\partial\tilde{\mathcal{M}}$,

$$h_{\mu\nu} = g_{\mu\nu} - n_\mu n_\nu; \quad h^{\mu\nu} = g^{\mu\nu} - n^\mu n^\nu.$$

(For simplicity, we assume that the 3-surface $\partial\tilde{\mathcal{M}}$ is everywhere spacelike and $g(\mathbf{n},\mathbf{n}) = 1$. For spacelike $\mathbf{n}$, the induced metric would be $h^{-1} = g^{-1} + \mathbf{n}\otimes\mathbf{n}$ and we would need to take $\sqrt{h}$ instead of $\sqrt{-h}$. Finally, the surface $\partial\tilde{\mathcal{M}}$ may consist of timelike and spacelike pieces, as long as it is nowhere null.) The remainder of this section is occupied by a calculation showing that $K \equiv h^{\mu\nu}n_{\mu;\nu}$ satisfies Eq. (5.35).

Remark: Note that the quantity K is equal to the trace of the extrinsic curvature of the 3-surface $\partial\tilde{\mathcal{M}}$, namely

$$K = \mathrm{Tr}_{(\mathbf{x},\mathbf{y})}K_{\partial\tilde{\mathcal{M}}}(\mathbf{x},\mathbf{y}) = g^{\mu\nu}K_{\mu\nu} = h^{\mu\nu}K_{\mu\nu}.$$

On the surface $\partial\tilde{\mathcal{M}}$, we have $h^{\mu\nu}n_{\mu;\nu} = g^{\mu\nu}n_{\mu;\nu} = \mathrm{div}\mathbf{n}$. However, the function $\mathrm{div}\mathbf{n}$ contains non-tangential derivatives of the metric and thus cannot be used for a variational principle where $\nabla\delta g$ is not zero.

We begin with Eq. (5.17),

$$g(\mathbf{n},\mathbf{q}) = n_\beta\left(\delta g^{\alpha\beta}{}_{;\alpha} - g_{\mu\nu}\delta g^{\mu\nu;\beta}\right) = -n^\beta g^{\mu\nu}\left(\delta g_{\mu\beta;\nu} - \delta g_{\mu\nu;\beta}\right). \tag{5.36}$$

Since the variation $\delta g^{\mu\nu}$ of the metric vanishes on $\partial\tilde{\mathcal{M}}$, it follows that any tangential derivative of δg vanishes. Since the inverse partial metric $h^{\alpha\beta}$ consists of tensor products of tangential vectors, we have

$$h^{\alpha\beta}\delta g_{\mu\nu;\alpha} = 0. \tag{5.37}$$

Substituting $g^{\mu\nu} = h^{\mu\nu} + n^\mu n^\nu$ into Eq. (5.36), we therefore find

$$g(\mathbf{n},\mathbf{q}) = -n^\beta h^{\mu\nu}\left(\delta g_{\mu\beta;\nu} - \delta g_{\mu\nu;\beta}\right) = n^\beta h^{\mu\nu}\delta g_{\mu\nu;\beta}.$$

On the other hand, we consider the variation δK due to the variation δg. Since $\delta g = 0$ on the surface $\partial\tilde{\mathcal{M}}$, the vector field $\mathbf{n}$ remains normalized and orthogonal to the surface, and the only varying quantity is the Levi-Civita connection ∇, thus

$$\delta K \equiv h^{\mu\nu}\left(\tilde{\nabla}_\mu n_\nu - \nabla_\mu n_\nu\right) = -\delta\Gamma^\alpha_{\mu\nu}h^{\mu\nu}n_\alpha.$$

Using Eqs. (5.15) and (5.37), we find

$$\delta K = -\frac{1}{2}h^{\alpha\beta}n^\mu\left(\delta g_{\mu\alpha;\beta} + \delta g_{\mu\beta;\alpha} - \delta g_{\alpha\beta;\mu}\right) = \frac{1}{2}h^{\alpha\beta}n^\mu\delta g_{a\beta;\mu}.$$

Therefore,

$$2\delta K = h^{\alpha\beta}n^\mu\delta g_{a\beta;\mu} = g(\mathbf{n},\mathbf{q})$$

as required.

Remark: Adding a boundary integral to the Einstein-Hilbert action is equivalent to adding a total divergence $\nabla_\mu(2Kn^\mu) \equiv \mathrm{div}(2K\mathbf{n})$ to the Lagrangian $\sqrt{-g}R$, *provided that* the vector field $\mathbf{n}$ is extended from the boundary $\partial\tilde{\mathcal{M}}$ to the *entire* domain $\tilde{\mathcal{M}}$ (note that K is a function of $\mathbf{n}$ and the metric g). More generally, one can add an arbitrary total divergence $\mathrm{div}\mathbf{B}$ of some vector field $\mathbf{B}$ to the Lagrangian, without changing the equations of motion. When the vector field $\mathbf{B}$ is chosen appropriately, the variation δS_G of the "corrected" action

$$S_G[g] \equiv \frac{1}{16\pi G}\int_{\mathcal{M}} [R - \mathrm{div}\mathbf{B}]\,\sqrt{-g}\mathrm{d}^4x$$

contains only volume and boundary terms with $\delta g^{\mu\nu}$. This indicates that $S_G[g]$ is first-order in the derivatives of g, as are the Lagrangians of most other field theories. However, the vector field $\mathbf{B}$ cannot be expressed as a function of the metric g alone.

It seems[4] that $\mathbf{B}$ can be expressed through a choice of four vector fields, one of which should coincide with $\mathbf{n}$ on $\partial\tilde{\mathcal{M}}$. Denote by $K[\mathbf{n}]$ the extrinsic curvature function evaluated on a normalized vector $\mathbf{n}$. Since we will need to vary the metric while keeping the vector $\mathbf{n}$ fixed, we need a formula for $K[\mathbf{n}]$ that includes the normalization of $\mathbf{n}$,

$$K[\mathbf{n}] = \mathrm{div}\mathbf{n} - \frac{g(\mathbf{n}, \nabla_{\mathbf{n}}\mathbf{n})}{g(\mathbf{n},\mathbf{n})}.$$

Choose an orthonormal frame $\{\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ in the entire domain $\tilde{\mathcal{M}}$. The vector fields $\{\mathbf{e}_\mu\}$ are orthonormal with respect to the metric g. Then define

$$\mathbf{B} = \sum_\mu 2K[\mathbf{e}_\mu]g(\mathbf{e}_\mu, \mathbf{e}_\mu)\mathbf{e}_\mu.$$

This specifies $\mathbf{B}$ as a function of g that contains first derivatives of g. Then one can show that $R - \mathrm{div}\mathbf{B}$ depends only on first derivatives of g. This can be shown, for instance, by considering a conformal transformation $g \to \tilde{g} \equiv e^{2\lambda}g$. One computes $\tilde{R} - \tilde{\mathrm{div}}\tilde{\mathbf{B}}$ and checks that the term containing $\Box\lambda$ in $\tilde{R}$ cancels with the terms $\sum_\mu \mathbf{e}_\mu \circ (\mathbf{e}_\mu \circ \lambda)$ coming from $\mathbf{B}$. **This explanation needs some more detail.**

Thus, the explicit form of $S_G[g]$ depends not only on g but also on an essentially arbitrary choice of the vector field $\mathbf{B}$ in the interior of the domain $\tilde{\mathcal{M}}$. Therefore, it is impossible to write the first-order action $S_G[g]$ as an explicit and *covariant* (i.e. coordinate-independent) function only of g. In a particular coordinate system, the vector field $\mathbf{B}$ may be chosen with particular numerical components so that the first-order Lagrangian *appears to be* a function only of the components $g_{\mu\nu}$ and $g_{\mu\nu,\alpha}$. However, this form of the first-order Lagrangian is not covariant (it depends on a chosen coordinate system).

5.2.5 The Hamiltonian for pure gravity

Now we shall compute the canonical momenta and the Hamiltonian for pure gravity (without matter) in general relativity.

For simplicity, we choose a time foliation such that the (timelike) vector field $\mathbf{n} \equiv g^{-1}dt$ normal to the surfaces $t = \text{const}$ is normalized, $g(\mathbf{n},\mathbf{n}) = 1$. This means that $\mathbf{n}$ is a *geodesic* vector field (why?); also, t is the proper time along the orbits of $\mathbf{n}$. This choice of the time foliation, called the **synchronous gauge**, may be impossible to perform globally due to focusing of the field $\mathbf{n}$. However, this choice is always possible locally, because one can emit timelike geodesics normal to an arbitrary initial spacelike surface and use the proper time along the geodesics as the coordinate t. It is also convenient to choose the local coordinates x^a within a surface $t = \text{const}$ such that the basis vectors $\mathbf{e}_a$ are connecting vectors for $\mathbf{n}$, i.e. $\mathcal{L}_{\mathbf{n}}\mathbf{e}_a = 0$. This completely removes the coordinate freedom ("fixes the gauge").

The partial metric h induced on the 3-surface $t = \text{const}$ is

$$h_{\mu\nu} = g_{\mu\nu} - n_\mu n_\nu.$$

Hence, in a basis $\{\mathbf{n}, \mathbf{a}, \mathbf{b}, \mathbf{c}\}$, where $\mathbf{a}, \mathbf{b}, \mathbf{c}$ are orthogonal to $\mathbf{n}$, the metric g has the matrix representation

$$g_{\mu\nu} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \end{pmatrix},$$

where the stars indicate the unknown nonzero components h_{ab} of the partial metric h. Note that $\sqrt{-g} = \sqrt{-h}$, so the covariant 4-volume element is $\sqrt{-g}\mathrm{d}^4x = \sqrt{-h}\mathrm{d}t \wedge \mathrm{d}^3x_a$, while the covariant 3-volume element on a surface $t = \text{const}$ is $\sqrt{-h}\mathrm{d}^3x_a$.

It is clear that the Lagrangian depends only on h, therefore we choose $h(t, x_a)$ to be the set of "generalized coordinates" in the sense of the Hamiltonian formalism. The next task is to find the dependence of the Lagrangian, $\sqrt{-g}R$, on $\partial h_{ab}/\partial t$ and to compute the corresponding canonical momenta p_{ab}.

The time derivative $\dot{h}_{ab} \equiv \partial h_{ab}/\partial t$ can be expressed in a geometric way as

$$\dot{h} \equiv \mathcal{L}_{\mathbf{n}}h,$$

since the basis vectors in the 3-surface commute with $\mathbf{n} \equiv \partial_t$, i.e. $\mathcal{L}_{\mathbf{n}}\mathbf{e}_a = 0$:

$$\frac{\partial h_{ab}}{\partial t} = \partial_t \circ (h(\mathbf{e}_a, \mathbf{e}_b)) = \mathcal{L}_{\mathbf{n}}\left(h(\mathbf{e}_a, \mathbf{e}_b)\right) = (\mathcal{L}_{\mathbf{n}}h) \circ (\mathbf{e}_a, \mathbf{e}_b).$$

Hence, for tangent vectors $\mathbf{x}, \mathbf{y}$ we have

$$\begin{aligned}(\mathcal{L}_{\mathbf{n}}h) \circ (\mathbf{x},\mathbf{y}) &= \mathcal{L}_{\mathbf{n}}\left(h(\mathbf{x},\mathbf{y})\right) - h(\mathcal{L}_{\mathbf{n}}\mathbf{x}, \mathbf{y}) - h(\mathbf{x}, \mathcal{L}_{\mathbf{n}}\mathbf{y}) \\ &= \nabla_{\mathbf{n}}h(\mathbf{x},\mathbf{y}) - h(\nabla_{\mathbf{n}}\mathbf{x} - \nabla_{\mathbf{x}}\mathbf{n}, \mathbf{y}) - h(\mathbf{x}, \nabla_{\mathbf{n}}\mathbf{y} - \nabla_{\mathbf{y}}\mathbf{n}) \\ &= h(\nabla_{\mathbf{x}}\mathbf{n}, \mathbf{y}) + h(\mathbf{x}, \nabla_{\mathbf{n}}\mathbf{y}) = 2K(\mathbf{x},\mathbf{y}), \qquad (5.38)\end{aligned}$$

where $K(\cdot,\cdot)$ is the extrinsic curvature of the 3-surface $t = \text{const}$. Therefore, $\dot{h}_{ab} = 2K_{ab}$, and the dependence of the Einstein-Hilbert action on $\dot{h}_{ab}$ will be found if we express $S_{EH}[g] = \int \mathrm{d}^4x\sqrt{-g}R$ through the extrinsic curvature K_{ab} of the surfaces $t = \text{const}$.

We already saw a relation of precisely this sort, namely the Gauss-Codazzi equation (5.33) that expresses the 4-curvature $R \equiv {}^{(4)}R$ through the 3-curvature ${}^{(3)}R$ and the extrinsic curvature K_{ab}. The only nontrivial calculation in this section will be to show that

$$\int \mathrm{d}^4x\sqrt{-g}{}^{(4)}R = \int \mathrm{d}^4x\sqrt{-g}\left({}^{(3)}R + K_{ab}K^{ab} - K^a_aK^b_b\right), \qquad (5.39)$$

up to total derivative terms that we omit. Note that Eq. (5.33) is very similar to Eq. (5.33), except for the different sign at the K terms. We shall derive Eq. (5.39) from Eq. (5.33) at the end of this section, and now let us assume that Eq. (5.39) is valid and determine the canonical momenta.

[4] I need to make this consideration more precise.

Since the Lagrangian density $\sqrt{-g}R$ depends on $\partial h/\partial t$ only through the terms containing K_{ab}, the canonical momentum p^{ab} is

$$p^{ab} \equiv \frac{1}{16\pi G}\frac{\partial\left(\sqrt{-g}R\right)}{\partial \dot{h}_{ab}} = \frac{\sqrt{-h}}{32\pi G}\frac{\partial}{\partial K_{ab}}\left(K_{ab}K^{ab} - K^a_a K^b_b\right).$$

Rewriting for convenience

$$K_{ab}K^{ab} - K^a_a K^b_b = K_{ab}K_{cd}\left(h^{ac}h^{bd} - h^{ab}h^{cd}\right),$$

we find after a simple calculation

$$p^{ab} = \frac{\sqrt{-h}}{16\pi G}\left(K^{ab} - K^c_c h^{ab}\right). \tag{5.40}$$

The inverse relation allows us to express the "velocity" $\dot{h}$ through the momentum p:

$$\dot{h}^{ab} = 2K^{ab} = \frac{32\pi G}{\sqrt{-h}}\left(p^{ab} - \frac{1}{2}p^c_c h^{ab}\right).$$

Now we can write the Hamiltonian for pure gravity:

$$\begin{aligned}
&\mathcal{H}[p_{ab}, h_{ab}] \\
&= \int_{t=\text{const}} \mathrm{d}^3x \left[p^{ab}\dot{h}_{ab} - \frac{\sqrt{-h}}{16\pi G}\left({}^{(3)}R + K_{ab}K^{ab} - K^a_a K^b_b\right)\right] \\
&= \int_{t=\text{const}} \mathrm{d}^3x \frac{\sqrt{-h}}{16\pi G}\left[K^{ab}K_{ab} - K^a_a K^b_b - {}^{(3)}R\right] \\
&= \int_{t=\text{const}} \mathrm{d}^3x \left[-\frac{\sqrt{-h}}{16\pi G}{}^{(3)}R + \frac{16\pi G}{\sqrt{-h}}\left(p^{ab}p_{ab} - \frac{1}{2}p^a_a p^b_b\right)\right].
\end{aligned} \tag{5.41}$$

The Hamilton equations of motion are cumbersome but follow straightforwardly,

$$\begin{aligned}
\frac{\partial h_{ab}}{\partial t} &= \frac{\partial H}{\partial p^{ab}} = \frac{32\pi G}{\sqrt{-h}}\left(p_{ab} - \frac{1}{2}h_{ab}p^c_c\right), \\
\frac{\partial p_{ab}}{\partial t} &= -\frac{\delta H}{\delta h^{ab}} \\
&= -\frac{\sqrt{-h}}{16\pi G}\frac{\delta {}^{(3)}R}{\delta h^{ab}} + \frac{\delta}{\delta h^{ab}}\left(p^{ab}p_{ab} - \frac{1}{2}p^a_a p^b_b\right)\frac{16\pi G}{\sqrt{-h}} \\
&= -\frac{\sqrt{-h}}{16\pi G}\left({}^{(3)}R_{ab} - \frac{1}{2}{}^{(3)}R h_{ab}\right) \\
&\quad + \frac{8\pi G}{\sqrt{-h}}\left[\left(p^{ab}p_{ab} - \frac{1}{2}p^a_a p^b_b\right)h_{ab} - 4\left(p_{ac}p^c_b - \frac{1}{2}p^c_c p_{ab}\right)\right].
\end{aligned}$$

Note that we used the three-dimensional Ricci tensor ${}^{(3)}R_{ab}$ when computing the variation of ${}^{(3)}R$.

The equations of motion in the Hamiltonian form are equivalent to Einstein's equations, but now it is clear that the metric $h_{ab}(t)$ and the extrinsic curvature $K_{ab}(t)$ can be found from initial data $h_{ab}(t_0)$, $K_{ab}(t_0)$ at an initial spacelike 3-surface $t = t_0$, as a solution of an appropriate Cauchy problem. This has applications in "numerical relativity," i.e. a numerical solution of the Einstein equations.

Derivation of Eq. (5.39)

Let us now derive Eq. (5.39) from Eq. (5.33).

The 3-dimensional curvature ${}^{(3)}R$ is defined as the 3-dimensional trace of the induced Riemann tensor ${}^{(3)}R(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d})$ with respect to $(\mathbf{a},\mathbf{c})$ and $(\mathbf{b},\mathbf{d})$. Since $\mathbf{n}$ is orthogonal to the 3-surfaces, the 3-dimensional trace ${}^{(3)}\mathrm{Tr}$ of any tensor A can be expressed through its 4-dimensional trace as

$${}^{(3)}\mathrm{Tr}_{(\mathbf{x},\mathbf{y})}A(\mathbf{x},\mathbf{y}) = {}^{(4)}\mathrm{Tr}_{(\mathbf{x},\mathbf{y})}A(\mathbf{x},\mathbf{y}) - A(\mathbf{n},\mathbf{n}).$$

The Gauss-Codazzi equation (5.33) is

$${}^{(3)}R(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{t}) = {}^{(3)}R(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{t}) + K(\mathbf{x},\mathbf{t})K(\mathbf{y},\mathbf{z}) - K(\mathbf{y},\mathbf{t})K(\mathbf{x},\mathbf{z}).$$

We begin by evaluating the 3-dimensional trace of the above equation over $(\mathbf{x},\mathbf{z})$:

$$\begin{aligned}
{}^{(3)}\mathrm{Tr}_{(\mathbf{x},\mathbf{z})}{}^{(3)}R(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{t}) &= {}^{(3)}\mathrm{Tr}_{(\mathbf{x},\mathbf{z})}{}^{(4)}R(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{t}) \\
&\quad + {}^{(3)}\mathrm{Tr}_{(\mathbf{x},\mathbf{z})}K(\mathbf{x},\mathbf{t})K(\mathbf{y},\mathbf{z}) - K(\mathbf{y},\mathbf{t})\mathrm{Tr}\,K \\
&= {}^{(4)}\mathrm{Tr}_{(\mathbf{x},\mathbf{z})}{}^{(4)}R(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{t}) - {}^{(4)}R(\mathbf{n},\mathbf{y},\mathbf{n},\mathbf{t}) \\
&\quad + {}^{(3)}\mathrm{Tr}_{(\mathbf{x},\mathbf{z})}K(\mathbf{x},\mathbf{t})K(\mathbf{y},\mathbf{z}) - K(\mathbf{y},\mathbf{t})K^a_a,
\end{aligned}$$

where we wrote $K^a_a \equiv {}^{(3)}\mathrm{Tr}_{(\mathbf{x},\mathbf{y})}K(\mathbf{x},\mathbf{y})$. Then we evaluate the 3-trace of the result over $(\mathbf{y},\mathbf{t})$, using the fact that $R(\mathbf{n},\mathbf{n},\cdot,\cdot) = 0$ and hence

$${}^{(3)}\mathrm{Tr}_{(\mathbf{x},\mathbf{z})}{}^{(4)}R(\mathbf{n},\mathbf{x},\mathbf{n},\mathbf{z}) = {}^{(4)}\mathrm{Tr}_{(\mathbf{x},\mathbf{z})}{}^{(4)}R(\mathbf{n},\mathbf{x},\mathbf{n},\mathbf{z}).$$

We find

$$\begin{aligned}
{}^{(3)}R \equiv {}^{(3)}\mathrm{Tr}_{(\mathbf{x},\mathbf{z})(\mathbf{y},\mathbf{t})}{}^{(3)}R(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{t}) &= {}^{(4)}R \\
&\quad - 2{}^{(4)}\mathrm{Tr}_{(\mathbf{x},\mathbf{z})}{}^{(4)}R(\mathbf{n},\mathbf{x},\mathbf{n},\mathbf{z}) + K_{ab}K^{ab} - K^a_a K^b_b.
\end{aligned}$$

It remains to compute the term ${}^{(4)}\mathrm{Tr}_{(\mathbf{x},\mathbf{z})}{}^{(4)}R(\mathbf{n},\mathbf{x},\mathbf{n},\mathbf{z})$. I shall present a computation in the index notation and also an index-free computation, for comparison. Let us first convert this term to the index notation:

$$\begin{aligned}
\mathrm{Tr}_{(\mathbf{x},\mathbf{z})}{}^{(4)}R(\mathbf{n},\mathbf{x},\mathbf{n},\mathbf{z}) &\equiv \mathrm{Tr}_{(\mathbf{x},\mathbf{z})}z^\delta g_{\gamma\delta}n^\alpha x^\beta \left[\nabla_\alpha,\nabla_\beta\right]n^\gamma \\
&= g^{\beta\gamma}g_{\gamma\delta}n^\alpha\left[\nabla_\alpha,\nabla_\beta\right]n^\gamma = n^\alpha\left[\nabla_\alpha,\nabla_\beta\right]n^\beta.
\end{aligned}$$

The last expression is simplified as

$$\begin{aligned}
n^\alpha n^\beta{}_{;\beta\alpha} - n^\alpha n^\beta{}_{;\alpha\beta} &= \left(n^\alpha n^\beta_{;\beta}\right)_{;\alpha} - n^\alpha_{;\alpha}n^\beta_{;\beta} - \left(n^\alpha n^\beta_{;\alpha}\right)_{;\beta} + n^\alpha_{;\beta}n^\beta_{;\alpha} \\
&= \mathrm{div}\,(\ldots) + K_{\alpha\beta}K^{\alpha\beta} - K^\alpha_\alpha K^\beta_\beta,
\end{aligned}$$

where we used the relation $K_{\alpha\beta} = n_{\alpha;\beta}$ and omitted the uninteresting total divergence terms. Finally, we obtain

$${}^{(3)}R \equiv {}^{(4)}R - K_{ab}K^{ab} + K^a_a K^b_b + \mathrm{div}\,(\ldots),$$

which is equivalent to Eq. (5.39).

In the index-free notation, the manipulations with traces are somewhat less transparent. We need to use the property that mute vectors have vanishing derivatives under the trace operation, as well as

$$A(\mathbf{x},\mathbf{y},\mathbf{z},\ldots) = \mathrm{Tr}_{(\mathbf{a},\mathbf{b})}g(\mathbf{a},\mathbf{x})A(\mathbf{b},\mathbf{y},\mathbf{z},\ldots)$$

(see Sec. 1.7.3). Also, we have $\nabla_{\mathbf{n}}\mathbf{n} = 0$ and $K(\mathbf{x},\mathbf{y}) = K(\mathbf{y},\mathbf{x})$. Thus,

$$\begin{aligned}
\mathrm{Tr}_{(\mathbf{x},\mathbf{z})}{}^{(4)}R(\mathbf{n},\mathbf{x},\mathbf{n},\mathbf{z}) &= \mathrm{Tr}_{(\mathbf{x},\mathbf{z})}g(\mathbf{z},\nabla_{\mathbf{n}}\nabla_{\mathbf{x}}\mathbf{n} + \nabla_{\nabla_{\mathbf{x}}\mathbf{n}}\mathbf{n}) \\
&= \mathrm{Tr}_{(\mathbf{a},\mathbf{b})(\mathbf{x},\mathbf{z})}\left[g(\mathbf{a},\mathbf{n})g(\mathbf{z},\nabla_{\mathbf{b}}\nabla_{\mathbf{x}}\mathbf{n}) + g(\mathbf{a},\nabla_{\mathbf{x}}\mathbf{n})g(\mathbf{z},\nabla_{\mathbf{b}}\mathbf{n})\right].
\end{aligned}$$

The second term in brackets yields directly

$$\mathrm{Tr}_{(\mathbf{a},\mathbf{b})(\mathbf{x},\mathbf{z})} g(\mathbf{a},\nabla_{\mathbf{x}}\mathbf{n}) g(\mathbf{z},\nabla_{\mathbf{b}}\mathbf{n}) = \mathrm{Tr}_{(\mathbf{a},\mathbf{b})(\mathbf{x},\mathbf{z})} K(\mathbf{a},\mathbf{x}) K(\mathbf{z},\mathbf{b}) = K_{ab}K^{ab}$$

while the first term in brackets is transformed as

$$g(\mathbf{b},\mathbf{n}) g(\mathbf{z},\nabla_{\mathbf{a}}\nabla_{\mathbf{x}}\mathbf{n}) = \nabla_{\mathbf{a}}\left[g(\mathbf{b},\mathbf{n}) g(\mathbf{z},\nabla_{\mathbf{x}}\mathbf{n})\right] - g(\mathbf{b},\nabla_{\mathbf{a}}\mathbf{n}) g(\mathbf{z},\nabla_{\mathbf{x}}\mathbf{n}),$$

which yields a total divergence and the term

$$-\mathrm{Tr}_{(\mathbf{a},\mathbf{b})(\mathbf{x},\mathbf{z})} g(\mathbf{b},\nabla_{\mathbf{a}}\mathbf{n}) g(\mathbf{z},\nabla_{\mathbf{x}}\mathbf{n}) = -\left[\mathrm{Tr}_{(\mathbf{a},\mathbf{b})} K(\mathbf{a},\mathbf{b})\right]^2 = -\left(K^a_a\right)^2.$$

Therefore we again recover Eq. (5.39).

5.2.6 Constraints in General Relativity

The Hamilton equations of motion are incomplete without the constraint equations which we neglected to derive earlier. These constraints are present because the Lagrangian L_{EH} is independent of the time derivatives of the components $g^{0\mu}$ of the metric. To see this explicitly, we can repeat the derivation of the Hamiltonian formulation without assuming that t is the proper time along geodesic lines of $\mathbf{n}$.

We have computed the Hamiltonian in the syncronous gauge for simplicity, but the calculation can be performed in an arbitrary gauge, i.e. with an arbitrary selection of the time foliation. Suppose that a function t is given on a spacetime, such that the contravariant gradient vector $g^{-1}dt$ is everywhere timelike (but not necessarily normalized). We may use the function t as the time coordinate and define equal-time 3-surfaces $t = \text{const}$. Let us define the normal vector to the 3-surfaces, $\mathbf{n} \equiv N\hat{g}^{-1}\mathrm{d}t$, where N is such that $g(\mathbf{n},\mathbf{n}) = 1$, namely $N^{-2} = g^{-1}(\mathrm{d}t,\mathrm{d}t)$. The local coordinates x_a on these 3-surfaces may be chosen arbitrarily. Given a choice of the local coordinates x_a, we may define the vector ∂_t which acts on functions $f(t,x_a)$ as $f \to \partial f/\partial t$, the partial derivative at fixed x_a. The vector ∂_t is not necessarily parallel to $\mathbf{n}$, and by construction we have

$$g(\mathbf{n},\partial_t) = Ng(\hat{g}^{-1}\mathrm{d}t,\partial_t) = N\,(\mathrm{d}t)\circ(\partial_t) = N,$$

hence in general we can decompose ∂_t as

$$\partial_t = N\mathbf{n} + \mathbf{s},$$

where $\mathbf{s}$ is a spacelike vector tangent to the surface $t = \text{const}$, which is thus equivalent to a 3-vector $\mathbf{s} = \sum_{a=1}^3 s^a \mathbf{e}_a$ in the local basis $\{\mathbf{e}_a\}$. The parameters N and s^a are called the **lapse function** and the **shift vector** for a chosen coordinate system. The lapse describes how much time "elapses" between two consecutive $t = \text{const}$ surfaces along the lines of $\mathbf{n}$, and the shift vector shows how much the local coordinate system shifts when we pass from one $t = \text{const}$ surface to another along $\mathbf{n}$.

Let us now determine the induced metric h within the 3-surfaces. The matrix $h_{ab} \equiv g(\mathbf{e}_a,\mathbf{e}_b)$ is related to the metric $g_{\mu\nu}$ in the basis $\{\partial_t,\mathbf{e}_a\}$ by the following matrix representation (verify this!),

$$g_{\mu\nu} = \begin{pmatrix} N^2 + h(\mathbf{s},\mathbf{s}) & s_1 & s_2 & s_3 \\ s_1 & h_{11} & h_{12} & h_{13} \\ s_2 & h_{21} & h_{22} & h_{23} \\ s_3 & h_{31} & h_{32} & h_{33} \end{pmatrix},$$

where $s_a \equiv h_{ab}s^a$. It follows from this representation that the determinants of g and h are related by $\sqrt{-g} = N\sqrt{-h}$. (See Calculation below.)

Calculation: Show that $\det g = N^2 \det h$ when $g_{\mu\nu}$ is related to h_{ab} by the above formula.

Solution: To compute $\det g$ directly, subtract from the first row of the matrix $g_{\mu\nu}$ the linear combination of the other three rows with coefficients s^a. This does not change the determinant, but the first row of the matrix simplifies to $(N^2,0,0,0)$. Hence

$$\begin{aligned}\det g_{\mu\nu} &= \det \begin{vmatrix} N^2 + s_a s^a & s_a \\ s_b & h_{ab} \end{vmatrix} \\ &= \det \begin{vmatrix} N^2 & 0 \\ s_b & h_{ab} \end{vmatrix} = N^2 \det h_{ab}.\end{aligned}$$

The action (5.39) involves ${}^{(3)}R$, which is independent of the choice of gauge, and the extrinsic curvature K_{ab}. The relationship between the extrinsic curvature $K(\mathbf{x},\mathbf{y}) = h(\mathbf{x},\nabla_{\mathbf{y}}\mathbf{n})$ and the metric derivatives $\partial h_{ab}/\partial t$ needs to be derived now. Since ∂_t is a connecting vector for $\mathbf{e}_a$, we have

$$\begin{aligned}\frac{\partial h_{ab}}{\partial t} &\equiv \partial_t \circ [h(\mathbf{e}_a,\mathbf{e}_b)] = \mathcal{L}_{\partial_t}[h(\mathbf{e}_a,\mathbf{e}_b)] \\ &= \left(\mathcal{L}_{\partial_t} h\right)\circ(\mathbf{e}_a,\mathbf{e}_b),\end{aligned}$$

and hence for arbitrary vectors $\mathbf{x},\mathbf{y}$ tangent to a 3-surface, we find, similarly to Eq. (5.38),

$$\left(\mathcal{L}_{\partial_t} h\right)\circ(\mathbf{x},\mathbf{y}) = h(\nabla_{\mathbf{x}}\partial_t,\mathbf{y}) + h(\mathbf{x},\nabla_{\mathbf{y}}\partial_t).$$

Since $h(\mathbf{n},\cdot) = 0$, we can simplify the above terms, e.g.

$$\begin{aligned}h(\nabla_{\mathbf{x}}\partial_t,\mathbf{y}) &= h(\nabla_{\mathbf{x}}(N\mathbf{n}+\mathbf{s}),\mathbf{y}) \\ &= Nh(\nabla_{\mathbf{x}}\mathbf{n},\mathbf{y}) + h(\nabla_{\mathbf{x}}\mathbf{s},\mathbf{y}) \\ &= NK(\mathbf{x},\mathbf{y}) + h({}^{(3)}\nabla_{\mathbf{x}}\mathbf{s},\mathbf{y}).\end{aligned}$$

In the last line, we replaced ∇ by ${}^{(3)}\nabla$ since, by definition,

$$h({}^{(3)}\nabla_{\mathbf{x}}\mathbf{y},\mathbf{z}) \equiv g(\nabla_{\mathbf{x}}\mathbf{y},\mathbf{z})$$

for tangent vectors $\mathbf{x},\mathbf{y},\mathbf{z}$. Hence,

$$\partial h_{ab}/\partial t = 2NK_{ab} + s_{a|b} + s_{b|a},$$

where we denote ${}^{(3)}\nabla_a$ by a vertical bar.

It now follows from Eq. (5.39) that the Lagrangian $L_{EH} = \sqrt{-g}R$ does not depend on the time derivatives of N and s_a. Therefore, we do not need to introduce the canonical momenta for these nondynamical variables. As before, the only canonical momenta are p_{ab} corresponding to the components h_{ab} of the induced 3-metric, and a simple calculation shows that p_{ab} is still related to K_{ab} by Eq. (5.40). Hence, the Hamiltonian is

$$\begin{aligned}\mathcal{H} = \int d^3x_a \frac{\sqrt{-h}}{16\pi G}\Big[&\left(K^{ab}K_{ab} - K^a_a K^b_b - {}^{(3)}R\right)N \\ &- \left(K^{ab} - K^c_c h^{ab}\right)_{|a} s_b\Big].\end{aligned}$$

A complete derivation of the Hamiltonian in an arbitrary gauge (including also a detailed treatment of all the boundary terms) can be found in the book [Poisson 2004], chapter 4. Here, we shall only determine the constraints of the theory.

The variables N and s_a enter the Hamiltonian as nondynamical Lagrange multipliers, therefore the constraints are

$$\begin{aligned}\frac{\delta\mathcal{H}}{\delta N} = 0 \;&\Rightarrow\; KK_{ab} - K^a_a K^b_b - {}^{(3)}R = 0, \\ \frac{\delta\mathcal{H}}{\delta s_a} = 0 \;&\Rightarrow\; \left(K^{ab} - K^c_c h^{ab}\right)_{|a} = 0.\end{aligned}$$

The constraint $\delta\mathcal{H}/\delta N = 0$ is called the **Hamiltonian constraint** or the **energy constraint**, while $\delta\mathcal{H}/\delta s_a = 0$ is called the **momentum constraint**. We find (perhaps with surprise) that the constraints make the Hamiltonian itself vanish, $\mathcal{H} = 0$.

Remark: A vanishing Hamiltonian is a characteristic feature of any theory which is invariant under arbitrary coordinate transformations. The constraint means that $\mathcal{H}[p_{ab}, h_{ab}] = 0$ for p_{ab}, h_{ab} that solve the equations of motion. It is important to realize that the constraint $\mathcal{H} = 0$ does not make the theory trivial; the Hamiltonian $\mathcal{H}[p_{ab}, h_{ab}]$ is a *nontrivial functional* of the canonical variables p_{ab}, h_{ab} and can be used to *derive* the equations of motion.

5.3 Quantum cosmology

Additional literature: [35].

The Hamiltonian description of General Relativity was developed to quantize that theory. We shall now describe the approach to quantization of gravity due to J. Wheeler and B. DeWitt. So far it was impossible to develop a full quantum theory of gravity based on the Hamiltonian description presented above. Technical difficulties are too great to derive in detail, for example, the quantum state that corresponds to a Schwarzschild spacetime. Therefore this approach to quantization of gravity remains formal.[5] However, as we shall now see, some interesting qualitative results *can* be obtained in the Wheeler-DeWitt approach.

5.3.1 Wave function of the universe

According to the general scheme of canonical quantization, we need to replace the coordinates q, p by operators $\hat{p}, \hat{q}$ satisfying the standard equal-time commutation relations. In the case of gravitation, we shall therefore write

$$\left[\hat{p}_{mn}(t, x_a), \hat{h}^{cd}(t, x_b)\right] = \delta^c_m \delta^d_n \delta^{(3)}(x_a - x_b).$$

It is easier to visualize the quantum theory in the Schrödinger picture. The quantum state in the "coordinate representation" is a **wave function** $\psi(t; q_i)$ of time t and the generalized coordinate q_i; the operator $\hat{q}_i$ acts as multiplication by q_i, and $\hat{p}_i = -\mathrm{i}\frac{\partial}{\partial q_i}$. In General Relativity, the role of q_i is played by the 3-metric $h_{ab}(x_c)$, stripped of the dependence on time. Therefore, the "wave function" is a functional $\Psi\left[t; h_{ab}(x_c)\right]$. Heuristically, one might expect that $|\Psi[t; h_{ab}]|^2$ is the probability of observing the metric $h_{ab}(x_c)$ at time t. (Below, we shall see that the interpretation of this wave function is not quite as straightforward as it may appear.)

We may also consider other matter fields $\phi_j(t, x_a)$ coupled to gravity. In the Hamiltonian formalism, these fields will be described by generalized coordinates $\phi_j(x_a)$ and the corresponding canonical momenta $p_j(x_a)$. Thus, the wave function of the full theory will be a functional of h_{ab} and ϕ_j. Such a functional $\Psi[t; h_{ab}(x_c), \phi_j(x_c)]$ is called the **wave function of the universe** because it describes (in principle) all the possible processes anywhere in the universe.

The space consisting of all the possible gravity and matter field configurations $\{h_{ab}(x_c), \phi_j(x_c)\}$ is called the **superspace**.[6] Thus, the wave function of the universe is a (complex-valued) function on superspace.

According to the standard rules of quantum mechanics, the wave function of the universe satisfies the Schrödinger equation,

$$\begin{aligned}\mathrm{i}\frac{\partial}{\partial t}\Psi[t; h_{ab}, \phi_j] &= \hat{\mathcal{H}}\Psi[t; h_{ab}, \phi_j] \\ &\equiv \mathcal{H}\left(\hat{p}_{ab}, h_{ab}; \hat{p}_j, \phi_j\right)\Psi[t; h_{ab}, \phi_j] \\ &= \mathcal{H}\left[\frac{\delta}{\delta h_{ab}(x_c)}, h_{ab}(x_c); \frac{\delta}{\delta\phi_j(x_c)}, \phi_j(x_c)\right]\Psi[t; h_{ab}, \phi_j].\end{aligned}$$

Here, the operator $\hat{\mathcal{H}}$ is the total Hamiltonian of the system, which is itself a functional of all the canonical variables. For instance, the Hamiltonian for pure gravity is given by Eq. (5.41). Note that the operators $\hat{p}_{ab}$ and $\hat{p}_j$ are replaced by *functional* derivatives with respect to h_{ab} and ϕ_j.

5.3.2 Wheeler-DeWitt equation

Since classical General Relativity is a constrained theory, the constraints must be passed on to the quantum theory as well. A comprehensive treatment of quantization for constrained systems is beyond the scope of these lectures; here we shall adopt the simplistic point of view that the constraint equations, which are of the form $C(q_i, p_i) = 0$, should be made operator-valued and imposed as operators on physical states:

$$C(\hat{q}_i, \hat{p}_i)\Psi(q_i) = 0.$$

In other words, quantum states Ψ satisfying the constraints are the only valid physical states of the system. It follows that the Hamiltonian constraint, $\mathcal{H} = 0$, is translated to the restriction

$$\hat{\mathcal{H}}\Psi[t; h_{ab}, \phi_j] = 0. \tag{5.42}$$

Then the Schrödinger equation yields

$$\mathrm{i}\frac{\partial}{\partial t}\Psi[t; h_{ab}, \phi_j] = 0.$$

Hence, the wave function of the universe is *time-independent* and satisfies the Hamiltonian constraint (5.42), which is called in this context the **Wheeler-DeWitt (WD) equation**. The time independence of the wave function may appear to be a puzzling feature of the theory. However, it is a necessary feature: In a generally covariant theory, the time parameter t is an arbitrary label on events in spacetime, and physically meaningful probabilities must be defined in terms of coordinate-independent functionals of h_{ab} and ϕ_j. We shall consider the interpretation of $\Psi[t; h_{ab}, \phi_j]$ in the next section.

Although we wrote the equations for quantized General Relativity, it remains difficult to extract any tangible results from these equations. Quite apart from the problem of solving the highly complicated equation (5.42), the question of the **operator ordering**: The Hamiltonian $\mathcal{H}$ is a nonlinear function

[5] Currently, a more promising approach to canonical quantization of General Relativity is a Hamiltonian approach based on a different choice of generalized coordinates, called "Ashtekar variables." The theory based on these variables is called "loop quantum gravity" and is beyond the scope of this book.

[6] Note that a point of superspace is a configuration $h_{ab}(x_c), \phi_j(x_c)$ which contains functions only of 3-dimensional coordinates x_c, not of time t. This can be visualized as an "instantaneous" field configuration on a 3-surface of constant time.

of h_{ab} and p_{ab}, containing terms such as $h_{ab}h_{cd}p^{ac}p^{bd}$, hence it is unclear how to order these noncommuting operators in the quantum Hamiltonian $\hat{\mathcal{H}}$. The quantum Hamiltonian is thus not a well-defined operator unless we adopt a particular prescription for the operator ordering. To determine the "correct" operator ordering, one needs to compute some predictions of the quantum theory with one or another operator ordering and to compare these predictions with experiments or with other known results. Such computations are generally too difficult, and thus the Wheeler-DeWitt equation (5.42) remains, in its full generality, a formula without application. (However, below we shall see that the Wheeler-DeWitt equation can be transformed into a differential equation and solved, if one restricts gravity and matter fields to spatially homogeneous configurations.)

5.3.3 Interpretation of the wave function

Even if one somehow computes the wave function of the universe Ψ, its interpretation is nontrivial, mainly because the functional $\Psi[h_{ab}, \phi_j]$ is independent from the time parameter t. This fact, however, does not mean that the theory always describes a *static* universe! A physically meaningful "time" must be defined not as the value of the parameter t, which can be changed by a coordinate transformation, but through some physical process (a "clock"). Since the wave function of the universe contains information about all the processes through its dependence on h_{ab} and ϕ_j, the theory can describe an evolving universe if an appropriate "clock process" could be found.

It is important to realize that the clock process must be realized by an *essentially classical* rather than by an essentially quantum physical system. In the quantum language, the generalized coordinate describing the clock process must exhibit very small quantum fluctuations around a large expectation value. If the universe were in a quantum state Ψ in which quantum fluctuations of every variable (including the metric) are significant, one could not expect to observe anything resembling a "flow of time." Therefore, even the *possibility* of an interpretation of the wave function of the universe depends on the existence of (nearly) classical systems in the universe.

Suppose that a nearly classical subsystem exists and is described by a variable ϕ_c. According to a standard result of quantum mechanics, the wave function of the subsystem is of the semiclassical (WKB) type,

$$\psi(\phi_c) \propto \exp\left(-\mathrm{i}S_{cl}(\phi_c)\right),$$

where $S_{cl}(\phi_c)$ is the value of the classical action on a classical trajectory $\phi(t)$, expressed as a function of the value of the variable ϕ_c at a final time,

$$S_{cl}(\phi_c) = \int_{t_0}^{t} L_{\phi_c}(\phi, \dot{\phi})dt, \quad \phi(t) = \phi_c.$$

The total wave function of the universe is a product of the semiclassical part $\psi(\phi_c)$ and the quantum part Ψ_q,

$$\Psi[h_{ab}, \phi_j] = \exp\left(-\mathrm{i}S_{cl}(\phi_c)\right)\Psi_q[\phi_c; h_{ab}, \phi_j].$$

Let us now see how the variable ϕ_c can be used as a clock process. The trajectory $\phi_c(t)$ of the classical subsystem is, at least locally, a monotonic function of t. Therefore, the values of ϕ_c can be used as "time" and then the quantum part of the wave function, $\Psi_q[\phi_c; h_{ab}, \phi_j]$, becomes a time-dependent wave function for the quantum variables h_{ab}, ϕ_j. One can show (see e.g. [35]) that a Schrödinger-type equation holds for this wave function, the variable ϕ_c playing the role of time, if the total wave function Ψ satisfies the WD equation.

5.3.4 "Minisuperspace"

The WD equation is a functional differential equation for a functional on a "superspace" and, as such, is too complicated to be solved in general. One can obtain specific results from the WD equation if one simplifies the problem sufficiently drastically.

Note that one of the main applications of General Relativity is **cosmology** where one considers only the gross features of the universe, that is, fields averaged over astronomically large scales. Astronomical observations offer a strong evidence that the universe around us is extremely homogeneous on large scales. Therefore, in cosmology one usually assumes that the 3-surfaces of constant time are spatially homogeneous. In other words, the 3-metric $h_{ab}(t)$ and the fields $\phi_j(t)$ depend only on the time t and are independent of the spatial coordinates x_c. Hence, we are motivated to consider the theory of quantized gravity with spatially homogeneous metric and fields. The corresponding simplification of the WD equation consists of reducing the infinite-dimensional superspace to a finite-dimensional space containing 3-metrics h_{ab} and field configurations ϕ_j that are spatially homogeneous, i.e. independent of the 3-coordinates x_c. The reduced superspace is called **minisuperspace**. The WD equation in minisuperspace becomes a (partial) differential equation for a wave function $\Psi(h_{ab}, \phi_j)$. The analysis of this simplified WD equation is the subject of **quantum cosmology**.

To be specific, let us consider a model of classical spacetime whose equal-time surfaces are homogeneous 3-spheres S^3 (hence, a closed universe!) with a fixed 3-metric γ. The spacetime metric is

$$g = N^2(t)\mathrm{d}t \otimes \mathrm{d}t - a^2(t)\gamma,$$

where $N(t)$ is the lapse function and $a(t)$ is an unknown function of time called the **scale factor**. Since the lapse function is nondynamical, the scale factor is the only physical variable in the gravitational sector of the model. Further, suppose there exists a scalar field $\phi(t)$ which is, again, independent of the spatial coordinates x_c. Thus the minisuperspace is two-dimensional and consists of two variables, a and ϕ. We shall now quantize this cosmological model using the WD equation.

We need to express the classical Hamiltonian (5.41) for gravity through the variable $a(t)$, and add the Hamiltonian for the scalar field ϕ,

$$\mathcal{H}_\phi = \int_{t=\text{const}} \mathrm{d}^3x_c N\sqrt{-h}\left[\frac{1}{2}p_\phi^2 + h^{ab}\phi_{,a}\phi_{,b} + V(\phi)\right],$$

where p_ϕ is the canonical momentum for ϕ, and

$$V(\phi) \equiv \Lambda + \frac{1}{2}m^2\phi^2 + O(\phi^3)$$

is a potential describing the vacuum energy density Λ, the mass m, and a possible self-interaction of the field ϕ. The partial metric $h = -a^2\gamma$, so $\sqrt{-h} = a^3\sqrt{\gamma}$. Since the field ϕ is

spatially homogeneous, we may integrate over 3-spheres, assuming that the metric γ is normalized so that

$$\int_{t=\text{const}} \mathrm{d}^3 x_c \sqrt{\gamma} = \sigma_3 = 2\pi^2,$$

where

$$\sigma_n \equiv \frac{2\pi^{n/2}}{\Gamma(n/2)}$$

is the n-volume of a unit n-sphere S^n. (This normalization of the 3-metric γ means that the radius of the 3-sphere is equal to 1.) Thus we obtain

$$\mathcal{H}_\phi = N a^3 \sigma_3 \left[\frac{1}{2} p_\phi^2 + V(\phi)\right].$$

Returning to the gravitational sector of the model, we need to compute the 3-curvature scalar ${}^{(3)}R$ and the extrinsic curvature tensor K_{ab}. The 3-curvature ${}^{(3)}R$ depends only on the intrinsic geometry of the 3-surface $t = \text{const}$, which is a 3-sphere S^3 with the natural metric and radius $a(t)$, and thus a space of constant curvature; the value of the curvature is $a^{-1}(t)$. The 3-dimensional Riemann tensor for a space of constant curvature was computed in Sec. 1.10.3, hence

$${}^{(3)}R(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d}) = a^{-2}\left(h(\mathbf{a},\mathbf{c})h(\mathbf{b},\mathbf{d}) - h(\mathbf{a},\mathbf{d})h(\mathbf{b},\mathbf{c})\right).$$

Using

$${}^{(3)}\mathrm{Tr}_{(\mathbf{a},\mathbf{b})} h(\mathbf{a},\mathbf{b}) = 3,$$

we find

$${}^{(3)}R = {}^{(3)}\mathrm{Tr}_{(\mathbf{a},\mathbf{c})(\mathbf{b},\mathbf{d})}\, {}^{(3)}R(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d}) = 6a^{-2}.$$

It remains to compute the extrinsic curvature. We may introduce local coordinates x_a on a surface $t = \text{const}$; however, the explicit form of these coordinates will not be necessary for the present calculation. It is sufficient to note that the basis 3-vectors $\mathbf{e}_a \equiv \partial_{x_a}$ commute with each other and with ∂_t. Hence, from Eq. (1.45) we get

$$g(\nabla_{\mathbf{e}_b}\partial_t, \mathbf{e}_c) = \frac{1}{2}\partial_t g(\mathbf{e}_b, \mathbf{e}_c) = \dot{a} a \gamma(\mathbf{e}_b, \mathbf{e}_c) = \frac{\dot{a}}{a} h_{bc}.$$

Since the normal vector $\mathbf{n}$ to the 3-surfaces $t = \text{const}$ is $\mathbf{n} = N^{-1}\partial_t$, the extrinsic curvature tensor is

$$K(\mathbf{x},\mathbf{y}) \equiv g(\nabla_{\mathbf{x}}\mathbf{n}, \mathbf{y}) = N^{-1} g(\nabla_{\mathbf{x}}\partial_t, \mathbf{y}) = N^{-1}\frac{\dot{a}}{a} h(\mathbf{x},\mathbf{y}).$$

Then we determine the relevant combination of the traces of K,

$$K_{bc}K^{bc} - K_b^b K_c^c = \left(\frac{\dot{a}}{Na}\right)^2 \left(h_{bc}h^{bc} - h_b^b h_c^c\right) = -6\left(\frac{\dot{a}}{Na}\right)^2.$$

The gravitational Lagrangian becomes

$$\begin{aligned}\mathcal{L}_G &= \int_{t=\text{const}} \mathrm{d}^3 x \frac{N\sqrt{-h}}{16\pi G}\left({}^{(3)}R + K_{ab}K^{ab} - K_a^a K_b^b\right) \\ &= \int_{t=\text{const}} \mathrm{d}^3 x \frac{N\sqrt{-h}}{16\pi G}\left(\frac{6}{a^2} - 6\left(\frac{\dot{a}}{Na}\right)^2\right).\end{aligned}$$

We can now integrate over the 3-volume of the 3-sphere $t = \text{const}$,

$$\int_{t=\text{const}} \mathrm{d}^3 x \sqrt{-h} = a^3 \sigma_3,$$

because all the terms are spatially homogeneous. Therefore,

$$\mathcal{L} = \frac{6Na^3\sigma_3}{16\pi G}\left(\frac{1}{a^2} - \frac{\dot{a}^2}{N^2 a^2}\right) = \frac{3\pi}{4G} Na\left(1 - \frac{1}{N^2}\dot{a}^2\right).$$

To compute the Hamiltonian, it remains to determine the canonical momentum p_a,

$$p_a = \frac{\partial \mathcal{L}}{\partial \dot{a}} = -\frac{3\pi}{2G}\frac{a\dot{a}}{N}.$$

Hence,

$$\dot{a} = -\frac{2G}{3\pi}\frac{N}{a} p_a,$$

and the gravitational Hamiltonian is

$$\begin{aligned}\mathcal{H}_G(p_a, a) &= p_a \dot{a} - \mathcal{L}_G \\ &= \frac{-3\pi}{4G} Na + \frac{N}{a} p_a^2 \left(-\frac{2G}{3\pi} + \frac{3\pi}{4G}\frac{4G^2}{9\pi^2}\right) \\ &= \frac{-3\pi}{4G} Na - \frac{G}{3\pi}\frac{N}{a} p_a^2.\end{aligned}$$

Finally, we can write the total Hamiltonian for the minisuperspace theory,

$$\begin{aligned}&\mathcal{H}\left(p_a, a, p_\phi, \phi\right) = \mathcal{H}_G + \mathcal{H}_\phi \\ &= \frac{N}{a}\left[-\frac{G}{3\pi} p_a^2 - \frac{3\pi}{4G} a^2 + a^4 \sigma_3 \left(\frac{1}{2} p_\phi^2 + V(\phi)\right)\right].\end{aligned}$$

Therefore, the "wave function of the mini-universe," $\Psi(a,\phi)$, satisfies the WD equation

$$\begin{aligned}&\left[-\frac{G}{3\pi}\hat{p}_a^2 - \frac{3\pi}{4G} a^2 + 2\pi^2 a^4 \left(\frac{1}{2}\hat{p}_\phi^2 + V(\phi)\right)\right]\Psi \\ &= \left[\frac{G}{3\pi}\hbar^2\partial_a^2 - \frac{3\pi}{4G} a^2 + \pi^2 a^4\left(-\hbar^2\partial_\phi^2 + 2V(\phi)\right)\right]\Psi = 0.\end{aligned}$$

This equation needs to be supplemented by boundary conditions. However, the choice of the boundary conditions is a subtle problem and we shall not discuss this issue here. In the present case, we shall simply choose the solution of the WD equation that will have a clear physical interpretation.

It is reasonable to suppose that the scale factor a has a semiclassical regime where it is the *large* radius of the 3-sphere $t = \text{const}$. (Our universe is *large*.) Therefore, in the semiclassical regime the value of a can be used as the "clock" variable. We can then express the wave function in the semiclassical form with respect to the variable a,

$$\Psi(a,\phi) = \exp\left(-\mathrm{i}\hbar^{-1} S(a)\right)\psi(a,\phi),$$

where the factor $\psi(a,\phi)$ is assumed to be a slow-varying function of a but quickly-varying function of the quantum variable ϕ, namely

$$\left|\frac{\partial\psi}{\partial a}\right| \sim \hbar\left|\frac{\partial\psi}{\partial\phi}\right|.$$

Then we can expand the WD equation in powers of $\hbar$, substituting the potential $V(\phi)$,

$$-\frac{G}{3\pi}\left(S'^2 + \frac{9\pi^2 a^2}{4G^2} - \frac{6\pi^3}{G} a^4 V_0\right)\psi \qquad (5.43)$$

$$-\frac{\mathrm{i}G\hbar}{3\pi}\left(S'' + 2S'\partial_a\right)\psi + \pi^2 a^4\left(-\hbar^2\partial_\phi^2 + m^2\phi^2\right)\psi = 0. \qquad (5.44)$$

The first line (5.43) above is of zeroth order in $\hbar$ and thus must be satisfied separately. This is an equation determining the behavior of the metric variable a. Its solution is

$$S(a) = \pm \int \mathrm{d}a \sqrt{\frac{6\pi^3}{G} a^4 V_0 - \frac{9\pi^2 a^2}{4G^2}}.$$

To visualize the physical interpretation of this solution, we note that Eq. (5.43) is formally analogous to the semiclassical form of the stationary Schrödinger equation for a one-dimensional particle in a potential

$$U(a) \equiv \frac{9\pi^2 a^2}{8G^2} - \frac{3\pi^3}{G} a^4 V_0.$$

This "particle" can tunnel from the initial state at $a = 0$ to the value

$$a = a_0 \equiv \sqrt{\frac{3}{8\pi G V_0}}.$$

The regime $a > a_0$ corresponds to a classical motion of the particle with the velocity $da/d\tau = \sqrt{2U(a)}$, while for $0 < a < a_0$ the motion is classically forbidden since $U(a) < 0$. The tunneling process is interpreted as the *creation of a closed universe* having the initial size a_0. Since the assumed initial state $a = 0$ corresponds to a sphere of zero radius or, in other words, to the absence of space, this process is also called **creation of a universe from nothing**.

Consider now the second part (5.44) of the WD equation. In the semiclassical regime, $S(a)$ is a slow-changing function of a, so we can disregard $|S''| \ll S'^2$. Since a is a well-defined semiclassical variable in the regime $a > a_0$, we may define the "time" τ according to the heuristic picture of a moving "particle" with the coordinate a. It is convenient to define

$$\mathrm{d}\tau = \frac{3\pi^3}{2G} \frac{a^4 \mathrm{d}a}{\sqrt{2U(a)}} = \frac{3\pi^3}{2G} \frac{a^4 \mathrm{d}a}{S'(a)},$$

and express a through τ in the range $a > a_0$. Then Eq. (5.44) can be rewritten as

$$\mathrm{i}\hbar \frac{\partial \psi}{\partial \tau} = \left(-\hbar^2 \partial_\phi^2 + m^2 \phi^2\right) \psi \equiv \hat{H}_\psi \psi.$$

This is a familiar *time-dependent* Schrödinger equation for a quantum system with the Hamiltonian $\hat{H}_\psi$.

In this way, we find that the time-independent "wave function of the universe" $\Psi(a, \phi)$, when properly interpreted in the regime when one of the variables is classical, yields a familiar picture of the evolution of quantum systems with time. Note that in the classically forbidden regime $0 < a < a_0$, no "time" can be defined, and a classical interpretation of the wave function Ψ is impossible.

6 Tetrad methods

The tetrad formalism is covered in the books [21, 33, 36] with varying degrees of clarity and detail.

Note: The material on vector bundles is standard **but needs to be expanded. Also, there should be more material on tetrad formalism: derivation of Einstein equation, as well as an explanation of tetrad formalism in terms of vector bundles (making geometric sense of tetrad indices, "covariant" exterior differential, etc.).**

6.1 Tetrad formalism

In the standard approach, General Relativity is formulated as a "theory of the metric." In other words, the main variable is the metric tensor g, while the connection and the curvature are expressed through the metric. An alternative formulation is through orthonormal bases of vector fields (called "tetrads"). This formulation is convenient for many calculations, and also serves to connect gravity with spinor field theory. Some approaches to quantum gravity are based on the tetrad formulation.

For convenience, I work in a four-dimensional manifold having a metric with Lorentzian signature $(+---)$. Generalizations to other dimensions and signatures are straightforward.

6.1.1 Tetrads

If a metric tensor g on a manifold $\mathcal{M}$ is specified, one can always choose an orthonormal basis $\{\mathbf{e}_0(p), \mathbf{e}_1(p), \mathbf{e}_2(p), \mathbf{e}_3(p)\}$ in each tangent space $T_p\mathcal{M}$. One of the tetrad vectors must be timelike and the other three must be spacelike (because the metric g has a Lorentzian signature). It is a convention to choose the vector $\mathbf{e}_0$ timelike and the other three spacelike, so that the scalar product matrix is

$$g(\mathbf{e}_a, \mathbf{e}_b) = \eta_{ab} \equiv \operatorname{diag}(1, -1, -1, -1),$$

where η_{ab} is understood as a fixed matrix (which is numerically the same as the standard metric in Minkowski spacetime). Note that the Latin indices label the basis vectors and run (for convenience) from 0 to 3. In this section I will indicate summations over these indices explicitly. I will work in a four-dimensional spacetime

Essentially we have chosen a particular set of vector fields $\{\mathbf{e}_a(p)\}$, defined at every point p. Such an orthonormal set of vector fields $\{\mathbf{e}_a(p)\}$ is called a **tetrad** or a **vierbein** in case of four-dimensional manifolds, and a **frame field** or **vielbein** in any number of dimensions.[1]

Remark: With a slight strain on the terminology, I call the basis $\{\mathbf{e}_a\}$ "orthonormal" even though $g(\mathbf{e}_a, \mathbf{e}_a) = -1$ for $a = 1, 2, 3$ as it must be since the metric has Lorentzian signature. ■

[1] The German words "vierbein" and "vielbein" are pronounced approximately as the English nonwords "fear-bine" and "feel-bine" respectively. The corresponding French term is "repère," pronounced close to "repair."

In a local coordinate system $\{x^\mu\}$ where the metric is specified as a set of components $g_{\mu\nu}(p)$, the components $e^a_\mu(p)$ of the tetrad may be found by the following procedure. At a fixed spacetime point p, consider the matrix $g_{\mu\nu}(p)$ as a bilinear form. This bilinear form may be diagonalized by a linear coordinate transformation, according to standard procedures of linear algebra. Suppose that $\{t, x, y, z\}$ are new coordinates in which the metric is diagonal at the point p, e.g.

$$g_{\mu\nu}(p) = \operatorname{diag}(A, -B, -C, -D) \equiv \begin{pmatrix} A & 0 & 0 & 0 \\ 0 & -B & 0 & 0 \\ 0 & 0 & -C & 0 \\ 0 & 0 & 0 & -D \end{pmatrix}$$

with $A, B, C, D > 0$. Then it is easy to write an orthogonal basis in the tangent space $T_p\mathcal{M}$, for instance

$$\mathbf{e}_0(p) = \frac{1}{\sqrt{A}}\partial_t, \quad \mathbf{e}_1(p) = \frac{1}{\sqrt{B}}\partial_x,$$
$$\mathbf{e}_2(p) = \frac{1}{\sqrt{C}}\partial_y, \quad \mathbf{e}_3(p) = \frac{1}{\sqrt{D}}\partial_z.$$

In this way, obviously generalizable to any number of dimensions, a vielbein may be determined at every point p of the manifold.

Conversely, if a tetrad $\{\mathbf{e}_a(p)\}$ is given at every point p of a manifold $\mathcal{M}$ then we may recover the metric tensor g explicitly in the following way. First we need to compute the four 1-forms $\boldsymbol{\theta}^a$ $(a = 0, 1, 2, 3)$ that comprise the dual basis to $\{\mathbf{e}_a\}$. Here and below we denote vectors and forms by boldface symbols. (It is convenient to write "a" as an upper index in $\boldsymbol{\theta}^a$, even though we presently do not consider it a tensor index.)

Construction of the dual basis: We work in the tangent space $T_p\mathcal{M}$ at a point p of a manifold $\mathcal{M}$. Since $\{\mathbf{e}_a(p)\}$ is a basis in $T_p\mathcal{M}$, any vector $\mathbf{v} \in T_p\mathcal{M}$ is uniquely decomposed as a linear combination

$$\mathbf{v} = \sum_a \lambda^a \mathbf{e}_a,$$

where λ^a are some numbers. When the basis $\{\mathbf{e}_a\}$ is fixed, the numbers λ^a are linear functions of $\mathbf{v}$, and thus are 1-forms which we denote $\boldsymbol{\theta}^a$; in other words, $\lambda^a \equiv \boldsymbol{\theta}^a(\mathbf{v})$. These 1-forms $\boldsymbol{\theta}^a$ yield the components of a vector in the basis $\{\mathbf{e}_a\}$, so that

$$\mathbf{v} = \sum_a \boldsymbol{\theta}^a(\mathbf{v})\mathbf{e}_a.$$

This relationship holds for any vector $\mathbf{v}$; sometimes this is written as a decomposition of the identity operator,

$$\hat{\mathbf{1}} = \sum_a \mathbf{e}_a \otimes \boldsymbol{\theta}^a. \tag{6.1}$$

The set of 1-forms $\{\boldsymbol{\theta}^a\}$ is called the **dual tetrad** and is a basis in the dual tangent space $T^*\mathcal{M}(p)$. Sometimes one calls $\boldsymbol{\theta}^a$ simply the "tetrad" for brevity.

In any case, it is straightforward to pass from $\{\mathbf{e}_a\}$ to $\{\theta^a\}$ and back. In a local coordinate system, the dual basis has components θ^a_μ which comprise a 4×4 matrix satisfying

$$\sum_a \theta^a_\mu v^\mu e^\nu_a = v^\nu$$

for any v^μ. Therefore,

$$\sum_a \theta^a_\mu e^\nu_a = \delta^\nu_\mu,$$

so the matrix of components θ^a_μ can be computed as the inverse matrix to e^μ_a. ■

Once the 1-forms $\boldsymbol{\theta}^a$, $a = 0,1,2,3$, are computed, the metric tensor g is expressed as

$$g(\mathbf{u},\mathbf{v}) = g\left(\sum_a \boldsymbol{\theta}^a(\mathbf{u})\mathbf{e}_a, \sum_b \boldsymbol{\theta}^b(\mathbf{v})\mathbf{e}_b\right) = \sum_{a,b}\eta_{ab}\boldsymbol{\theta}^a(\mathbf{u})\boldsymbol{\theta}^b(\mathbf{v}). \tag{6.2}$$

Rewritten in a more condensed notation, the above formula becomes

$$g = \sum_{a,b}\eta_{ab}\boldsymbol{\theta}^a\otimes\boldsymbol{\theta}^b.$$

In components,

$$g_{\mu\nu} = \sum_{a,b}\eta_{ab}\theta^a_\mu\theta^b_\nu.$$

This is an explicit formula that expresses the metric tensor g through the dual tetrad. Analogously, the inverse metric g^{-1} is expressed through the tetrad vectors $\{\mathbf{e}_a\}$ as

$$g^{-1} = \sum_{a,b}\eta^{ab}\mathbf{e}_a\otimes\mathbf{e}_b.$$

Thus, in order to define a metric structure on a manifold, one may specify a tetrad $\{\mathbf{e}_a\}$ or alternatively a dual tetrad $\{\boldsymbol{\theta}^a\}$ instead of specifying the metric tensor g.

Remarks:

- The tetrad $\{\mathbf{e}_a(p)\}$ can be viewed as a map from a fixed, four-dimensional spacetime $\mathbb{R}^4$ to the tangent space $T_p\mathcal{M}$. A vector

  $$v^a \equiv \left\{v^0, v^1, v^2, v^3\right\} \in \mathbb{R}^4$$

  is mapped to the tangent vector

  $$\mathbf{v} \equiv \sum_a v^a \mathbf{e}_a(p) \in T_p\mathcal{M}.$$

  This map $v^a \to \mathbf{v}$ is compatible with the scalar product in $\mathbb{R}^4$ defined by the matrix η_{ab} and the scalar product in $T_p\mathcal{M}$ defined by the metric g, namely

  $$\sum_{a,b}\eta_{ab}u^a v^b = g(\mathbf{u},\mathbf{v}) \text{ if } u^a \to \mathbf{u},\ v^a \to \mathbf{v}.$$

  Therefore one refers to the abstract space $\mathbb{R}^4$ and the "metric" η_{ab} as the **fiducial Minkowski spacetime** and metric. This fiducial flat spacetime can be interpreted as the instantaneous reference frame of a certain observer. This observer's four-velocity instantaneously coincides with the timelike vector $\mathbf{e}_0(p)$ and the observer's spatial axes are chosen along the directions $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$.

- There are infinitely many possible tetrads corresponding to one and the same metric g. A given tetrad $\{\mathbf{e}_a\}$ can be transformed via a Lorentz rotation,

  $$\mathbf{e}_a(p) \to \tilde{\mathbf{e}}_a(p) \equiv \sum_b \Lambda^b_a(p)\mathbf{e}_b(p),$$

  without changing the metric g. Here, $\Lambda^b_a(p)$ is a matrix representing a Lorentz transformation in the fiducial Minkowski spacetime. This matrix $\Lambda^b_a(p)$ may be chosen arbitrarily at every point p and thus is called a **local** Lorentz transformation. Conversely, *any* two orthonormal bases $\{\mathbf{e}_a\}$ and $\{\tilde{\mathbf{e}}_a\}$ are connected by a Lorentz transformation. Therefore, the local Lorentz transformations $\Lambda^b_a(p)$ represent the entire freedom of choosing the tetrad field for a fixed metric g.

- A tetrad is essentially a set of four vector fields which must be linearly independent at every point. Since it is impossible to have globally nonzero vector fields on some manifolds (e.g. on a 2-sphere), the tetrad fields $\{\mathbf{e}_a(p)\}$ might have to be defined only locally. In other words, different tetrad fields will have to be defined within different charts covering the manifold.

- It is clear that the 1-forms $\boldsymbol{\theta}^a$ can be expressed through $\mathbf{e}_a$ and the metric as follows,

  $$\boldsymbol{\theta}^a(\mathbf{v}) \equiv \boldsymbol{\theta}^a \circ \mathbf{v} = \eta_{aa}g(\mathbf{e}_a,\mathbf{v}) = \sum_b \eta^{ab}g(\mathbf{e}_b,\mathbf{v}),$$

  where $\mathbf{v}$ is an arbitrary vector. In the notation of Sec. 1.5.5, we may rewrite this as

  $$\boldsymbol{\theta}^a = \eta_{aa}\hat{g}\mathbf{e}_a = \sum_b \eta^{ab}\hat{g}\mathbf{e}_b.$$

  (There is no summation over a in the expressions $\eta_{aa}\hat{g}\mathbf{e}_a$ above.) Often it is convenient to raise and lower the Latin (tetrad) indices using the fiducial Minkowski metric η^{ab}. Thus, one defines

  $$\boldsymbol{\theta}_a \equiv \eta_{aa}\boldsymbol{\theta}^a = \sum_b \eta_{ab}\boldsymbol{\theta}^b \tag{6.3}$$

  and writes equations such as

  $$g(\mathbf{e}_a,\mathbf{v}) = \boldsymbol{\theta}_a(\mathbf{v}). \tag{6.4}$$

  However, I stress that the dual tetrad $\{\boldsymbol{\theta}^a\}$ can be calculated from $\{\mathbf{e}_a\}$ *without* using this formula. The metric tensor g is already fixed once we declare that a given basis $\{\mathbf{e}_a\}$ is orthonormal.

- The dual tetrad $\{\boldsymbol{\theta}^a\}$ is orthonormal with respect to the *inverse* metric,

  $$g^{-1}(\boldsymbol{\theta}^a,\boldsymbol{\theta}^b) = \eta^{ab}.$$

  Thus $\{\boldsymbol{\theta}^a(p)\}$ is an orthonormal basis in the cotangent space $T^*\mathcal{M}(p)$. ■

Holonomic and nonholonomic. Sometimes the frame basis $\{\mathbf{e}_a\}$ or the dual basis $\{\theta^a\}$ are called **nonholonomic** bases, while the coordinate basis $\{\partial/\partial x^\mu\}$ is called **holonomic**. This terminology will not be used in the present text; instead I call $\{\mathbf{e}_a\}$ an orthonormal frame basis, a tetrad, and a vielbein.

The term "holonomic" seems to come from mechanics textbooks. In theoretical mechanics, there is a notion of holonomic

and nonholonomic constraint equations. A **constraint equation** is a relation of the form

$$\sum_j \frac{dq^j}{dt} A_j(\mathbf{q}) = 0,$$

where $\mathbf{q} \equiv \{q^j\} \in \mathbb{R}^n$ is a time-dependent vector of "generalized coordinates" and $A_j(\mathbf{q})$ are some coefficients. A constraint equation is called **holonomic** if it can be rewritten as a total time derivative of a function, i.e. if there exists a function $\alpha(\mathbf{q})$ such that the constraint equation is simply $d\alpha(\mathbf{q})/dt = 0$, in other words if

$$A_j = \frac{\partial \alpha(\mathbf{q})}{\partial q^j}, \quad j = 1, ...n.$$

One can reformulate these statements more transparently by introducing the auxiliary 1-form $\mathbf{A} \equiv \sum_j A_j \mathrm{d}q^j$. Then the constraint equation is rewritten as $\mathbf{A} \circ \dot{\mathbf{q}} = 0$, where $\dot{\mathbf{q}}$ is the tangent vector to the trajectory $\mathbf{q}(t)$. The constraint is holonomic if there exists a scalar function $\alpha(\mathbf{q})$ such that $\mathbf{A} = \mathrm{d}\alpha$; in the standard terminology, the constraint is holonomic when the 1-form $\mathbf{A}$ is **exact** (equal to a differential of some function). So one calls a basis $\{\boldsymbol{\theta}^a\}$ "holonomic" if there exist functions f^a such that $\boldsymbol{\theta}^a = \mathrm{d}f^a$, and "nonholonomic" otherwise. These functions could serve as a local coordinate system $\{f^0, f^1, f^2, f^3\}$ naturally adapted to the given basis $\{\boldsymbol{\theta}^a\}$.

It is easy to recognize a holonomic basis by checking that $\mathrm{d}\boldsymbol{\theta}^a = 0$ for all a. For example, a basis of 1-forms $\{t^2\mathrm{d}t, x\mathrm{d}x + y\mathrm{d}y, x\mathrm{d}y + y\mathrm{d}x, \mathrm{d}z\}$ is holonomic.

However, when an *orthonormal* basis is holonomic, it means that there exist coordinates $\{x^a\}$ in which the 1-forms $\mathrm{d}x^a$ are orthogonal. In other words, these are Cartesian coordinates and the metric is simply the flat Euclidean (or Minkowski) metric. It is impossible to choose an orthonormal holonomic basis of 1-forms in a nonflat spacetime.

The condition on the basis vectors $\mathbf{e}_a$ that indicates a holonomic basis is mutual commutation. If $\{\boldsymbol{\theta}^a\}$ is a holonomic basis, then there exist a coordinate system $\{x^a\}$ where $\boldsymbol{\theta}^a = \mathrm{d}x^a$. Hence, the dual basis $\{\mathbf{e}_a\}$ will consist of vectors $\mathbf{e}_a = \partial_{x^a}$ and hence $[\mathbf{e}_a, \mathbf{e}_b] = 0$ for all a, b. This is the property that indicates a holonomic basis. In a nonflat spacetime, an *orthonormal* basis $\{\mathbf{e}_a\}$ will be necessarily nonholonomic.

It is true that in almost all cases we will be working in a nonflat spacetime, and hence the vielbein will be nonholonomic. However, to me this is not a sufficient motivation to call $\{\mathbf{e}_a\}$ a nonholonomic basis. What is important for our considerations is the *orthonormality* of the chosen frame basis $\{\mathbf{e}_a\}$ with respect to the metric g. Calling $\{\mathbf{e}_a\}$ merely a nonholonomic basis may create an impression that it is sufficient to select *some* nonholonomic basis, whether orthonormal or not. However, an arbitrary non-orthonormal, nonholonomic basis is much less useful in calculations than an orthonormal basis. For this reason, we call $\{\mathbf{e}_a\}$ an "orhonormal frame basis" (or a **vielbein**) rather than a "nonholonomic basis."

6.1.2 Examples

Example 1: Consider the Schwarzschild metric (1.38). The task is to determine a tetrad basis and a dual tetrad basis.

Since the metric is in a diagonal form, a possible choice of the tetrad is

$$\mathbf{e}_0 = \left(1 - \frac{2M}{r}\right)^{-1/2} \partial_t, \quad \mathbf{e}_1 = \left(1 - \frac{2M}{r}\right)^{1/2} \partial_r,$$
$$\mathbf{e}_2 = \frac{1}{r}\partial_\theta, \quad \mathbf{e}_3 = \frac{1}{r\sin\theta}\partial_\phi.$$

Note that the tetrad is undefined (singular) if $r = 0$, $r = 2M$, or $\sin\theta = 0$ because at these locations the coordinate system (t, r, θ, ϕ) and/or the metric g are singular. This is an example of a tetrad defined only *locally*, that is, only within a certain limited coordinate patch, rather than globally in the entire spacetime. More than one coordinate patch must be used to cover the entire Schwarzschild spacetime.

The dual basis $\{\boldsymbol{\theta}^a\}$ corresponding to the tetrad $\{\mathbf{e}^a\}$ is

$$\boldsymbol{\theta}^0 = \left(1 - \frac{2M}{r}\right)^{1/2} \mathrm{d}t, \quad \boldsymbol{\theta}^1 = \left(1 - \frac{2M}{r}\right)^{-1/2} \mathrm{d}r,$$
$$\boldsymbol{\theta}^2 = r\mathrm{d}\theta, \quad \boldsymbol{\theta}^3 = r\sin\theta \mathrm{d}\phi.$$

■

Example 2: Suppose now that we are given a two-dimensional spacetime with local coordinates $\{t, x\}$ and the frame fields

$$\mathbf{e}_0 = \partial_t, \quad \mathbf{e}_1 = e^{-Ht}\partial_x.$$

The task is to determine the metric tensor g such that this frame is orthonormal.

We can recover the metric by first determining the dual frame,

$$\boldsymbol{\theta}^0 = \mathrm{d}t, \quad \boldsymbol{\theta}^1 = e^{Ht}\mathrm{d}x,$$

and then writing

$$g = \sum_{a,b} \eta_{ab} \boldsymbol{\theta}^a \otimes \boldsymbol{\theta}^b = \mathrm{d}t \otimes \mathrm{d}t - e^{2Ht}\mathrm{d}x \otimes \mathrm{d}x$$
$$\equiv dt^2 - e^{2Ht}dx^2.$$

This metric is the two-dimensional version of de Sitter metric (1.39). ■

6.1.3 Hodge duality

Using the metric g, one determines the Levi-Civita tensor $\boldsymbol{\varepsilon}$, and then one can establish a one-to-one linear map between n-forms and $(4-n)$-forms. This map is called the **Hodge duality** operation. Since this operation is used relatively little in the present text (it rarely yields a computational advantage), I only sketch the construction on examples and list some basic properties of the Hodge duality. [2]

We work with a four-dimensional manifold where the metric g and the Levi-Civita tensor $\boldsymbol{\varepsilon}$ are known. First consider a 1-form ω. The Hodge duality map yields a 3-form ${}^*\omega$ called the **Hodge dual** to ω; this operation is also called the **Hodge star**. The 3-form $*\omega$ is defined by its action on arbitrary vectors $\mathbf{x}, \mathbf{y}, \mathbf{z}$:

$$({}^*\omega) \circ (\mathbf{x}, \mathbf{y}, \mathbf{z}) \equiv \boldsymbol{\varepsilon}(\hat{g}^{-1}\omega, \mathbf{x}, \mathbf{y}, \mathbf{z}).$$

Somewhat more awkwardly, this definition can be rewritten as

$$*\omega = \iota_{\hat{g}^{-1}\omega}\boldsymbol{\varepsilon}.$$

It is clear that $\omega \mapsto *\omega$ is a linear map in ω (if we regard the metric g and the tensor $\boldsymbol{\varepsilon}$ as fixed).

[2] I was motivated to make this excursion into the Hodge duality by reading the lecture notes [13], chapter 6, where I learned about Statement 6.1.3.2(a).

Example: Consider a four-dimensional Minkowski space with the orthonormal basis of 1-forms $\{dt, dx, dy, dz\}$. The Levi-Civita tensor is

$$\varepsilon = dt \wedge dx \wedge dy \wedge dz.$$

Let us compute the Hodge star applied to the 1-form dt. Using the vector $\hat{g}^{-1}dt = \partial_t$, we find

$$*(dt) = \iota_{\partial_t}\varepsilon = dx \wedge dy \wedge dz.$$

Similarly, we have

$$*(dx) = \iota_{\partial_x}\varepsilon = -dt \wedge dy \wedge dz.$$

Heuristically, $*\omega$ must contain the "complement" of ω, so that $\omega \wedge *\omega$ is proportional to ε. ■

Another convenient way to express $*\omega$ is to use an orthonormal basis $\{\mathbf{e}_a\}$. Since

$$\hat{g}^{-1}\omega = \sum_{a,b} \eta^{ab} \left(\iota_{\mathbf{e}_b}\omega\right) \mathbf{e}_a,$$

we can write

$$*\omega = \sum_a \eta^{aa} \left(\iota_{\mathbf{e}_a}\omega\right) \left(\iota_{\mathbf{e}_a}\varepsilon\right).$$

This expression can be also used as a definition of $*\omega$. By construction, this definition is independent of the choice of the basis $\{\mathbf{e}_a\}$, as long as this basis is orthonormal with respect to the fixed metric g and positively oriented with respect to ε. (We merely rewrote a basis-free definition in terms of an arbitrary orthonormal basis.) The basis independence can be made manifest by expressing $*\omega$ through the trace operation as follows,

$$*\omega = \mathrm{Tr}_{(\mathbf{a},\mathbf{b})} \left(\iota_{\mathbf{a}}\omega\right) \left(\iota_{\mathbf{b}}\varepsilon\right).$$

It is clear that this can be used as a definition of $*\omega$ equivalent to the definitions above.

Similarly, a 2-form $\omega(\mathbf{a}, \mathbf{b})$ is converted to its Hodge dual $*\omega$, which is also a 2-form. This 2-form can be defined using an orthonormal basis $\{\mathbf{e}_a\}$ as follows,

$$*\,\omega \equiv \frac{1}{2!}\sum_{a,b} \eta^{aa}\eta^{bb} \left(\iota_{\mathbf{e}_b}\iota_{\mathbf{e}_a}\omega\right) \left(\iota_{\mathbf{e}_b}\iota_{\mathbf{e}_a}\varepsilon\right). \tag{6.5}$$

As before, this definition is independent of the choice of the orthonormal basis $\{\mathbf{e}_a\}$. This can also be seen with the following trick. Any 2-form ω can be decomposed into a linear combination of exterior products of 1-forms. Since the Hodge duality is a linear map, we only need to consider a single exterior product $\omega = \phi \wedge \chi$, where ϕ and χ are 1-forms. The 2-form $*(\phi \wedge \chi)$ is then expressed as

$$(*(\phi \wedge \chi)) \circ (\mathbf{x}, \mathbf{y}) = \varepsilon(\hat{g}^{-1}\phi, \hat{g}^{-1}\chi, \mathbf{x}, \mathbf{y}).$$

It is straightforward to verify that this expression is equivalent to Eq. (6.5). When the definition of $*\omega$ is written in this basis-free manner, it becomes manifest that $*\omega$ depends only on the metric g but not on the choice of the basis $\{\mathbf{e}_a\}$. However, calculations using an explicit basis $\{\mathbf{e}_a\}$ are much more convenient since one does not need to decompose all n-forms into linear combinations of exterior products of 1-forms. Sometimes also the index-free trace notation is convenient,

$$*\omega \equiv \frac{1}{2!}\mathrm{Tr}_{(\mathbf{a},\mathbf{a}')(\mathbf{b},\mathbf{b}')} \left(\iota_{\mathbf{b}'}\iota_{\mathbf{a}'}\omega\right) \left(\iota_{\mathbf{b}}\iota_{\mathbf{a}}\varepsilon\right).$$

So far we have seen the definitions of $*\omega$ for 1-forms and 2-forms. To express the Hodge dual of an n-form Ω, one can build an expression analogous to Eq. (6.5) but with n sets of vectors and the factor $1/n!$ in front:

$$\begin{aligned} *\Omega &\equiv \frac{1}{n!}\sum_{a_1,\ldots,a_n} \eta^{a_1 a_1}\ldots\eta^{a_n a_n} \left(\iota_{\mathbf{e}_{a_n}}\ldots\iota_{\mathbf{e}_{a_1}}\Omega\right) \left(\iota_{\mathbf{e}_{a_n}}\ldots\iota_{\mathbf{e}_{a_1}}\varepsilon\right) \\ &= \frac{1}{n!}\mathrm{Tr}_{(\mathbf{a}_1,\mathbf{b}_1)\ldots(\mathbf{a}_n,\mathbf{b}_n)} \left(\iota_{\mathbf{a}_n}\ldots\iota_{\mathbf{a}_1}\Omega\right) \left(\iota_{\mathbf{b}_n}\ldots\iota_{\mathbf{b}_1}\varepsilon\right). \end{aligned}$$

Here we wrote the arguments $\mathbf{e}_{a_j}$ in the reverse order for convenience; recall that

$$\iota_{\mathbf{e}_{a_n}}\ldots\iota_{\mathbf{e}_{a_1}}\Omega \equiv \Omega(\mathbf{e}_{a_1}, \ldots, \mathbf{e}_{a_n}).$$

As before, this definition of $*\Omega$ depends only on the metric g and on ε but not on the choice of the orthonormal basis $\{\mathbf{e}_a\}$. In four dimensions, the Hodge star establishes a linear map from 0-forms (scalars) to 4-forms, from 1-forms to 3-forms, etc. In particular, the 4-form ε is mapped into a scalar. The numerical factor $1/n!$ leads to the relation $*\varepsilon = -1$:

$$\begin{aligned} *\varepsilon &= \frac{1}{4!}\sum_{a,b,c,d} \eta^{aa}\ldots\eta^{dd} \left(\iota_{\mathbf{e}_d}\ldots\iota_{\mathbf{e}_a}\varepsilon\right) \left(\iota_{\mathbf{e}_d}\ldots\iota_{\mathbf{e}_a}\varepsilon\right) \\ &= \eta^{00}\eta^{11}\eta^{22}\eta^{33} = \det\eta_{ab} = -1. \end{aligned}$$

Example: The Hodge dual to a scalar function ("0-form") f is a 4-form $*f = f\varepsilon$; taking the Hodge dual again, one gets $*(f\varepsilon) = -f$. Therefore, $**f = -f$ for scalar functions f.

Let $\{\theta^a\}$ $(a = 0, 1, 2, 3)$ be an orthonormal basis of 1-forms; then we have

$$\begin{aligned} *\theta^0 &= \iota_{\mathbf{e}_0}\left(\theta^0 \wedge \theta^1 \wedge \theta^2 \wedge \theta^3\right) = \theta^1 \wedge \theta^2 \wedge \theta^3; \\ *\theta^3 &= \iota_{-\mathbf{e}_3}\left(\theta^0 \wedge \theta^1 \wedge \theta^2 \wedge \theta^3\right) = \theta^0 \wedge \theta^1 \wedge \theta^2; \\ *(\theta^0 \wedge \theta^1) &= \iota_{-\mathbf{e}_1}\iota_{\mathbf{e}_0}\left(\theta^0 \wedge \theta^1 \wedge \theta^2 \wedge \theta^3\right) = -\theta^2 \wedge \theta^3; \\ *(\theta^2 \wedge \theta^3) &= \theta^0 \wedge \theta^1; \quad *(\theta^1 \wedge \theta^2 \wedge \theta^3) = -\theta^0; \quad \text{etc.} \end{aligned}$$

In four dimensions, one can derive a more convenient formula for the Hodge dual to a 4-form Ω by noting that $\varepsilon(\mathbf{e}_a, \ldots, \mathbf{e}_d) \neq 0$ only when all the indices a, b, c, d are different:

$$\begin{aligned} *\Omega &= \frac{1}{4!}\sum_{a,b,c,d} \eta^{aa}\ldots\eta^{dd} \left(\iota_{\mathbf{e}_d}\ldots\iota_{\mathbf{e}_a}\Omega\right) \left(\iota_{\mathbf{e}_d}\ldots\iota_{\mathbf{e}_a}\varepsilon\right) \\ &= (\det\eta_{ab})\,\Omega(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4) = -\Omega(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4). \end{aligned}$$

It is clear that this formula extends from four-dimensional to N-dimensional space,

$$*\,\Omega = (\det\eta_{ab})\,\Omega(\mathbf{e}_1, \ldots, \mathbf{e}_N), \tag{6.6}$$

where η_{ab} is the canonical form of the N-dimensional metric (i.e. the form where the metric is diagonal and has only ± 1 on the diagonal). ■

The action of the Hodge star on basis 1-forms $\{\theta^a\}$, as illustrated by the examples above, can be summarized by the

following formulas:

$$*1 \equiv \varepsilon = \frac{1}{4!}\sum_{a,b,c,d}\varepsilon^{abcd}\theta_a \wedge \theta_b \wedge \theta_c \wedge \theta_d,$$

$$*\theta^a = \frac{1}{3!}\sum_{b,c,d}\varepsilon^{abcd}\theta_b \wedge \theta_c \wedge \theta_d,$$

$$*(\theta^a \wedge \theta^b) = \frac{1}{2!}\sum_{c,d}\varepsilon^{abcd}\theta_c \wedge \theta_d,$$

$$*(\theta^a \wedge \theta^b \wedge \theta^c) = \sum_{d}\varepsilon^{abcd}\theta_d,$$

where ε^{abcd} is not a tensor but a totally antisymmetric array of numbers (the Levi-Civita symbol) defined such that $\varepsilon^{0123} = 1$. (It is important to note that the 1-forms θ_a differ from θ^a by the sign factor η^{aa}.)

The example above shows that $**f = -f$ for a scalar f. Since $\varepsilon = *1$, we also have $**\varepsilon = -\varepsilon$. A similar property holds for any n-form ω.

Statement 6.1.3.1: For any n-form ω in N-dimensional manifold with canonical metric η_{ab}, one has

$$**\omega = (\det \eta_{ab})\,(-1)^{(N-n)n}\,\omega.$$

(Proof on page 180.) ■

It is not particularly easy to perform calculations with the Hodge star operation. For instance, there are no general identities relating the Hodge star to other operations, such as the exterior differential or the Lie derivative; so in general $*(\omega \wedge \psi) \neq *\omega \wedge *\psi$, $\mathrm{d}*\omega \neq *\mathrm{d}\omega$, $\mathcal{L}_{\mathbf{x}}*\omega \neq *\mathcal{L}_{\mathbf{x}}\omega$, etc. When doing calculations with the Hodge star, one needs to use either an explicit basis $\{\mathbf{e}_a\}$, or the index-free trace formalism.

It follows from the example above that

$$\theta^0 \wedge *\theta^0 = -\varepsilon; \quad (\theta^0 \wedge \theta^1) \wedge *(\theta^0 \wedge \theta^1) = -\varepsilon; \quad \text{etc.}$$

On the other hand, it is easy to see that

$$\theta^0 \wedge *\theta^1 = 0; \quad (\theta^0 \wedge \theta^1) \wedge *(\theta^0 \wedge \theta^2) = 0; \quad \text{etc.}$$

It seems that the combination $\omega_1 \wedge *\omega_2$ is nonzero on basis n-forms only when the two forms are equal. This combination is indeed a useful function of two n-forms. The following statement lists some properties of this function.

Statement 6.1.3.2: **(a)** If ω_1 and ω_2 are two n-forms in an N-dimensional manifold, one can compute the N-form $\omega_1 \wedge (*\omega_2)$ that is symmetric under interchange of ω_1 and ω_2:

$$\omega_1 \wedge *\omega_2 = \omega_2 \wedge *\omega_1.$$

Thus the scalar $*(\omega_1 \wedge *\omega_2)$ is a symmetric bilinear function in the space of n-forms. This bilinear function can be viewed as a natural scalar product in the space of n-forms. In particular, if $\{\theta^a\}$ is an orthonormal basis of 1-forms, then an orthonormal basis in the space of n-forms is the set of exterior products

$$\{\theta^{a_1} \wedge ... \wedge \theta^{a_n}\},$$

with all possible sets of indices $\{a_1, ..., a_n\}$ for which these exterior products are nonzero.

(b) If ω_1, ω_2 are two n-forms, the combination $\omega_1 \wedge *\omega_2$ is proportional to the volume element with the following coefficient,

$$\omega_1 \wedge *\omega_2 = \frac{\mathrm{Vol}}{n!}\mathrm{Tr}_{(\mathbf{a}_1,\mathbf{b}_1)...(\mathbf{a}_n,\mathbf{b}_n)}\omega_1(\mathbf{a}_1, ..., \mathbf{a}_n)\omega_2(\mathbf{b}_1, ..., \mathbf{b}_n). \tag{6.7}$$

(c) Suppose ω_1 and ω_2 are 1-forms corresponding to vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, i.e. $\omega_1 = \hat{g}\mathbf{v}_1$, $\omega_2 = \hat{g}\mathbf{v}_2$.Then the scalar product defined in part **(a)** coincides with the natural scalar product of 1-forms given by the inverse metric g^{-1}. More precisely,

$$\omega_1 \wedge *\omega_2 = g^{-1}(\omega_1, \omega_2)\,\mathrm{Vol} = g(\mathbf{v}_1, \mathbf{v}_2)\,\mathrm{Vol}.$$

(d) If Ω is a 2-form and ω_1, ω_2 are 1-forms corresponding to vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, one has

$$\Omega \wedge *(\omega_1 \wedge \omega_2) = \Omega(\mathbf{v}_1, \mathbf{v}_2)\,\mathrm{Vol}.$$

(Proof on page 180.) ■

Remark: It is interesting to note what happens if ω_1 is an n_1-form and ω_2 is an n_2-form with $n_1 \neq n_2$. In that case, $\omega_1 \wedge *\omega_2$ is an $(n_1 + N - n_2)$-form while $\omega_2 \wedge *\omega_1$ is an $(n_2 + N - n_1)$-form. If $n_1 \neq n_2$, either $n_1 + N - n_2 > N$ or $n_2 + N - n_1 > N$. Hence, one of the expressions $\omega_1 \wedge *\omega_2$ or $\omega_2 \wedge *\omega_1$ is identically zero as an n-form with $n > N$, while the other expression is not necessarily zero. It is clear that Statement 6.1.3.2 cannot be extended to this case. ■

Statement 6.1.3.2: If Ω is a 2-form and ω is the 1-form corresponding to the vector $\mathbf{v} = \hat{g}^{-1}\omega$, then

$$\omega \wedge *\Omega = -*(\iota_{\mathbf{v}}\Omega), \tag{6.8}$$

where both sides are 3-forms.

Hint: Use Statement 6.1.3.2. (Proof on page 181.) ■

Practice problem: Verify Eq. (6.8) for $\Omega = \theta^0 \wedge \theta^1$, $\omega = \theta^1$, and for $\Omega = \theta^2 \wedge \theta^3$, $\omega = \theta^2$. ■

6.1.4 Levi-Civita connection

Given a metric g, one may determine the Levi-Civita connection ∇ explicitly (see Eq. (1.45) in Sec. 1.6.6). It turns out that the Levi-Civita connection can be expressed directly through the tetrad. We begin by assuming that the metric tensor g and the Levi-Civita connection ∇ are known, and then derive the formula for the covariant derivative of the tetrad field. Subsequently, it will be clear how to recover the connection from the tetrad without explicitly involving the metric g.

Given a tetrad $\{\mathbf{e}_a\}$, we would like to be able to compute the covariant derivative $\nabla_{\mathbf{u}}\mathbf{v}$ for arbitrary vector fields $\mathbf{u}, \mathbf{v}$. Since $\{\mathbf{e}_a\}$ is a basis, every vector field $\mathbf{v}$ can be expressed through $\{\mathbf{e}_a\}$ as

$$\mathbf{v} = \sum_a v^a \mathbf{e}_a,$$

where $v^a \equiv v^a(p)$ is a set of four *scalar* functions (the Latin index is *not* a tensor index). Therefore it is sufficient to be able to calculate $\nabla_{\mathbf{u}}\mathbf{v}$ for $\mathbf{u} = \mathbf{e}_a$ and $\mathbf{v} = \mathbf{e}_b$, where $a, b = 0, 1, 2, 3$. According to Eq. (1.45), we have

$$(\nabla_{\mathbf{x}}\mathbf{y}, \mathbf{z}) = \frac{1}{2}\big(\mathbf{x} \circ g(\mathbf{y}, \mathbf{z}) + \mathbf{y} \circ g(\mathbf{x}, \mathbf{z}) - \mathbf{z} \circ g(\mathbf{x}, \mathbf{y}) - g(\mathbf{x}, [\mathbf{y}, \mathbf{z}]) - g(\mathbf{y}, [\mathbf{x}, \mathbf{z}]) + g(\mathbf{z}, [\mathbf{x}, \mathbf{y}])\big).$$

This formula enables us to calculate $g(\nabla_{\mathbf{e}_a}\mathbf{e}_b, \mathbf{e}_c)$ if we can evaluate terms such as $\mathbf{e}_a \circ g(\mathbf{e}_b, \mathbf{e}_c)$ and $g(\mathbf{e}_a, [\mathbf{e}_b, \mathbf{e}_c])$. Since $g(\mathbf{e}_b, \mathbf{e}_c) = \eta_{bc}$ is (by construction) a fixed numerical matrix, the directional derivatives of η_{ab} are zero, so $\mathbf{e}_a \circ g(\mathbf{e}_b, \mathbf{e}_c) = 0$. Thus we have

$$g(\nabla_{\mathbf{e}_a}\mathbf{e}_b, \mathbf{e}_c) = \frac{g(\mathbf{e}_c, [\mathbf{e}_a, \mathbf{e}_b]) - g(\mathbf{e}_a, [\mathbf{e}_b, \mathbf{e}_c]) - g(\mathbf{e}_b, [\mathbf{e}_a, \mathbf{e}_c])}{2}.$$

We may rewrite this formula using the *lower-index* version of the dual tetrad $\boldsymbol{\theta}_c$ [see Eqs. (6.3) and (6.4)] as

$$\boldsymbol{\theta}_c \circ \nabla_{\mathbf{e}_a} \mathbf{e}_b = \frac{\boldsymbol{\theta}_c \circ [\mathbf{e}_a, \mathbf{e}_b] - \boldsymbol{\theta}_a \circ [\mathbf{e}_b, \mathbf{e}_c] - \boldsymbol{\theta}_b \circ [\mathbf{e}_a, \mathbf{e}_c]}{2}. \quad (6.9)$$

Since $\boldsymbol{\theta}^c \circ \mathbf{v}$ is the c-th component of a vector $\mathbf{v}$ in the basis $\{\mathbf{e}_a\}$, the formula (6.9) completely determines the vector $\nabla_{\mathbf{e}_a} \mathbf{e}_b$ if the basis vectors $\mathbf{e}_a$ and their commutators $[\mathbf{e}_a, \mathbf{e}_b]$ are known.

We point out that the essential information required for computing the covariant derivative is the commutators $[\mathbf{e}_a, \mathbf{e}_b]$. In general, these commutators can be expressed as

$$[\mathbf{e}_a, \mathbf{e}_b] = \sum_c C^c{}_{ab} \mathbf{e}_c,$$

where $C^c{}_{ab} = -C^c{}_{ba}$ are some scalar coefficients (which are, of course, functions of the spacetime point p). The coefficients $C^c{}_{ab}$ can be explicitly computed from $\{\mathbf{e}_a\}$ without need to involve the metric g; recall that the commutator $[\mathbf{u}, \mathbf{v}]$ is a geometric operation defined independently of the metric. Then one has complete information about the Levi-Civita connection ∇ through Eq. (6.9). The formula (6.9) can be rewritten as

$$\eta_{cc} \boldsymbol{\theta}^c \circ \nabla_{\mathbf{e}_a} \mathbf{e}_b \equiv \boldsymbol{\theta}_c \circ \nabla_{\mathbf{e}_a} \mathbf{e}_b = \frac{1}{2} \left(C_{c\,ab} - C_{b\,ac} - C_{a\,bc} \right), \quad (6.10)$$

where $C_{c\,ab}$ is obtained from $C^c{}_{ab}$ by lowering the first index through η_{ac}. Alternatively, we may write

$$\nabla_{\mathbf{e}_a} \mathbf{e}_b = \frac{1}{2} \sum_c \left(C_{c\,ab} - C_{b\,ac} - C_{a\,bc} \right) \eta_{cc} \mathbf{e}_c. \quad (6.11)$$

At this point, one can compute covariant derivatives $\nabla_{\mathbf{u}} \mathbf{v}$ of arbitrary vector fields $\mathbf{u}, \mathbf{v}$ by expressing them through the basis $\{\mathbf{e}_a\}$ and using the Leibnitz rule.

Example: We have a two-dimensional spacetime with local coordinates $\{t, x\}$ and the given orthonormal frame

$$\mathbf{e}_0 = \partial_t, \quad \mathbf{e}_1 = e^{-Ht} \partial_x.$$

The task is to compute $\nabla_{\mathbf{u}} \mathbf{v}$, where $\mathbf{u}$ and $\mathbf{v}$ are ∂_t or ∂_x.

Let us first determine $\nabla_{\mathbf{e}_a} \mathbf{e}_b$ for all a, b. We begin by calculating the commutators $[\mathbf{e}_a, \mathbf{e}_b]$. The only nontrivial commutator is

$$[\mathbf{e}_0, \mathbf{e}_1] = -H \mathbf{e}_1.$$

It follows that $C^1{}_{01} = -C^1{}_{10} = -H$ are the only nonzero coefficients among $C^c{}_{ab}$. Lowering the first index of $C^c{}_{ab}$, we find $C_{1\,01} = -C_{1\,10} = H$. Then we use Eq. (6.11) and compute

$$\begin{aligned}
\nabla_{\mathbf{e}_0} \mathbf{e}_0 &= 0, \quad \text{(no nonzero } C_{c\,ab}\text{)} \\
\nabla_{\mathbf{e}_0} \mathbf{e}_1 &= \frac{1}{2} \left(C_{1\,01} - C_{1\,01} \right) \eta_{11} \mathbf{e}_1 = 0, \\
\nabla_{\mathbf{e}_1} \mathbf{e}_0 &= \frac{1}{2} \left(C_{1\,10} - C_{1\,01} \right) \eta_{11} \mathbf{e}_1 = H \mathbf{e}_1, \\
\nabla_{\mathbf{e}_1} \mathbf{e}_1 &= \frac{1}{2} \left(-C_{1\,10} - C_{1\,10} \right) \eta_{00} \mathbf{e}_0 = H \mathbf{e}_0.
\end{aligned}$$

Finally, we use the Leibnitz rule:

$$\begin{aligned}
\nabla_{\partial_t} \partial_x &= \nabla_{\mathbf{e}_0} \left(e^{Ht} \mathbf{e}_1 \right) = \left(\mathbf{e}_0 \circ e^{Ht} \right) \mathbf{e}_1 + e^{Ht} \nabla_{\mathbf{e}_0} \mathbf{e}_1 \\
&= \left(\partial_t e^{Ht} \right) e^{-Ht} \partial_x = H \partial_x.
\end{aligned}$$

Similarly, we find $\nabla_{\partial_t} \partial_t = 0$ and $\nabla_{\partial_x} \partial_x = H e^{2Ht} \partial_t$. ■

Remark: In the Koszul formula (1.45), terms with scalar products vanish for an orthonormal basis; however, terms with commutators remain. If we choose the basis vectors as coordinate derivatives ∂_μ (this is the choice usually made in calculations using index notation), the terms with commutators will vanish but terms with scalar products will remain. At first glance, it is not obvious which choice of basis has more advantages in calculations. However, it turns out that in many cases calculations are shorter in an orthonormal basis. ■

6.1.5 Connection as a set of 1-forms

In the previous section we have seen that the information about Levi-Civita connection ∇ is carried entirely by the commutators $[\mathbf{e}_a, \mathbf{e}_b]$ and can thus be recovered without explicitly involving the metric g. A more elegant (and practically more useful) description of the Levi-Civita connection is obtained by considering the dual tetrad $\{\boldsymbol{\theta}^a\}$.

Just as the commutator of vector fields is defined geometrically without involving the metric, the exterior differential $\mathrm{d}\boldsymbol{\theta}^a$ is a 2-form defined geometrically by

$$\mathrm{d}\boldsymbol{\theta}^a(\mathbf{u}, \mathbf{v}) = \mathbf{u} \circ \boldsymbol{\theta}^a(\mathbf{v}) - \mathbf{v} \circ \boldsymbol{\theta}^a(\mathbf{u}) - \boldsymbol{\theta}^a([\mathbf{u}, \mathbf{v}]).$$

Substituting the tetrad vectors $\mathbf{e}_b$ and $\mathbf{e}_c$ instead of $\mathbf{u}, \mathbf{v}$, we have

$$\mathrm{d}\boldsymbol{\theta}^a(\mathbf{e}_b, \mathbf{e}_c) = -\boldsymbol{\theta}^a([\mathbf{e}_b, \mathbf{e}_c]) \equiv -C^a{}_{bc}.$$

So the 2-forms $\mathrm{d}\boldsymbol{\theta}^a$ carry essentially the same information as the commutators of the tetrad vectors. Hence, we expect to be able to express the Levi-Civita connection through the dual tetrad $\boldsymbol{\theta}^a$ and its differentials $\mathrm{d}\boldsymbol{\theta}^a$. To do this, one uses some clever tricks, which we will now show.

Since $\{\boldsymbol{\theta}^a\}$ is a basis in the dual space, the 2-form $\mathrm{d}\boldsymbol{\theta}^a$ can be always decomposed as

$$\mathrm{d}\boldsymbol{\theta}^a = \frac{1}{2} \sum_{b,c} X^a{}_{bc} \boldsymbol{\theta}^b \wedge \boldsymbol{\theta}^c$$

with some coefficients $X^a{}_{bc} = -X^a{}_{cb}$. It is easy to see that these coefficients are directly related to $C^a{}_{bc}$ (see the following Calculation).

Calculation: Show that $X^a{}_{bc} = -C^a{}_{bc}$.

Solution: Consider the 2-form $\mathrm{d}\boldsymbol{\theta}^a$ applied to arbitrary vectors $\mathbf{u}, \mathbf{v}$. Decompose $\mathbf{u}$ and $\mathbf{v}$ through the tetrad and compute

$$\begin{aligned}
(\mathrm{d}\boldsymbol{\theta}^a) \circ (\mathbf{u}, \mathbf{v}) &= (\mathrm{d}\boldsymbol{\theta}^a) \circ \left(\sum_b u^b \mathbf{e}_b, \sum_c v^c \mathbf{e}_c \right) \\
&= \sum_{b,c} u^b v^c \, (\mathrm{d}\boldsymbol{\theta}^a) \circ (\mathbf{e}_b, \mathbf{e}_c) = -\sum_{b,c} C^a{}_{bc} u^b v^c.
\end{aligned}$$

On the other hand, $u^b \equiv \boldsymbol{\theta}^b \circ \mathbf{u}$ and $v^c \equiv \boldsymbol{\theta}^c \circ \mathbf{v}$, so

$$\begin{aligned}
\left(\sum_{b,c} X^a{}_{bc} \boldsymbol{\theta}^b \wedge \boldsymbol{\theta}^c \right) \circ (\mathbf{u}, \mathbf{v}) &= \sum_{b,c} X^a{}_{bc} \left(u^b v^c - u^c v^b \right) \\
&= 2 \sum_{b,c} X^a{}_{bc} u^b v^c.
\end{aligned}$$

Since $\mathbf{u}, \mathbf{v}$ are arbitrary vectors, we obtain $X^a{}_{bc} = -C^a{}_{bc}$. ■

Let us now see how to recover the Levi-Civita connection. So far we have the property

$$\mathrm{d}\boldsymbol{\theta}^c = -\frac{1}{2} \sum_{a,b} C^c_{ab} \boldsymbol{\theta}^a \wedge \boldsymbol{\theta}^b, \quad (6.12)$$

which shows that the coefficients $C^c{}_{ab}$ can be easily computed by decomposing $\mathrm{d}\boldsymbol{\theta}^c$ into the basis of $\boldsymbol{\theta}^a \wedge \boldsymbol{\theta}^b$. However, according to Eq. (6.10), covariant derivatives of vectors involve a certain combination of $C^c{}_{ab}$,

$$\boldsymbol{\theta}_c \circ \nabla_{\mathbf{e}_a} \mathbf{e}_b = \frac{1}{2}\left(C_{c\,ab} - C_{b\,ac} - C_{a\,bc}\right),$$

rather than simply $C^c{}_{ab}$. (Here and below we freely lower and raise all Latin indices by implicitly using the fiducial Minkowski metric η_{ab}.) Now we note that the scalar expression $\boldsymbol{\theta}^c \circ \nabla_{\mathbf{u}} \mathbf{e}_b$ linearly depends on the vector $\mathbf{u}$ and is thus equal to some 1-form applied to $\mathbf{u}$. Let us *denote* that 1-form by $\boldsymbol{\omega}^c{}_b$; in other words, by definition, we set

$$\boldsymbol{\omega}^c{}_b(\mathbf{u}) \equiv \boldsymbol{\theta}^c \circ \nabla_{\mathbf{u}} \mathbf{e}_b.$$

Another way to restate the definition is to say that the 1-form $\boldsymbol{\omega}^c{}_b$ satisfies

$$\nabla_{\mathbf{u}} \mathbf{e}_b = \sum_c \boldsymbol{\omega}^c{}_b(\mathbf{u}) \mathbf{e}_c \tag{6.13}$$

for any vector $\mathbf{u}$. The 1-forms $\boldsymbol{\omega}^c{}_b$ are called the **connection 1-forms** or the **spin connection** corresponding to the dual tetrad $\boldsymbol{\theta}^a$.

It is clear from Eq. (6.10) that the 1-forms $\boldsymbol{\omega}_{c\,b}$ (that is, $\boldsymbol{\omega}^c{}_b$ with the first index lowered) satisfy

$$\boldsymbol{\omega}_{c\,b} \equiv \frac{1}{2}\sum_a \left(C_{c\,ab} - C_{b\,ac} - C_{a\,bc}\right) \boldsymbol{\theta}^a. \tag{6.14}$$

Thus, the connection forms $\boldsymbol{\omega}_{c\,b}$ may be computed straightforwardly from the commutator coefficients $C^c{}_{ab}$. After computing $\boldsymbol{\omega}_{c\,b}$, covariant derivatives of arbitrary vector fields can be expressed as follows,

$$\nabla_{\mathbf{u}} \mathbf{e}_b = \sum_c \boldsymbol{\omega}^c{}_b(\mathbf{u}) \mathbf{e}_c,$$

$$\nabla_{\mathbf{u}} \mathbf{v} = \sum_b \nabla_{\mathbf{u}} \left(v^b \mathbf{e}_b\right) = \sum_b \left(\mathbf{u} \circ v^b\right) \mathbf{e}_b + \sum_{b,c} v^b \boldsymbol{\omega}^c{}_b(\mathbf{u}) \mathbf{e}_c.$$

In terms of the components v^c of a vector field $\mathbf{v}$, we find the formula

$$\left(\nabla_{\mathbf{u}} \mathbf{v}\right)^c \equiv \boldsymbol{\theta}^c \circ \nabla_{\mathbf{u}} \mathbf{v} = \mathbf{u} \circ v^c + \sum_b \boldsymbol{\omega}^c{}_b(\mathbf{u}) v^b. \tag{6.15}$$

It is clear that $\boldsymbol{\omega}^c{}_b$ play the role of the Christoffel symbols in the formula for the covariant derivative.

The formula (6.14) is not necessarily the quickest way to compute $\boldsymbol{\omega}_{c\,b}$. Comparing it with Eq. (6.12), we notice that it might help to compute the following expression,

$$\sum_b \boldsymbol{\omega}_{c\,b} \wedge \boldsymbol{\theta}^b = \frac{1}{2}\sum_{a,b} \left(C_{c\,ab} - C_{b\,ac} - C_{a\,bc}\right) \boldsymbol{\theta}^a \wedge \boldsymbol{\theta}^b$$
$$= \frac{1}{2}\sum_{a,b} C_{c\,ab} \boldsymbol{\theta}^a \wedge \boldsymbol{\theta}^b.$$

(The simplification in the second line happens because the terms symmetric in a, b cancel). Then we can raise the index c, use Eq. (6.12), and obtain the following relationship,

$$\mathrm{d}\boldsymbol{\theta}^c = -\sum_b \boldsymbol{\omega}^c{}_b \wedge \boldsymbol{\theta}^b. \tag{6.16}$$

The relationship (6.16) is called the **first Cartan structure equation**.

An immediate practical importance of this equation is that it can be used, instead of Eq. (6.14), to *determine* the 1-forms $\boldsymbol{\omega}^c{}_b$. Of course, Eq. (6.16) alone does not specify $\boldsymbol{\omega}^c{}_b$ uniquely. For instance, the 1-forms $\frac{1}{2}\sum_a C^c{}_{ab} \boldsymbol{\theta}^a$ are also a solution of Eq. (6.16), although these 1-forms do not coincide with $\boldsymbol{\omega}^c{}_b$. More generally, if $\boldsymbol{\omega}^c{}_b$ is any solution of Eq. (6.16), one may add the following term,

$$\boldsymbol{\omega}^c{}_b \to \boldsymbol{\omega}^c{}_b + \sum_a B^c{}_{ab} \boldsymbol{\theta}^a,$$

where $B^c{}_{ab} = B^c{}_{ba}$ is an arbitrary array of coefficients symmetric in (a, b). The new $\boldsymbol{\omega}^c{}_b$ will still be a solution of Eq. (6.16). However, in addition to Eq. (6.16) the connection 1-forms $\boldsymbol{\omega}^c{}_b$ must satisfy the antisymmetry requirement,

$$\boldsymbol{\omega}_{c\,b} = -\boldsymbol{\omega}_{b\,c}, \tag{6.17}$$

which can be easily read from Eq. (6.14). It turns out[3] that the two requirements (6.16), (6.17) are sufficient to specify $\boldsymbol{\omega}^c{}_b$ uniquely, so that one does not need to use Eq. (6.14). In practice, it is often quicker to guess the solution of Eqs. (6.16), (6.17) than to follow the straightforward but longer formula (6.14).

Example: Consider a two-dimensional spacetime with local coordinates $\{t, x\}$ and a frame field $\{\mathbf{e}_0, \mathbf{e}_1\}$ defined as

$$\mathbf{e}_0 = e^{-f(t)} \partial_t, \quad \mathbf{e}_1 = e^{-h(t)} \partial_x,$$

where $f(t), h(t)$ are scalar functions depending only on t (but not on x). The task is to compute the connection 1-forms $\omega_{c\,b}$ and the covariant derivatives $\nabla_{\partial_t} \partial_x$ and $\nabla_{\partial_x} \partial_x$.

Solution: The metric corresponding to the given frame field is

$$g = e^{2f(t)} dt^2 - e^{2h(t)} dx^2.$$

We begin by defining the dual frame,

$$\boldsymbol{\theta}^0 = e^{f(t)} \mathrm{d}t, \quad \boldsymbol{\theta}^1 = e^{h(t)} \mathrm{d}x,$$

and the differentials,

$$\mathrm{d}\boldsymbol{\theta}^0 = 0, \quad \mathrm{d}\boldsymbol{\theta}^1 = \dot{h} e^{h(t)} \mathrm{d}t \wedge \mathrm{d}x = -\dot{h} e^{h(t) - f(t)} \mathrm{d}x \wedge \boldsymbol{\theta}^0,$$

where $\dot{h} \equiv \partial_t h(t)$. Note that we intentionally wrote $\mathrm{d}\boldsymbol{\theta}^1$ in the form $(...) \wedge \boldsymbol{\theta}^0$ rather than $(...) \wedge \boldsymbol{\theta}^1$, trying to satisfy the antisymmetry property (6.17) which entails $\boldsymbol{\omega}^0{}_0 = \boldsymbol{\omega}^1{}_1 = 0$. Then we can *try to guess* a solution $\tilde{\boldsymbol{\omega}}^c{}_b$ of Eq. (6.16) as follows,

$$\tilde{\boldsymbol{\omega}}^0{}_0 = \tilde{\boldsymbol{\omega}}^0{}_1 = 0, \quad \tilde{\boldsymbol{\omega}}^1{}_0 = \dot{h} e^{h(t) - f(t)} \mathrm{d}x, \quad \tilde{\boldsymbol{\omega}}^1{}_1 = 0.$$

Despite our efforts, the 1-forms $\tilde{\boldsymbol{\omega}}^a{}_b$ are not yet the correct connection forms because they do not satisfy the antisymmetry property (6.17). To correct this, we may add a multiple of $\mathrm{d}\boldsymbol{\theta}^0$ to $\tilde{\boldsymbol{\omega}}^1{}_0$ and a multiple of $\mathrm{d}\boldsymbol{\theta}^1$ to $\tilde{\boldsymbol{\omega}}^0{}_1$. Presently, it is clear that we only need to add $\dot{h} e^{h(t) - f(t)} \mathrm{d}x$ to $\tilde{\boldsymbol{\omega}}^0{}_1$. So the correct set of the connection 1-forms is

$$\boldsymbol{\omega}^0{}_1 = \boldsymbol{\omega}_{01} = -\boldsymbol{\omega}_{10} = \boldsymbol{\omega}^1{}_0 = \dot{h} e^{h(t) - f(t)} \mathrm{d}x.$$

The covariant derivatives are now computed using the formula (6.15). Consider the vector $\mathbf{v} = \partial_x = e^{h(t)} \mathbf{e}_1$. In the orthonormal frame $\{\mathbf{e}_1\}$ the vector $\mathbf{v}$ has a single nonzero component, $v^1 = e^{h(t)}$. Therefore Eq. (6.15) is rewritten as

$$\left(\nabla_{\mathbf{u}} \mathbf{v}\right)^c = \mathbf{u} \circ v^c + \boldsymbol{\omega}^c{}_1(\mathbf{u}) v^1.$$

[3] See Sec. 6.1.6 for a general proof of this statement.

A direct evaluation of components yields

$$(\nabla_{\mathbf{u}}\mathbf{v})^0 = \omega^0{}_1(\mathbf{u})e^{h(t)} = \dot{h}e^{2h(t)-f(t)}\mathbf{u}\circ x,$$
$$(\nabla_{\mathbf{u}}\mathbf{v})^1 = \mathbf{u}\circ e^{h(t)} = \dot{h}e^{h(t)}\mathbf{u}\circ t.$$

Therefore, for arbitrary $\mathbf{u}$,

$$\nabla_{\mathbf{u}}\mathbf{v} = \dot{h}e^{h(t)}\left\{e^{h(t)-f(t)}\left(\mathbf{u}\circ x\right)\mathbf{e}_0 + \left(\mathbf{u}\circ t\right)\mathbf{e}_1\right\}.$$

Substituting $\mathbf{u} = \partial_t$ or $\mathbf{u} = \partial_x$, we find

$$\nabla_{\partial_t}\partial_x = \dot{h}e^{h(t)}\mathbf{e}_1 = \dot{h}\partial_x;$$
$$\nabla_{\partial_x}\partial_x = \dot{h}e^{2h(t)-f(t)}\mathbf{e}_0 = \dot{h}e^{2h(t)-2f(t)}\partial_t.$$

■

6.1.6 *Solving equations for n-forms

In Sec. 6.1.5 we found that the connection 1-forms ω_{ab} can be found by solving the Cartan structure equation (6.16). In general, there are several tricks for finding general solutions of linear equations for n-forms in the frame basis.

The following statement shows how to solve the equations such as the Cartan structure equation in a more general case.

Statement 6.1.6.1: If any set of 2-forms A_c is given (where c is a frame index) and $\{\theta^c\}$ is a dual frame basis then there exists a unique set of 1-forms χ_{ab} such that

$$A_a + \sum_b \chi_{ab}\wedge\theta^b = 0; \quad \chi_{ab} = -\chi_{ba}.$$

The 1-forms χ_{ab} can be expressed by the formula

$$\chi_{ab} = \sum_c \chi_{ab\,c}\theta^c, \quad \chi_{ab\,c} \equiv \frac{1}{2}\left(A_{a\,bc} - A_{b\,ac} - A_{c\,ab}\right), \tag{6.18}$$

where $A_{c\,ab}$ are the coefficients in the decomposition

$$A_c = \frac{1}{2}\sum_{a,b} A_{c\,ab}\theta^a\wedge\theta^b, \quad A_{c\,ab} = -A_{c\,ba}.$$

Idea of proof: Guess the formula for a solution χ_{ab} and then show that the difference $\tilde{\chi}_{ab} - \chi_{ab}$ of two possible solutions is zero. (Proof on page 181.) ■

A consequence of Statement 6.1.6.1 for $A_c = \mathrm{d}\theta_c$, $\chi_{ab} \equiv \omega_{ab}$ is that the system of equations (6.16), (6.17) has the unique solution (6.14).

The formula (6.18) can be also rewritten in a component-free manner, using the frame bases $\{\mathbf{e}_a\}$ and $\{\theta^a\}$ instead:

$$\chi_{ab} = \frac{1}{2}\left(\iota_{\mathbf{e}_b}A_a - \iota_{\mathbf{e}_a}A_b - \sum_c A_c(\mathbf{e}_a,\mathbf{e}_b)\theta^c\right). \tag{6.19}$$

It is clear that this expression is antisymmetric in (a,b). The corresponding solution for the connection 1-forms can be written as

$$\omega_{ab} = \frac{1}{2}\iota_{\mathbf{e}_b}\mathrm{d}\theta_a - \frac{1}{2}\iota_{\mathbf{e}_a}\mathrm{d}\theta_b - \frac{1}{2}\sum_c\left(\iota_{\mathbf{e}_b}\iota_{\mathbf{e}_a}\mathrm{d}\theta_c\right)\theta^c. \tag{6.20}$$

It is instructive to check directly that $A_a + \sum_b \chi_{ab}\wedge\theta^b = 0$ holds with χ_{ab} defined by Eq. (6.19). We compute

$$2\sum_b \chi_{ab}\wedge\theta^b = \sum_b\left(\iota_{\mathbf{e}_b}A_a\right)\wedge\theta^b - \sum_b\left(\iota_{\mathbf{e}_a}A_b\right)\wedge\theta^b - \sum_{b,c} A_c(\mathbf{e}_a,\mathbf{e}_b)\theta^c\wedge\theta^b. \tag{6.21}$$

To proceed, we note that expressions such as

$$\sum_b\left(\iota_{\mathbf{e}_b}A_a\right)\wedge\theta^b$$

are reminiscent of the decomposition (6.1). The following statement shows how such expressions can be simplified.

Statement 6.1.6.2: Suppose $\mathbf{\Omega}$ is any n-form and $\{\mathbf{e}_a\}$ and $\{\theta^a\}$ is a basis and a dual basis for which the decomposition (6.1) holds. (These bases must be dual to each other but do not actually have to be orthonormal.) Then one has the decomposition

$$\mathbf{\Omega} = \frac{1}{n}\sum_a \theta^a\wedge\left(\iota_{\mathbf{e}_a}\mathbf{\Omega}\right). \tag{6.22}$$

In other words,

$$\frac{1}{n}\sum_a \theta^a\wedge\iota_{\mathbf{e}_a}$$

is the identity operator on n-forms.

Proof of Statement 6.1.6.2: Let us apply the n-form at the right-hand side of Eq. (6.22) to a set of n arbitrary vectors $\{\mathbf{v}_1,...,\mathbf{v}_n\}$. By definition (1.21) of the exterior product, we have

$$\left(\sum_a \theta^a\wedge\iota_{\mathbf{e}_a}\mathbf{\Omega}\right)\circ(\mathbf{v}_1,...,\mathbf{v}_n)$$
$$= \sum_{j=1}^n(-1)^{j-1}\sum_a\theta^a(\mathbf{v}_j)\mathbf{\Omega}\left(\mathbf{e}_a,\mathbf{v}_1,...,\hat{\mathbf{v}}_j,...,\mathbf{v}_n\right),$$

where the hat over $\hat{\mathbf{v}}_j$ means that the j-th vector is absent from the list of arguments. Since the n-form $\boldsymbol{\omega}$ is linear in each argument, we may substitute

$$\sum_a\theta^a(\mathbf{v}_j)\mathbf{\Omega}\left(\mathbf{e}_a,\mathbf{v}_1,...\right) = \sum_a\mathbf{\Omega}\left(\theta^a(\mathbf{v}_j)\mathbf{e}_a,\mathbf{v}_1,...\right)$$
$$= \mathbf{\Omega}\left(\mathbf{v}_j,\mathbf{v}_1,...\right).$$

Hence,

$$\left(\sum_a \theta^a\wedge\iota_{\mathbf{e}_a}\mathbf{\Omega}\right)\circ(\mathbf{v}_1,...,\mathbf{v}_n)$$
$$= \sum_{j=1}^n(-1)^{j-1}\mathbf{\Omega}\left(\mathbf{v}_j,\mathbf{v}_1,...,\hat{\mathbf{v}}_j,...,\mathbf{v}_n\right)$$
$$= \sum_{j=1}^n\mathbf{\Omega}\left(\mathbf{v}_1,...,\mathbf{v}_n\right) = n\mathbf{\Omega}\left(\mathbf{v}_1,...,\mathbf{v}_n\right).$$

This shows that Eq. (6.22) holds. ■

Using Statement 6.1.6.2 with $n = 2$, we can transform Eq. (6.21) as follows,

$$2\sum_b\chi_{ab}\wedge\theta^b = -2A_a - \sum_b\left(\iota_{\mathbf{e}_a}A_b\right)\wedge\theta^b + \sum_c\left(\sum_b\iota_{\mathbf{e}_b}\left(\iota_{\mathbf{e}_a}A_c\right)\theta^b\right)\wedge\theta^c$$
$$= -2A_a - \sum_c\left(\iota_{\mathbf{e}_a}A_c\right)\wedge\theta^c + \sum_c\left(\iota_{\mathbf{e}_a}A_c\right)\wedge\theta^c$$
$$= -2A_a.$$

Thus $A_a + \sum_b\chi_{ab}\wedge\theta^b = 0$.

A useful alternative way of writing the formula (6.19) is shown in the following calculation.

Calculation 6.1.6.3: The formula (6.19) can be rewritten equivalently as

$$\chi_{a\,b} = \iota_{\mathbf{e}_b} A_a - \iota_{\mathbf{e}_a} A_b - \frac{1}{2}\iota_{\mathbf{e}_b}\iota_{\mathbf{e}_a}\sum_c \theta^c \wedge A_c. \tag{6.23}$$

Details: We may interchange the order of θ^c and $\iota_{\mathbf{e}_a}$ in Eq. (6.19) by using the property of $\iota_{\mathbf{x}}$,

$$\iota_{\mathbf{x}}\left(\theta^c \wedge \mathbf{\Omega}\right) = \left(\iota_{\mathbf{x}}\theta^c\right)\mathbf{\Omega} - \theta^c \wedge \iota_{\mathbf{x}}\mathbf{\Omega},$$

which holds for any n-form $\mathbf{\Omega}$ and for any vector $\mathbf{x}$. Since $\iota_{\mathbf{e}_a}\theta^c = \delta^c_a$, we find

$$\begin{aligned}\iota_{\mathbf{e}_a}\sum_c \theta^c \wedge A_c &= \sum_c \delta^c_a A_c - \sum_c \theta^c \wedge \iota_{\mathbf{e}_a} A_c \\ &= A_a - \sum_c \theta^c \wedge \iota_{\mathbf{e}_a} A_c\end{aligned}$$

and

$$\begin{aligned}\iota_{\mathbf{e}_b}\iota_{\mathbf{e}_a}\sum_c \theta^c \wedge A_c &= \iota_{\mathbf{e}_b}\left(A_a - \sum_c \theta^c \wedge \iota_{\mathbf{e}_a} A_c\right) \\ &= \iota_{\mathbf{e}_b} A_a - \iota_{\mathbf{e}_a} A_b + \sum_c \theta^c \wedge \iota_{\mathbf{e}_b}\iota_{\mathbf{e}_a} A_c. \end{aligned}\tag{6.24}$$

Since

$$\iota_{\mathbf{e}_b}\iota_{\mathbf{e}_a} A_c \equiv A_c(\mathbf{e}_a, \mathbf{e}_b),$$

the formula (6.19) can be rewritten as Eq. (6.23). ■

Remark: The spin connection $\omega_{a\,b}$ can be found from the formula (6.23) by substituting $A_c \equiv \mathrm{d}\theta_c$. Due to the Frobenius theorem, the expression $\theta^c \wedge A_c = \theta^c \wedge \mathrm{d}\theta_c$ is equal to zero if the 1-forms θ^c correspond to hypersurface-orthogonal vector fields $\mathbf{e}_c$. Since the orthonormal frame $\{\mathbf{e}_c\}$ frequently consists of hypersurface-orthogonal vectors, calculations are simplified when the formula (6.23) is used. ■

As an aside, we note that the formula (6.23) cannot hold for 1-forms A_c (any 1-form A_c is decomposed uniquely as $\sum_a A_{ac}\theta^a$ and the coefficients A_{ac} are not necessarily antisymmetric), but it can be generalized from 2-forms $\mathbf{A}_a$ to n-forms $\mathbf{\Omega}_a$.

Statement 6.1.6.4: Given a set of n-forms A_c (where c is a vielbein index and $n \geq 3$), there exists a set of $(n-1)$-forms $B_{b\,a}$ such that

$$A_a - \sum_b \theta^b \wedge B_{b\,a} = 0; \quad B_{a\,b} = -B_{b\,a}. \tag{6.25}$$

A suitable set of $B_{b\,a}$ is determined by the equivalent expressions

$$B_{b\,a} = \frac{1}{n-1}\left(\iota_{\mathbf{e}_b} A_a - \iota_{\mathbf{e}_a} A_b\right) - \frac{1}{n\,(n-1)}\iota_{\mathbf{e}_b}\iota_{\mathbf{e}_a}\sum_c \theta^c \wedge A_c \tag{6.26}$$

$$= \frac{1}{n}\left(\iota_{\mathbf{e}_b} A_a - \iota_{\mathbf{e}_a} A_b\right) - \frac{1}{n\,(n-1)}\sum_c \theta^c \wedge \left(\iota_{\mathbf{e}_b}\iota_{\mathbf{e}_a} A_c\right). \tag{6.27}$$

Unlike the case $n = 2$, the set $B_{b\,a}$ satisfying Eq. (6.25) is not unique. (Proof on page 181.) ■

6.2 Applications of tetrad formalism

After developing powerful techniques for manipulating the orthonormal frames, let us see what calculations are possible in this formalism. To make the notation more concise, I will now adopt the Einstein summation convention for the frame indices; for instance, I will now write $\mathbf{e}_a \circ \theta^a(\mathbf{u})$ instead of $\sum_a \mathbf{e}_a \circ \theta^a(\mathbf{u})$. (In Sec. 6.1 I did not use the Einstein summation convention but wrote all summations explicitly, to make the presentation more clear.) The frame indices are raised and lowered using the fiducial Minkowski metric η_{ab}.

Examples: Here are some previously derived identities rewritten in this notation:

$$\begin{aligned} g(\mathbf{u},\mathbf{v}) &= \theta^a(\mathbf{u})\theta_a(\mathbf{v}), \quad g = \theta^a \otimes \theta_a; \\ \hat{g}\mathbf{u} &= g(\mathbf{u},\mathbf{e}_a)\theta^a = \theta_a(\mathbf{u})\theta^a, \\ \mathrm{Tr}_{(\mathbf{a},\mathbf{b})} X(\mathbf{a},\mathbf{b}) &= X(\mathbf{e}_a,\mathbf{e}^a), \\ \omega &= \theta^a \omega(\mathbf{e}_a), \end{aligned}$$

where ω is a 1-form. ■

6.2.1 Computing geodesic equations

The geodesic equation $\nabla_{\mathbf{u}}\mathbf{u} = 0$ is used to find trajectories of particles in GR. We now show how to use the vielbein formalism to derive the required differential equations for the particle worldline $\{x^\mu(\tau)\}$ in a local coordinate system.[4]

Assuming that $\{x^\mu\}$ is a given local coordinate system in an n-dimensional manifold, we are interested in determining a trajectory $\gamma \equiv \{x^\mu(\tau)\}$ such that $\nabla_{\dot\gamma}\dot\gamma = 0$. Suppose that an orthonormal frame $\{\mathbf{e}_a\}$, the dual frame $\{\theta^a\}$, and the corresponding connection 1-forms $\omega_{a\,b}$ are already known. We would like to determine the unknown functions $\{x^\mu(\tau)\}$. The tangent vector $\dot\gamma$ can be expressed as

$$\dot\gamma = u^a \mathbf{e}_a, \quad u^a \equiv \theta^a \circ \dot\gamma.$$

The coefficients u^a will be particular expressions involving $x^\mu(\tau)$ and $\dot{x}^\mu(\tau)$. Then we use Eq. (6.13) and compute

$$\begin{aligned}\nabla_{\dot\gamma}\dot\gamma &= \nabla_{\dot\gamma}\left(u^a \mathbf{e}_a\right) \\ &= \left(\nabla_{\dot\gamma} u^a\right)\mathbf{e}_a + u^b \mathbf{e}_c \left(\omega^c{}_b \circ \dot\gamma\right) \\ &= \left(\nabla_{\dot\gamma} u^a + u^b \omega^a{}_b \circ \dot\gamma\right)\mathbf{e}_a.\end{aligned}$$

Since $\{\mathbf{e}_a\}$ is a basis, we obtain n equations

$$\nabla_{\dot\gamma} u^a + \sum_b u^b \omega^a{}_b \circ \dot\gamma = 0, \quad a = 1, \ldots, n.$$

These equations are second-order in derivatives of $x^\mu(\tau)$, because u^a contains $\dot{x}^\mu$, while the derivative along the curve, $\nabla_{\dot\gamma} u^a$, produces $\ddot{x}^\mu$ when it acts on $\dot{x}^\mu$ contained in u^a.

The geodesic equations can be further simplified, so that precomputing $\omega^a{}_b$ is not required.

Calculation 6.2.1.1: The geodesic equation for a vector field $\mathbf{u}$ can be written as

$$\mathbf{u} \circ \left(\theta_a(\mathbf{u})\right) + \theta^b(\mathbf{u})\iota_{\mathbf{e}_a}\iota_{\mathbf{u}}\left(\mathrm{d}\theta_b\right) = 0, \quad a = 1, \ldots, n. \tag{6.28}$$

[4] The idea to explore geodesic equations in tetrad formalism came to me after reading the section entitled "Computing with adapted frames and examples" of the book [20].

One can derive the formula (6.28) by using the explicit relationship between $\omega^a{}_b$ and θ^a, for instance, in the form (6.20). However, it is faster to use the Koszul formula (1.45) for computing the scalar product $g(\nabla_{\mathbf{u}}\mathbf{u}, \mathbf{e}_a)$, which must vanish for every $a = 1, ..., n$. We find (without assuming $g(\mathbf{u}, \mathbf{u}) = \text{const}$)

$$\begin{aligned} g(\nabla_{\mathbf{u}}\mathbf{u}, \mathbf{e}_a) &= \frac{1}{2}\left(2\mathbf{u} \circ g(\mathbf{u}, \mathbf{e}_a) - \mathbf{e}_a \circ g(\mathbf{u}, \mathbf{u}) - 2g(\mathbf{u}, [\mathbf{u}, \mathbf{e}_a])\right) \\ &\overset{1}{=} \mathbf{u} \circ \theta_a(\mathbf{u}) - \mathbf{e}_a \circ \frac{g(\mathbf{u}, \mathbf{u})}{2} + \theta^b(\mathbf{u})\theta_b([\mathbf{e}_a, \mathbf{u}]) \\ &\overset{2}{=} \mathbf{u} \circ \theta_a(\mathbf{u}) + \theta^b(\mathbf{u})\left(\mathrm{d}\theta_b(\mathbf{u}, \mathbf{e}_a) + \mathbf{e}_a \circ \theta_b(\mathbf{u})\right) \\ &\quad - \mathbf{e}_a \circ \frac{g(\mathbf{u}, \mathbf{u})}{2} \\ &\overset{3}{=} \mathbf{u} \circ \theta_a(\mathbf{u}) + \theta^b(\mathbf{u})\mathrm{d}\theta_b(\mathbf{u}, \mathbf{e}_a), \end{aligned}$$

where $\overset{1}{=}$ is due to Eqs. (6.2) and (6.4); $\overset{2}{=}$ is due to the definition (1.23) and the fact that $\mathbf{u} \circ \theta_b(\mathbf{e}_a) = \mathbf{u} \circ \eta_{ab} = 0$; and $\overset{3}{=}$ is due to

$$\theta^b(\mathbf{u})\left(\mathbf{e}_a \circ \theta_b(\mathbf{u})\right) = \frac{1}{2}\mathbf{e}_a \circ \left(\theta^b(\mathbf{u})\theta_b(\mathbf{u})\right) = \frac{1}{2}\mathbf{e}_a \circ g(\mathbf{u}, \mathbf{u}).$$

Of course, $g(\mathbf{u}, \mathbf{u}) = \text{const}$ for a geodesic field $\mathbf{u}$, but we did not use that fact in the present derivation. As a result, the left-hand side of Eq. (6.28) is strictly equivalent to $g(\nabla_{\mathbf{u}}\mathbf{u}, \mathbf{e}_a)$ even for non-geodesic vector fields $\mathbf{u}$. ■

In practical calculations, it is frequently convenient to use the "conservation law" $g(\dot\gamma, \dot\gamma) = \text{const}$ in place of one of the n geodesic equations. (The first-order equation $g(\dot\gamma, \dot\gamma) = \text{const}$ is always a consequence of the second-order geodesic equations.)

Example: Let us determine the geodesic curves in the two-dimensional metric

$$g = e^{2f(x)}dt^2 - e^{2h(x)}dx^2.$$

The obvious orthonormal frame consists of the vectors $\mathbf{e}_0 = e^{-f}\partial_t$, $\mathbf{e}_1 = e^{-h}\partial_x$ and the corresponding 1-forms $\theta^0 = e^f \mathrm{d}t$, $\theta^1 = e^h \mathrm{d}x$. According to the geodesic equation (6.28), a curve $\gamma(\tau) \equiv \{t(\tau), x(\tau)\}$ with the tangent vector

$$\dot\gamma(\tau) \equiv \left\{\frac{dt(\tau)}{d\tau}, \frac{dx(\tau)}{d\tau}\right\} \equiv \dot t(\tau)\partial_t + \dot x(\tau)\partial_x$$

is geodesic if the following two equations hold,

$$\dot\gamma \circ (\theta_a(\dot\gamma)) + \theta^b(\dot\gamma)\,(\mathrm{d}\theta_b) \circ (\dot\gamma, \mathbf{e}_a) = 0, \quad a = 0, 1.$$

(The overdot denotes $d/d\tau$ everywhere.) Presently, $\mathrm{d}\theta^0 = f' e^f \mathrm{d}x \wedge \mathrm{d}t$ and $\mathrm{d}\theta^1 = 0$, while

$$\theta^0(\dot\gamma) = \theta_0(\dot\gamma) = e^f \dot t, \quad \theta^1(\dot\gamma) = -\theta_1(\dot\gamma) = e^h \dot x,$$

so the two geodesic equations are

$$\begin{aligned} \dot\gamma \circ \left(e^f \dot t\right) + e^f \dot t \left(f' e^f \mathrm{d}x \wedge \mathrm{d}t\right) \circ \left(\dot\gamma, e^{-f}\partial_t\right) = 0, \qquad [a = 0] \\ \dot\gamma \circ \left(-e^h \dot x\right) + e^f \dot t \left(f' e^f \mathrm{d}x \wedge \mathrm{d}t\right) \circ \left(\dot\gamma, e^{-h}\partial_x\right) = 0. \qquad [a = 1] \end{aligned}$$

Simplifying these equations, while keeping in mind that $\dot\gamma$ acts on $\dot t$ and $\dot x$ simply as $d/d\tau$, we find

$$\begin{aligned} \ddot t + 2f'\dot x \dot t = 0, \qquad [a = 0] \\ \ddot x + h'\dot x^2 + f' e^{2f-2h}\dot t^2 = 0. \qquad [a = 1] \end{aligned}$$

These are second-order equations for a general geodesic curve $\{t(\tau), x(\tau)\}$.

We now use the property $g(\dot\gamma, \dot\gamma) = \text{const}$ to simplify further calculations. To be definite, let us look for timelike geodesics normalized by

$$g(\dot\gamma, \dot\gamma) = e^{2f}\dot t^2 - e^{2h}\dot x^2 = 1.$$

Presently it is clear that the equation for $x(\tau)$ contains only $\dot t^2$ and functions of x and $\dot x$. Therefore, let us substitute $\dot t^2$ through $\dot x^2$ into that equation,

$$\begin{aligned} \ddot x + h'\dot x^2 + f' e^{-2h}\left(1 + e^{2h}\dot x^2\right) \\ = \ddot x + \left(h' + f'\right)\dot x^2 + f' e^{-2h} = 0. \end{aligned}$$

We obtained a closed second-order equation for $x(\tau)$, and we can proceed to solve this equation by standard methods. Reducing to the first-order equation by a substitution $\dot x = v(x)$, we find

$$vv' + \left(h' + f'\right)v^2 + f' e^{-2h} = 0,$$

where the prime stands for d/dx. The last equation can be integrated to

$$\frac{1}{2}v^2 e^{2(h+f)} = \frac{1}{2}C^2 - \frac{1}{2}e^{2f},$$

where $C^2 > 0$ is an integration constant. Finally, the solution of the geodesic equation for $x(\tau)$ is expressed in the form of an indefinite integral,

$$\int^{x(\tau)} \frac{dx}{v(x)} = \int^{x(\tau)} \frac{e^{h+f}dx}{\sqrt{C^2 - e^{2f}}} = \tau - \tau_0.$$

Then the solution for $t(\tau)$ can be found by integrating the equation

$$\dot t = e^{-f}\sqrt{1 + e^{2h}\dot x^2} = Ce^{-2f(x(\tau))}.$$

In some cases (for some functions f and h), these equations can be integrated analytically and the geodesics obtained in closed form. ■

6.2.2 Determining Killing vectors

A Killing vector field $\mathbf{k}$ satisfies $\mathcal{L}_{\mathbf{k}}g = 0$ (see Sec. 1.6.7); a conformal Killing vector satisfies $\mathcal{L}_{\mathbf{k}}g = 2\lambda g$ (see Sec. 3.1.3). Sometimes one would like to determine whether a Killing vector or a conformal Killing vector exists for a given metric. One can perform the necessary calculations in the vielbein formalism.

Since the metric g can be expressed through the vielbein (see Sec. 6.1.1) as

$$g = \eta_{ab}\theta^a \otimes \theta^b,$$

while η_{ab} is a constant numeric matrix, we have

$$\mathcal{L}_{\mathbf{k}}g = \eta_{ab}\left[(\mathcal{L}_{\mathbf{k}}\theta^a) \otimes \theta^b + \theta^a \otimes \mathcal{L}_{\mathbf{k}}\theta^b\right].$$

The Lie derivatives $\mathcal{L}_{\mathbf{k}}\theta^a$ can be computed easily through the Cartan homotopy formula (1.24),

$$\mathcal{L}_{\mathbf{k}}\theta^a = \mathrm{d}\left(\theta^a \circ \mathbf{k}\right) + \iota_{\mathbf{k}}\mathrm{d}\theta^a.$$

Then one can derive the Killing equation easily.

Example: Consider a metric

$$g = e^{2f(t)}dt^2 - e^{2h(t)}dx^2.$$

Let us determine whether

$$\mathbf{k} \equiv a(t,x)\partial_t + b(t,x)\partial_x$$

can be a Killing vector for g, where $a(t,x)$ and $b(t,x)$ are unknown functions.

The dual frame is

$$\boldsymbol{\theta}^0 = e^{f(t)}\mathrm{d}t, \quad \boldsymbol{\theta}^1 = e^{h(t)}\mathrm{d}x.$$

The differentials are

$$\mathrm{d}\boldsymbol{\theta}^0 = 0, \quad \mathrm{d}\boldsymbol{\theta}^1 = e^h\dot{h}\mathrm{d}t \wedge \mathrm{d}x.$$

So we compute

$$\mathcal{L}_{\mathbf{k}}\boldsymbol{\theta}^0 = \mathrm{d}\left(\boldsymbol{\theta}^0 \circ \mathbf{k}\right) + \iota_{\mathbf{k}}\mathrm{d}\boldsymbol{\theta}^0 = \mathrm{d}\left(e^f a\right),$$
$$\mathcal{L}_{\mathbf{k}}\boldsymbol{\theta}^1 = \mathrm{d}\left(\boldsymbol{\theta}^1 \circ \mathbf{k}\right) + \iota_{\mathbf{k}}\mathrm{d}\boldsymbol{\theta}^1 = \mathrm{d}\left(e^h b\right) + e^h\dot{h}\left(a\mathrm{d}x - b\mathrm{d}t\right).$$

Note that

$$\mathrm{d}(e^f a) = \left(e^f a\right)_{,t}\mathrm{d}t + \left(e^f a\right)_{,x}\mathrm{d}x,$$

and similarly for $\mathrm{d}(e^h b)$. So the Killing equation, $\mathcal{L}_{\mathbf{k}}g = 0$, becomes

$$\mathcal{L}_{\mathbf{k}}\boldsymbol{\theta}^0 \otimes \boldsymbol{\theta}^0 + \boldsymbol{\theta}^0 \otimes \mathcal{L}_{\mathbf{k}}\boldsymbol{\theta}^0 - \mathcal{L}_{\mathbf{k}}\boldsymbol{\theta}^1 \otimes \boldsymbol{\theta}^1 - \boldsymbol{\theta}^1 \otimes \mathcal{L}_{\mathbf{k}}\boldsymbol{\theta}^1 = 0.$$

This is rewritten (with implied symmetric tensor products) as

$$\begin{aligned}&\left[\left(e^f a\right)_{,t}dt + \left(e^f a\right)_{,x}dx\right]e^f dt\\ &- \left[\left(e^h b\right)_{,t}dt + \left(e^h b\right)_{,x}dx + e^h\dot{h}\left(adx - bdt\right)\right]e^h dx\\ &= 0.\end{aligned}$$

Gathering terms with dt^2, $dt\,dx$, and dx^2, we obtain the equations

$$\begin{aligned}\left(e^f a\right)_{,t} &= 0,\\ \left(e^f a\right)_{,x}e^f - \left(e^h b\right)_{,t}e^h + b\dot{h}e^{2h} &= 0,\\ \left(e^h b\right)_{,x} + a\dot{h}e^h &= 0.\end{aligned}$$

These equations can be simplified to

$$\begin{aligned}\left(e^f a\right)_{,t} &= 0,\\ a_{,x} &= b_{,t}e^{2h-2f},\\ b_{,x} &= -a\dot{h}.\end{aligned}$$

One solution is $a = 0, b = \text{const}$, which is the obvious Killing vector $\mathbf{k} = \partial_x$. If $a \neq 0$, we may solve the first equation as $a = q(x)e^{-f}$ with $q \neq 0$, differentiate the second equation by x, the third equation by t, and divide by q to obtain

$$\frac{q_{,xx}}{q} = -e^{2h-2f}\left(\ddot{h} - \dot{h}\dot{f}\right).$$

Since the right-hand side is a function of t while the left-hand side is a function of x, solutions exist only if both sides are equal to a constant. Unless the functions $f(t), h(t)$ are such that $e^{2h-2f}\left(\ddot{h} - \dot{h}\dot{f}\right) = K = \text{const}$, no other Killing vectors exist besides $\mathbf{k} = \partial_x$. ■

6.2.3 Curvature as a set of 2-forms

The connection forms $\boldsymbol{\omega}^a{}_b$ contain first derivatives of the metric and are analogous to Christoffel symbols. The Riemann tensor also can be conveniently represented as a set of 2-forms.

Consider the definition of the Riemann tensor,

$$R(\mathbf{u},\mathbf{v})\mathbf{w} \equiv \nabla_{\mathbf{u}}\nabla_{\mathbf{v}}\mathbf{w} - \nabla_{\mathbf{v}}\nabla_{\mathbf{u}}\mathbf{w} - \nabla_{[\mathbf{u},\mathbf{v}]}\mathbf{w},$$

and recall that this is a transformation-valued antisymmetric bilinear form, if viewed as a function of the vectors $\mathbf{u}, \mathbf{v}$. For fixed $\mathbf{u}$ and $\mathbf{v}$, the object $R(\mathbf{u},\mathbf{v})$ is a linear transformation that can be described, as usual, by a matrix of coefficients in the basis $\{\mathbf{e}_a\}$. Let us denote this matrix by $R^a{}_b(\mathbf{u},\mathbf{v})$; so we set, by definition,

$$\boldsymbol{R}^a{}_b(\mathbf{u},\mathbf{v}) \equiv \boldsymbol{\theta}^a \circ R(\mathbf{u},\mathbf{v})\mathbf{e}_b. \tag{6.29}$$

Then the transformation $R(\mathbf{u},\mathbf{v})$ acts on the vector $\mathbf{e}_b$ as

$$R(\mathbf{u},\mathbf{v})\mathbf{e}_b = \mathbf{e}_a\boldsymbol{R}^a{}_b(\mathbf{u},\mathbf{v}).$$

The 2-forms $\boldsymbol{R}^a{}_b$ completely describe the curvature tensor and are called the **curvature 2-forms**.

It is often convenient to lower the index of the curvature 2-forms and to consider the array

$$\boldsymbol{R}_{a\,b} \equiv \eta_{ac}\boldsymbol{R}^c{}_b.$$

The relationship between the 2-forms $\boldsymbol{R}_{a\,b}$ and the covariant Riemann tensor $R(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d})$ can then be expressed as

$$R(\mathbf{x},\mathbf{y},\mathbf{e}_a,\mathbf{e}_b) = \boldsymbol{R}_{b\,a}(\mathbf{x},\mathbf{y}) = \iota_{\mathbf{y}}\iota_{\mathbf{x}}\boldsymbol{R}_{b\,a}.$$

Remark: There is an obvious arbitrariness in the definition of the 2-forms $\boldsymbol{R}^a{}_b$: one could interchange the indices a and b, writing $\boldsymbol{R}_{a\,b}$ instead of $\boldsymbol{R}_{b\,a}$ above. This is the freedom of choosing the overall sign of the Riemann tensor. While not essential, this freedom leads to annoying sign discrepancies between different textbooks. ■

The curvature 2-forms can be computed directly and efficiently from the connection 1-forms $\boldsymbol{\omega}^c{}_b$ using the formula (6.30) below. To derive that formula, let us begin by expressing $R(\mathbf{u},\mathbf{v})\mathbf{e}_b$ with help of Eq. (6.13). We find

$$\begin{aligned}\nabla_{\mathbf{u}}\nabla_{\mathbf{v}}\mathbf{e}_b &= \nabla_{\mathbf{u}}\left(\boldsymbol{\omega}^a{}_b(\mathbf{v})\mathbf{e}_a\right)\\ &= \left(\mathbf{u} \circ \boldsymbol{\omega}^a{}_b(\mathbf{v})\right)\mathbf{e}_a + \boldsymbol{\omega}^a{}_b(\mathbf{v})\nabla_{\mathbf{u}}\mathbf{e}_a\\ &= \left(\mathbf{u} \circ \boldsymbol{\omega}^a{}_b(\mathbf{v})\right)\mathbf{e}_a + \boldsymbol{\omega}^a{}_b(\mathbf{v})\boldsymbol{\omega}^c{}_a(\mathbf{u})\mathbf{e}_c;\\ \nabla_{[\mathbf{u},\mathbf{v}]}\mathbf{e}_b &= \mathbf{e}_a\boldsymbol{\omega}^a{}_b([\mathbf{u},\mathbf{v}]).\end{aligned}$$

We can now compute the coefficient at $\mathbf{e}_a$ of $R(\mathbf{u},\mathbf{v})\mathbf{e}_b$ (while doing that, we need to relabel the index $a \to c$):

$$\begin{aligned}\boldsymbol{\theta}^a \circ R(\mathbf{u},\mathbf{v})\mathbf{e}_b &= \mathbf{u} \circ \boldsymbol{\omega}^a{}_b(\mathbf{v}) - \mathbf{v} \circ \boldsymbol{\omega}^a{}_b(\mathbf{u}) - \boldsymbol{\omega}^a{}_b([\mathbf{u},\mathbf{v}])\\ &\quad + \boldsymbol{\omega}^c{}_b(\mathbf{v})\boldsymbol{\omega}^a{}_c(\mathbf{u}) - \boldsymbol{\omega}^c{}_b(\mathbf{v})\boldsymbol{\omega}^a{}_c(\mathbf{u})\\ &= \left(\mathrm{d}\boldsymbol{\omega}^a{}_b\right) \circ (\mathbf{u},\mathbf{v}) + \left(\boldsymbol{\omega}^a{}_c \wedge \boldsymbol{\omega}^c{}_b\right) \circ (\mathbf{u},\mathbf{v}),\end{aligned}$$

where we used the standard definitions of $\mathrm{d}\boldsymbol{\omega}^a{}_b$ and $\boldsymbol{\omega}^a{}_c \wedge \boldsymbol{\omega}^c{}_b$ for 1-forms $\boldsymbol{\omega}^a{}_b$. The result is called the **second Cartan structure equation**,

$$\boldsymbol{R}^a{}_b = \mathrm{d}\boldsymbol{\omega}^a{}_b + \boldsymbol{\omega}^a{}_c \wedge \boldsymbol{\omega}^c{}_b. \tag{6.30}$$

Remarks: The curvature 2-forms $\boldsymbol{R}_{ab}$ (with the first index lowered through η_{ac}) are obviously antisymmetric in a, b because $\boldsymbol{\omega}_{ab}$ are.

The Cartan structure equations make computations of the curvature tensor significantly shorter. The simplification is due to several factors: firstly, the connection forms $\boldsymbol{\omega}_{ab}$, being antisymmetric in a, b, have fewer nonzero coefficients than the Christoffel symbols $\Gamma^{\lambda}_{\alpha\beta}$, which are symmetric in α, β. Secondly, many terms are cancelled automatically during the calculation of exterior differential and exterior product in Eq. (6.30). ■

Calculation: Compute the connection and the curvature forms for the two-dimensional metric $g = e^{2f}dt^2 - e^{2h}dx^2$, where $f(t,x)$ and $h(t,x)$ are arbitrary functions of t and x. Compute the curvature 2-form of the two-dimensional de Sitter metric by setting $f(t,x) = 0, h(t,x) = Ht$.

Solution: The dual frame is

$$\boldsymbol{\theta}^0 = e^f \mathrm{d}t, \quad \boldsymbol{\theta}^1 = e^h \mathrm{d}x.$$

The differentials $\mathrm{d}\boldsymbol{\theta}^a$ are

$$\mathrm{d}\boldsymbol{\theta}^0 = \mathrm{d}f \wedge e^f \mathrm{d}t = f_{,x} e^f \mathrm{d}x \wedge \mathrm{d}t, \quad d\boldsymbol{\theta}^1 = h_{,t} e^h \mathrm{d}t \wedge \mathrm{d}x.$$

The only nonzero connection forms are $\boldsymbol{\omega}^0{}_1$ and $\boldsymbol{\omega}^1{}_0$, so after some tries we get

$$\boldsymbol{\omega}^0{}_1 = \boldsymbol{\omega}_{01} = -\boldsymbol{\omega}_{10} = \boldsymbol{\omega}^1{}_0 = f_{,x} e^{f-h} \mathrm{d}t + h_{,t} e^{h-f} \mathrm{d}x.$$

The only nonzero curvature 2-forms are $\boldsymbol{R}^0{}_1 = \boldsymbol{R}^1{}_0$ and are found as

$$\boldsymbol{R}^0{}_1 = \mathrm{d}\boldsymbol{\omega}^0{}_1 + \sum_c \boldsymbol{\omega}^0{}_c \wedge \boldsymbol{\omega}^c{}_1;$$

however, the sum over c falls out because $\boldsymbol{\omega}^0{}_0 = \boldsymbol{\omega}^1{}_1 = 0$. Thus the only nontrivial component of the curvature 2-forms is

$$\begin{aligned}\boldsymbol{R}^0{}_1 = \mathrm{d}\boldsymbol{\omega}^0{}_1 &= \left(f_{,x} e^{f-h}\right)_{,x} \mathrm{d}x \wedge \mathrm{d}t + \left(h_{,t} e^{h-f}\right)_{,t} \mathrm{d}t \wedge \mathrm{d}x \\ &= \left[\left(h_{,t} e^{h-f}\right)_{,t} - \left(f_{,x} e^{f-h}\right)_{,x}\right] \mathrm{d}t \wedge \mathrm{d}x.\end{aligned}$$

Note that the calculation is significantly simplified due to the use of forms.

For the de Sitter case, we find

$$\boldsymbol{R}^0{}_1 = H^2 e^{Ht} \mathrm{d}t \wedge \mathrm{d}x.$$

■

6.2.4 Ricci tensor and Ricci scalar

The Ricci tensor Ric is defined as the trace of the Riemann tensor,

$$\mathrm{Ric}\,(\mathbf{a}, \mathbf{b}) \equiv \mathrm{Tr}_{(\mathbf{x},\mathbf{y})} R(\mathbf{a}, \mathbf{x}, \mathbf{b}, \mathbf{y}).$$

In the tetrad formulation, the first two arguments of $R()$ are the implicit arguments of the 2-forms $\boldsymbol{R}^a{}_b$, so by Eq. (6.29) we have

$$R(\mathbf{a}, \mathbf{x}, \mathbf{e}_b, \mathbf{y}) \equiv g(R(\mathbf{a}, \mathbf{x})\mathbf{e}_b, \mathbf{y}) = g(\mathbf{e}_a, \mathbf{y}) \boldsymbol{R}^a{}_b(\mathbf{a}, \mathbf{x}).$$

We can use the basis $\{\mathbf{e}_a\}$ to compute the trace:

$$\begin{aligned}\mathrm{Ric}\,(\mathbf{a}, \mathbf{e}_b) &= \mathrm{Tr}_{(\mathbf{x},\mathbf{y})} R(\mathbf{a}, \mathbf{x}, \mathbf{e}_b, \mathbf{y}) = R(\mathbf{a}, \mathbf{e}^c, \mathbf{e}_b, \mathbf{e}_c) \\ &= \boldsymbol{R}^c{}_b(\mathbf{a}, \mathbf{e}_c).\end{aligned}$$

Therefore, the Ricci tensor is adequately described by the set of 1-forms

$$\begin{aligned}\mathrm{Ric}_b &\equiv -\iota_{\mathbf{e}_c} \boldsymbol{R}^c{}_b = \iota_{\mathbf{e}^c} \boldsymbol{R}_{bc}; \\ \mathrm{Ric}\,(\mathbf{x}, \mathbf{y}) &= \boldsymbol{\theta}^b(\mathbf{y}) \iota_{\mathbf{x}} \mathrm{Ric}_b = \iota_{\mathbf{x}} g(\mathbf{e}^b, \mathbf{y}) \mathrm{Ric}_b.\end{aligned}$$

Here it is important to recall that $\mathbf{e}^b$ differs from $\mathbf{e}_b$ and is defined by $\mathbf{e}^b \equiv \eta^{ab}\mathbf{e}_a$.

Similarly, the Ricci scalar is computed as the trace

$$R = \mathrm{Tr}_{(\mathbf{x},\mathbf{y})} \mathrm{Ric}\,(\mathbf{x}, \mathbf{y}) = \iota_{\mathbf{e}^b} \mathrm{Ric}_b = R_{ab}\left(\mathbf{e}^b, \mathbf{e}^a\right). \tag{6.31}$$

These equations offer an efficient method for practical calculations with the Ricci tensor and scalar.

Calculation 6.2.4.1: Using the tetrad formalism, we will now compute the change in the Ricci tensor and Ricci scalar due to a conformal transformation of the metric.

Let us denote by $\{\boldsymbol{\theta}^a\}$ the dual basis in the metric g. The corresponding dual basis in the metric $\tilde{g} \equiv e^{2\lambda} g$ is $\left\{e^{\lambda}\boldsymbol{\theta}^a\right\}$. Suppose that $\boldsymbol{\omega}_{ab}$ are the connection 1-forms for the basis $\{\boldsymbol{\theta}^a\}$. We now need to compute the modified connection forms $\tilde{\boldsymbol{\omega}}_{ab}$. After that, the modified Riemann and Ricci tensors will be determined.

The connection forms $\tilde{\boldsymbol{\omega}}_{ab}$ are found using the first Cartan structure equation (6.16):

$$\begin{aligned}\mathrm{d}\boldsymbol{\theta}^c &= -\boldsymbol{\omega}^c{}_b \wedge \boldsymbol{\theta}^b, \\ \mathrm{d}\left(e^{\lambda}\boldsymbol{\theta}^c\right) &= -\tilde{\boldsymbol{\omega}}^c{}_b \wedge e^{\lambda}\boldsymbol{\theta}^b.\end{aligned}$$

It follows that

$$(\mathrm{d}\lambda) \wedge \boldsymbol{\theta}_c = -\left(\tilde{\boldsymbol{\omega}}_{cb} - \boldsymbol{\omega}_{cb}\right) \wedge \boldsymbol{\theta}^b.$$

We can now use Eq. (6.23) to express the 1-forms

$$\delta\tilde{\boldsymbol{\omega}}_{cb} \equiv \tilde{\boldsymbol{\omega}}_{cb} - \boldsymbol{\omega}_{cb}$$

explicitly through $\{\boldsymbol{\theta}^a\}$ and $\mathrm{d}\lambda$:

$$\delta\boldsymbol{\omega}_{ab} = \iota_{\mathbf{e}_b}(\mathrm{d}\lambda \wedge \boldsymbol{\theta}_a) - \iota_{\mathbf{e}_a}(\mathrm{d}\lambda \wedge \boldsymbol{\theta}_b) - \frac{1}{2}\iota_{\mathbf{e}_b}\iota_{\mathbf{e}_a}(\boldsymbol{\theta}^c \wedge \mathrm{d}\lambda \wedge \boldsymbol{\theta}_c).$$

The last term vanishes since $\boldsymbol{\theta}^c \wedge \boldsymbol{\theta}_c = 0$, and after a simplification,

$$\iota_{\mathbf{e}_b}(\mathrm{d}\lambda \wedge \boldsymbol{\theta}_a) = \iota_{\mathbf{e}_b}(\mathrm{d}\lambda) \wedge \boldsymbol{\theta}_a - (\mathrm{d}\lambda)\,\eta_{ab} = (\mathbf{e}_b \circ \lambda)\,\boldsymbol{\theta}_a - \eta_{ab}\mathrm{d}\lambda,$$

we find

$$\delta\boldsymbol{\omega}_{ab} = (\mathbf{e}_b \circ \lambda)\,\boldsymbol{\theta}_a - (\mathbf{e}_a \circ \lambda)\,\boldsymbol{\theta}_b. \tag{6.32}$$

The Riemann tensor is expressed through $\boldsymbol{\omega}_{ab}$ via the second Cartan structure equation (6.30). Hence, the curvature 2-forms $\tilde{\boldsymbol{R}}_{ab}$ for the metric $\tilde{g}$ are related to the 2-forms $\boldsymbol{R}_{ab}$ for the metric g by

$$\begin{aligned}\tilde{\boldsymbol{R}}_{ab} &= \boldsymbol{R}_{ab} + \mathrm{d}\delta\boldsymbol{\omega}_{ab} \\ &\quad + \delta\boldsymbol{\omega}_{ac} \wedge \boldsymbol{\omega}^c{}_b + \boldsymbol{\omega}_{ac} \wedge \delta\boldsymbol{\omega}^c{}_b + \delta\boldsymbol{\omega}_{ac} \wedge \delta\boldsymbol{\omega}^c{}_b.\end{aligned}$$

Now we need to substitute $\delta\boldsymbol{\omega}_{ab}$ from Eq. (6.32). The individual terms can be rewritten as follows,

$$\mathrm{d}\delta\boldsymbol{\omega}_{ab} = (\mathrm{d}\,(\mathbf{e}_b \circ \lambda)) \wedge \boldsymbol{\theta}_a + (\mathbf{e}_b \circ \lambda)\,\mathrm{d}\boldsymbol{\theta}_a - [_{a\leftrightarrow b}],$$

where I indicated by $[_{a\leftrightarrow b}]$ the repetition of previous terms with the indices a and b interchanged. Further, we use the first Cartan structure equation to simplify

$$\begin{aligned}\delta\boldsymbol{\omega}_{ac} \wedge \boldsymbol{\omega}^c{}_b &= (\mathbf{e}_c \circ \lambda)\,\boldsymbol{\theta}_a \wedge \boldsymbol{\omega}^c{}_b - (\mathbf{e}_a \circ \lambda)\,\boldsymbol{\theta}_c \wedge \boldsymbol{\omega}^c{}_b \\ &= (\mathbf{e}_c \circ \lambda)\,\boldsymbol{\theta}_a \wedge \boldsymbol{\omega}^c{}_b + (\mathbf{e}_a \circ \lambda)\,\mathrm{d}\boldsymbol{\theta}_b.\end{aligned}$$

Using the identities

$$(\mathbf{e}_c \circ \lambda)\,\boldsymbol{\theta}^c = \boldsymbol{\theta}^c \iota_{\mathbf{e}_c}(\mathrm{d}\lambda) = \mathrm{d}\lambda,$$
$$(\mathbf{e}_c \circ \lambda)\,(\mathbf{e}^c \circ \lambda) = \mathrm{Tr}_{(\mathbf{a},\mathbf{b})}\,(\iota_{\mathbf{a}}\mathrm{d}\lambda)\,(\iota_{\mathbf{b}}\mathrm{d}\lambda) = g^{-1}(\mathrm{d}\lambda, \mathrm{d}\lambda),$$
$$\boldsymbol{\theta}^c \wedge \boldsymbol{\theta}_c = 0,$$

(the first line above follows from Statement 6.1.6.2), we find

$$\begin{aligned}
&\delta\boldsymbol{\omega}_{a\,c} \wedge \delta\boldsymbol{\omega}^c{}_b \\
&= ((\mathbf{e}_c \circ \lambda)\,\boldsymbol{\theta}_a - (\mathbf{e}_a \circ \lambda)\,\boldsymbol{\theta}_c) \wedge ((\mathbf{e}_b \circ \lambda)\,\boldsymbol{\theta}^c - (\mathbf{e}^c \circ \lambda)\,\boldsymbol{\theta}_b) \\
&= [(\mathbf{e}_b \circ \lambda)\,\boldsymbol{\theta}_a - (\mathbf{e}_b \circ \lambda)\,\boldsymbol{\theta}_b] \wedge \mathrm{d}\lambda - g^{-1}(\mathrm{d}\lambda, \mathrm{d}\lambda)\,\boldsymbol{\theta}_a \wedge \boldsymbol{\theta}_b \\
&= \delta\boldsymbol{\omega}_{a\,b} \wedge \mathrm{d}\lambda - g^{-1}(\mathrm{d}\lambda, \mathrm{d}\lambda)\,\boldsymbol{\theta}_a \wedge \boldsymbol{\theta}_b.
\end{aligned}$$

Putting the terms together and simplifying, we obtain

$$\begin{aligned}
\tilde{\boldsymbol{R}}_{a\,b} &= \boldsymbol{R}_{a\,b} + (\mathrm{d}\,(\mathbf{e}_b \circ \lambda)) \wedge \boldsymbol{\theta}_a - (\mathrm{d}\,(\mathbf{e}_a \circ \lambda)) \wedge \boldsymbol{\theta}_b \\
&\quad + (\mathbf{e}_c \circ \lambda)\,[\boldsymbol{\theta}_a \wedge \boldsymbol{\omega}^c{}_b - \boldsymbol{\theta}_b \wedge \boldsymbol{\omega}^c{}_a] \\
&\quad + \delta\boldsymbol{\omega}_{a\,b} \wedge \mathrm{d}\lambda - g^{-1}(\mathrm{d}\lambda, \mathrm{d}\lambda)\,\boldsymbol{\theta}_a \wedge \boldsymbol{\theta}_b. \qquad (6.33)
\end{aligned}$$

At this point we note that the terms $\mathrm{d}\,(\mathbf{e}_c \circ \lambda)$ involve second derivatives of λ. It is useful to introduce explicitly the tensor consisting of the second *covariant* derivatives of λ. This tensor is called the **Hessian** of the function λ and denoted H_λ. The Hessian is a symmetric bilinear form defined by the equivalent formulas

$$H_\lambda(\mathbf{a}, \mathbf{b}) = \iota_{\mathbf{a}}(\nabla_{\mathbf{b}}\mathrm{d}\lambda) = \nabla_{\mathbf{a}}\nabla_{\mathbf{b}}\lambda - \nabla_{\nabla_{\mathbf{a}}\mathbf{b}}\lambda.$$

It follows from the definition that $H_\lambda(\mathbf{a}, \mathbf{b})$ contains only derivatives of λ and of the metric, but no derivatives of $\mathbf{a}$ or $\mathbf{b}$.

Since λ is a given function, we may assume that the Hessian H_λ is a known tensor. We can now express the terms $\mathrm{d}\,(\mathbf{e}_b \circ \lambda)$ through the Hessian. First, we use the Leibnitz rule for the derivative $\nabla_{\mathbf{a}}$ of a 1-form $\mathrm{d}\lambda$ applied to a vector $\mathbf{b}$,

$$\begin{aligned}
\nabla_{\mathbf{a}}\,[(\mathrm{d}\lambda) \circ \mathbf{b}] &= (\nabla_{\mathbf{a}}\mathrm{d}\lambda) \circ \mathbf{b} + (\mathrm{d}\lambda) \circ \nabla_{\mathbf{a}}\mathbf{b} \\
&= H_\lambda(\mathbf{a}, \mathbf{b}) + (\nabla_{\mathbf{a}}\mathbf{b}) \circ \lambda.
\end{aligned}$$

Substituting the basis vectors $\mathbf{a} = \mathbf{e}_a$ and $\mathbf{b} = \mathbf{e}_b$, we find

$$\nabla_{\mathbf{e}_a}\,[(\mathrm{d}\lambda) \circ \mathbf{e}_b] = \iota_{\mathbf{e}_a}\mathrm{d}\,(\mathbf{e}_b \circ \lambda) = H_\lambda(\mathbf{e}_a, \mathbf{e}_b) + (\nabla_{\mathbf{e}_a}\mathbf{e}_b) \circ \lambda.$$

Using Eq. (6.13), we can express $\nabla_{\mathbf{e}_a}\mathbf{e}_b$ through the connection 1-form $\boldsymbol{\omega}_{a\,b}$ and obtain

$$\iota_{\mathbf{e}_a}\mathrm{d}\,(\mathbf{e}_b \circ \lambda) = \iota_{\mathbf{e}_a}\iota_{\mathbf{e}_b}H_\lambda + \iota_{\mathbf{e}_a}\sum_c \boldsymbol{\omega}^c{}_b\,(\mathbf{e}_c \circ \lambda).$$

By dropping the insertion $\iota_{\mathbf{e}_a}$ of the arbitrary basis vector $\mathbf{e}_a$, we can rewrite the above as an equality of 1-forms,

$$\mathrm{d}\,(\mathbf{e}_b \circ \lambda) = \iota_{\mathbf{e}_b}H_\lambda + \boldsymbol{\omega}^c{}_b\,(\mathbf{e}_c \circ \lambda).$$

This relationship is now used to simplify Eq. (6.33) further,

$$\begin{aligned}
\tilde{\boldsymbol{R}}_{a\,b} &= \boldsymbol{R}_{a\,b} + (\iota_{\mathbf{e}_b}H_\lambda) \wedge \boldsymbol{\theta}_a - (\iota_{\mathbf{e}_a}H_\lambda) \wedge \boldsymbol{\theta}_b \\
&\quad + \delta\boldsymbol{\omega}_{a\,b} \wedge \mathrm{d}\lambda - g^{-1}(\mathrm{d}\lambda, \mathrm{d}\lambda)\,\boldsymbol{\theta}_a \wedge \boldsymbol{\theta}_b. \qquad (6.34)
\end{aligned}$$

Using Eq. (6.34), we can now compute the Ricci tensor,

$$\tilde{\mathrm{Ric}}_b = \iota_{\tilde{\mathbf{e}}^a}\tilde{\boldsymbol{R}}_{b\,a} = e^{-\lambda}\iota_{\mathbf{e}^a}\tilde{\boldsymbol{R}}_{b\,a}.$$

The following auxiliary calculations are helpful,

$$\begin{aligned}
\iota_{\mathbf{e}^a}(\boldsymbol{\theta}_a \wedge \boldsymbol{\theta}_b) &= (N-1)\,\boldsymbol{\theta}_b; \\
\iota_{\mathbf{e}^a}\delta\boldsymbol{\omega}_{a\,b} &= \iota_{\mathbf{e}^a}\,[(\mathbf{e}_b \circ \lambda)\,\boldsymbol{\theta}_a - (\mathbf{e}_a \circ \lambda)\,\boldsymbol{\theta}_b] \\
&= (N-1)\,(\mathbf{e}_b \circ \lambda); \\
(\mathbf{e}^a \circ \lambda)\,\delta\boldsymbol{\omega}_{a\,b} &= (\mathbf{e}^a \circ \lambda)\,((\mathbf{e}_b \circ \lambda)\,\boldsymbol{\theta}_a - (\mathbf{e}_a \circ \lambda)\,\boldsymbol{\theta}_b) \\
&= (\mathbf{e}_b \circ \lambda)\,\mathrm{d}\lambda - g^{-1}(\mathrm{d}\lambda, \mathrm{d}\lambda)\boldsymbol{\theta}_b;
\end{aligned}$$

where N is the dimension of the spacetime,

$$N = \iota_{\mathbf{e}^a}\boldsymbol{\theta}_a.$$

Then we find

$$\begin{aligned}
e^{\lambda}\tilde{\mathrm{Ric}}_b = \iota_{\mathbf{e}^a}\tilde{\boldsymbol{R}}_{b\,a} &= \mathrm{Ric}_b \\
&\quad + \iota_{\mathbf{e}^a}\left[(\iota_{\mathbf{e}_a}H_\lambda) \wedge \boldsymbol{\theta}_b - (\iota_{\mathbf{e}_b}H_\lambda) \wedge \boldsymbol{\theta}_a\right] \\
&\quad + \iota_{\mathbf{e}^a}\left[\delta\boldsymbol{\omega}_{b\,a} \wedge \mathrm{d}\lambda - g^{-1}(\mathrm{d}\lambda, \mathrm{d}\lambda)\,\boldsymbol{\theta}_b \wedge \boldsymbol{\theta}_a\right] \\
&= \mathrm{Ric}_b + (N-2)\,\iota_{\mathbf{e}_b}H_\lambda + H_\lambda(\mathbf{e}_a, \mathbf{e}^a)\boldsymbol{\theta}_b \\
&\quad - (N-2)\,(\mathbf{e}_b \circ \lambda)\,\mathrm{d}\lambda + (N-2)\,g^{-1}(\mathrm{d}\lambda, \mathrm{d}\lambda)\boldsymbol{\theta}_b.
\end{aligned}$$

The trace of the Hessian H_λ appearing in the last expression is traditionally called the (covariant) **D'Alembertian** of λ and denoted by $\Box\lambda$,

$$\mathrm{Tr}_{(\mathbf{a},\mathbf{b})}H_\lambda(\mathbf{a}, \mathbf{b}) = H_\lambda(\mathbf{e}_a, \mathbf{e}^a) \equiv \Box\lambda.$$

The Ricci scalar is given by

$$\tilde{R} = \iota_{\tilde{\mathbf{e}}^b}\tilde{\mathrm{Ric}}_b = e^{-\lambda}\iota_{\mathbf{e}^b}\tilde{\mathrm{Ric}}_b,$$

so we finally obtain

$$\begin{aligned}
e^{2\lambda}\tilde{R} &= \iota_{\mathbf{e}^b}e^{\lambda}\tilde{\mathrm{Ric}}_b \\
&= R + 2\,(N-1)\,\Box\lambda + (N-1)\,(N-2)\,g^{-1}(\mathrm{d}\lambda, \mathrm{d}\lambda).
\end{aligned}$$

The formulas for $\tilde{\mathrm{Ric}}$ and $\tilde{R}$ coincide with those obtained in Calculation 1.8.4.1. The calculation in the tetrad formalism is somewhat shorter and less tedious than the calculation in vector notation (see details on page 177). The trade-off is that one needs to introduce and use new techniques, such as connection 1-forms, the exterior product, and the insertion operator. ■

6.2.5 Einstein-Hilbert action in tetrads

The dynamics of General Relativity is described by Einstein equations, which are derived from the action principle. In Section 5.1.2 the action was defined as a functional of the metric tensor g. The metric was treated as a dynamical variable that satisfies certain equations of motion.

An alternative approach is to regard the tetrad of 1-forms $\{\boldsymbol{\theta}^a\}$ as the primary dynamical variable. The metric tensor can be expressed through $\{\boldsymbol{\theta}^a\}$ as $g = \sum_a \boldsymbol{\theta}^a \otimes \boldsymbol{\theta}_a$. However, it is also useful to work directly with the tetrad, without using the metric tensor. In this section I will explain how the Einstein-Hilbert action is expressed and the Einstein equations are derived in the tetrad formalism.[5]

The standard way of expressing the Einstein-Hilbert action in local coordinates is

$$S_{\mathrm{EH}} = \frac{1}{16\pi G}\int R\sqrt{-g}\mathrm{d}^4x,$$

[5] I learned some relevant details regarding this derivation from the course notes [15], lecture 15.

where R is the Ricci scalar and G is Newton's constant. This formula can be rewritten with help of the volume 4-form,

$$S_{\text{EH}} = \frac{1}{16\pi G}\int_{\mathcal{M}} R\,\text{Vol}.$$

Now we would like to express the 4-form $R\,\text{Vol}$ in terms of the tetrad $\{\theta^a\}$. This can be done using the following trick. We note that R is a double trace of the Riemann tensor, and that a multiple trace can be seen in Eq. (6.7). Therefore, it will be possible to express the 4-form $R\,\text{Vol}$ directly through the curvature 2-forms. Using Eq. (6.31) and Statement 6.1.3.2(d), we find

$$R\,\text{Vol} = R_{ab}(\mathbf{e}^b, \mathbf{e}^a)\,\text{Vol} = R_{ab}\wedge *(\theta^b\wedge\theta^a).$$

Using the relationship (see Sec. 6.1.3)

$$*(\theta^a\wedge\theta^b) = \frac{1}{2!}\varepsilon^{abcd}\theta_c\wedge\theta_d,$$

we obtain

$$R\,\text{Vol} = -\frac{1}{2}\varepsilon^{abcd}R_{ab}\wedge\theta_c\wedge\theta_d.$$

Hence, the Einstein-Hilbert action can be rewritten as

$$\begin{aligned} S_{\text{EH}} &= \frac{1}{16\pi G}\int_{\mathcal{M}} R_{ab}\wedge *\left(\theta^b\wedge\theta^a\right) \\ &= -\frac{1}{32\pi G}\int_{\mathcal{M}} \varepsilon^{abcd}R_{ab}\wedge\theta_c\wedge\theta_d. \end{aligned}$$

6.3 Connections on vector bundles

6.3.1 Vector bundles as generalization of tangent bundles

The tangent bundle to a manifold can be viewed as a collection of vector spaces $T_p\mathcal{M}$ attached to each point p of the manifold $\mathcal{M}$. These vector spaces are attached to points in a special way, so that the relation between tangent spaces $T_p\mathcal{M}$ and $T_{p'}\mathcal{M}$ for neighbor points p, p' can be nontrivial.

This construction is generalized by replacing the tangent space $T_p\mathcal{M}$ by a different, perhaps more complicated, vector space V. The result is the concept of a **vector bundle**, which is the union of copies of vector spaces $V(p)$ attached to each point p of a manifold $\mathcal{M}$. All the vector spaces $V(p)$ should have the same dimension. The manifold $\mathcal{M}$ is then called the **base manifold** and the vector space $V(p)$ is called the **fiber** of the bundle at point p. A function $\phi : p \to V(p)$ on $\mathcal{M}$ with values in the fibers $V(p)$ is called a **section** of the bundle. For example, sections of the tangent bundle $T\mathcal{M}$ are *vector fields* $\mathbf{v}(p)$ since the value of a vector field at a point p is a vector from the space $T_p\mathcal{M}$. Sections of more general vector bundles can be visualized as V-valued functions $f(p)$ on the manifold, such that the value of $f(p)$ belongs to the fiber space $V(p)$.

The reason we need to consider more general vector bundles than $T\mathcal{M}$ is that it is convenient to represent physical fields in (classical) field theory as sections of vector bundles. Fields are usually vector-valued (or tensor-valued) functions of spacetime, e.g. a Dirac spinor field ψ represents particles of spin $\frac{1}{2}$ and has values in a four-dimensional complex vector space. When we represent a field by a section of a vector bundle, the spacetime is the base manifold $\mathcal{M}$ and the value of the field at a point belongs to some vector space which is the fiber at that point. For example, spinor fields $\psi(p)$ can be thought of as sections of a "spinor bundle" with fibers $\mathbb{C}^4$; tensor fields $A^{\mu\nu}(p)$ are sections of a vector bundle with the fibers $T_p\mathcal{M}\otimes T_p\mathcal{M}$. Vector bundles give us a mathematical construction that describes all the physical fields in a unified language.

6.3.2 Examples of bundles

A direct product $\mathcal{M}\times V$, where $\mathcal{M}$ is any manifold and V is any vector space, is obviously a vector bundle; such bundles are called **trivial**. Many of the vector bundles used in physics are trivial, but sometimes the bundles turn out to be nontrivial, i.e. not isomorphic to a direct product of a base and a fiber.

For example, consider the tangent bundle TS^2 to a sphere S^2. This bundle has the fiber $\mathbb{R}^2$ and the base S^2. If the bundle TS^2 were trivial, one would be able to find a **trivialization** of TS^2, i.e. a smooth, one-to-one map $\mu : S^2\times\mathbb{R}^2 \to TS^2$ which preserves the linear structure in the fibers. The existence of such a map would imply that the image of a fixed nonzero vector from $\mathbb{R}^2$, say the basis vector $\mathbf{e}_1$ with components $(1,0)$, is a smooth and everywhere nonzero vector field $\mu(p;\mathbf{e}_1)$ on S^2. However, such a vector field does not exist, as shown by the following statement, which is a standard theorem of topology ("one cannot comb a sphere").

Statement 6.3.2.1: There is no smooth, everywhere nonzero tangent vector field to a sphere S^2. (Proof on this page.) ■

Proof of Statement 6.3.2.1: Consider a stereographic projection of the sphere S^2 onto a plane. The sphere is realized as the set $\{(x,y,z) : x^2+y^2+z^2 = R^2\}$ in $\mathbb{R}^3$, and the projection is described explicitly by

$$(x,y,z) \to \frac{x+iy}{R-z} \equiv \zeta,$$

where the plane is parametrized by a single complex coordinate ζ. It is clear that the north pole $(0,0,R)$ is mapped onto infinity,

$$\lim_{x,y\to 0}\frac{x+iy}{R-\sqrt{R^2-x^2-y^2}} = \infty,$$

while the south pole $(0,0,-R)$ is mapped onto $\zeta = 0$ (the center of the plane). Suppose we have a smooth and everywhere nonzero vector field $\mathbf{v}$ on the sphere; then the image of $\mathbf{v}$ under the projection is a smooth, everywhere nonzero vector field $\tilde{\mathbf{v}}$ on the plane. By the smoothness assumption, $\mathbf{v}$ is almost constant and nonzero in a sufficiently small neighborhood of the north pole. We may trace the direction of $\mathbf{v}$ around a small closed contour γ surrounding the pole, and compute the accumulated angle swept by $\mathbf{v}$. This accumulated angle will be zero since $\mathbf{v}$ is almost constant. However, after the stereographic projection the small contour γ will become a large closed contour γ_1 on the plane. Since the interior of γ on the sphere will be mapped to the exterior of γ_1 on the plane, the direction of $\mathbf{v}$ is mirrored and the angle swept by $\tilde{\mathbf{v}}$ when traced around $\tilde{\gamma}$ will be 4π. However, the contour γ_1 can be continuously deformed into a small contour γ_2 surrounding the center $\zeta = 0$ of the plane. The angle swept by $\tilde{\mathbf{v}}$ around γ_2 is zero since $\tilde{\mathbf{v}}$ must be approximately constant in a sufficiently small neighborhood of $\zeta = 0$. However, the angle swept by a vector field around a closed contour is always an integer multiple of 2π and hence must remain constant if $\tilde{\gamma}$ is deformed continuously and if $\tilde{\mathbf{v}}$ is smooth and everywhere nonzero. Thus the angle swept by $\tilde{\mathbf{v}}$ around γ_1 is zero and

cannot be equal to 4π. This gives a contradiction. Therefore, either the vector field $\mathbf{v}$ is not smooth around the poles (invalidating the first step of the proof), or $\mathbf{v}$ goes to zero or is non-smooth at another point, which makes the accumulated angle jump discontinuously from 4π to 0 during the deformation of the contours $\gamma_1 \to \gamma_2$. ■

So it follows from Statement 6.3.2.1 that the tangent bundle TS^2 is nontrivial. However, the tangent bundle of a 3-sphere, TS^3, is trivial. To show this, we first note that the sphere S^3 is the same manifold as the special unitary group $SU(2)$ (see the discussion in Sec. 1.2.2). Tangent vectors on $SU(2)$ correspond to infinitesimal unitary transformations, which are always of the form $1 + h$, where h is a "small" anti-Hermitian matrix, $h = -h^\dagger$. The space of all anti-Hermitian matrices is isomorphic to $\mathbb{R}^3$. A trivialization that maps TS^3 into $S^3 \times \mathbb{R}^3$ can be described explicitly as follows.[6] A pair $(A, h) \in SU(2) \times \mathbb{R}^3$, where both are represented by 2×2 matrices, is mapped into the tangent vector

$$\mathbf{v} \circ f(g) \equiv \left.\frac{\partial}{\partial s}\right|_{s=0} f\left(A(1+sh)A^{-1}\right),$$

where $f(A)$ is a function on $\mathcal{M} \equiv SU(2)$, $A \in SU(2)$, h is a 2×2 anti-Hermitian matrix, and $1 + sh$ is a 2×2 matrix representing an "infinitesimal" unitary transformation for "small" s.

6.3.3 Covariant derivatives on vector bundles

A covariant derivative on a vector bundle is a generalization of directional derivative, so that for a section ϕ and a vector field $\mathbf{u}$, the derivative $\nabla_{\mathbf{u}}\phi$ is another section. We have seen that the covariant derivative ∇ acts on various tensors in different ways (the covariant derivatives of a vector and a tensor are given by different formulae). Using the picture of bundles, we say that tensors of different ranks are sections of different bundles, and it is natural that different bundles have different covariant derivatives. (Up to now, we have been using the symbol ∇ to mean different connections, depending on the rank of tensor it acts on.)

A general way to write the covariant derivative on a vector bundle is

$$\nabla_{\mathbf{u}}\mathbf{t} = \partial_{\mathbf{u}}\mathbf{t} + \Gamma(\mathbf{u})\mathbf{t},$$

where ∂ is the "partial derivative" connection defined using a particular local coordinate system on the base manifold, i.e. $\partial_{\mathbf{u}} \equiv u^\mu \partial_\mu$, and Γ is a transformation-valued 1-form. The value of Γ is a linear transformation within a fiber and Γ is called the **connection 1-form**.

6.3.4 Gauge theories and associated bundles

Gauge field theories (such as electrodynamics, electroweak theory, and chromodynamics) use vector bundles with an additional structure. Namely, a certain Lie group G is acting in each fiber, such that the equations of the theory are invariant under local transformations of the group. For example, the gauge group G is $U(1)$ for electromagnetism, $SU(2)$ for the electroweak theory, and $SU(3)$ for chromodynamics. The group G is then called the **gauge group** and the vector bundle is called **associated** to this group (this is not a rigorous definition but is intended only as a qualitative hint). For an associated bundle, the connection 1-form Γ has values in a representation of the Lie algebra $\mathfrak{g}$ of the group G. This of course limits the possible connections Γ, leaving only the physically relevant ones. **Why not give a more rigorous definition?**

6.3.5 Tangent bundle as associated bundle

Let us try to see if the tangent bundle $T\mathcal{M}$ can be also viewed as a bundle of which a gauge group acts. The gauge group in that case would be the orthogonal group $SO(n)$ that acts fiberwise, preserving the metric g in each fiber. (For a metric with the Lorentzian signature, the group will be $SO(3,1)$, called also the **Lorentz group**.) The connection 1-form Γ is then required to give a transformation which is one of the transformations from the representation of the Lie algebra $\mathfrak{so}(n)$, i.e. for any vector $\mathbf{v}$ the transformation $\Gamma(\mathbf{v})$ must be an antisymmetric linear transformation such that

$$g(\Gamma(\mathbf{v})\mathbf{x}, \mathbf{y}) + g(\mathbf{x}, \Gamma(\mathbf{v})\mathbf{y}) = 0.$$

It is easy to see that this condition coincides with the condition (1.41) expressing the compatibility of the connection $\nabla = \partial + \Gamma$ with the metric. Therefore the statement that the connection ∇ on the tangent bundle comes from the orthogonal gauge group is the same as the assumption of the compatibility of ∇ with the metric. This is just one example of how physical assumptions can be expressed in the language of gauge groups. **This explanation needs to be made more clear. What exactly is invariant under a gauge transformation? Is Γ invariant?? (It isn't.) In fact, the above equation $g(\Gamma(\mathbf{v})\mathbf{x}, \mathbf{y}) + g(\mathbf{x}, \Gamma(\mathbf{v})\mathbf{y}) = 0$ seems to be incorrect!!!**

[6] It was proved in differential topology [4] that the only spheres with trivial tangent bundles are S^1, S^3, and S^7. Moreover, no even-dimensional sphere S^{2n} admits a continuous and everywhere nonzero tangent vector field [30]. (This information is not actually used in GR.)

7 Spinors

Additional literature: [27, 32].

The purpose of this chapter is to explain the definition of spinors and spinor fields, and to show some basic uses of spinors in general relativity.

7.1 Introducing spinors

A **spinor** is a two-dimensional complex-valued vector, $\alpha \in S$, where $S \equiv \mathbb{C}^2$ is an auxiliary space (the space of spinors).[1] The necessity to introduce spinors became apparent only after the development of quantum mechanics: It turned out that particles such as electrons or quarks must be described by spinor-valued fields, rather than by vector or tensor-valued fields. Below I shall explain the motivation for introducing spinors in more detail. Right now, I only give a rough heuristic picture.

Vectors can be visualized as "directed magnitudes," which literally means a direction along a curve leading from a point of the manifold to a neighbor point. Tensors are defined algebraically through vectors. Every tensor remains unchanged under a rotation by 2π because every point returns to its original place, and tensors are defined through directions between points. Now, it turns out that one can define a special kind of "directed magnitude" called a spinor that remains unchanged only under a 4π rotation but *changes sign* under a 2π rotation. Clearly, such an object cannot be a vector or a tensor, or any other object defined through points of a manifold, because every point in the neighborhood of a given center of rotation will return to its initial position after a 2π rotation. Thus, a spinor cannot be defined through vectors of the tangent space. Spinors belong to some auxiliary vector space in which the rotation group acts in a special way. To understand how a rotation can act in such an unusual way, we need to study the rotation group in some more detail.

7.1.1 Definition of quaternions

Let us consider rotations in a three-dimensional Euclidean space $\mathbb{R}^3$. The set of all orthogonal transformations of $\mathbb{R}^3$, i.e. the set of all linear transformations $\hat{R}$ such that $\hat{R}\hat{R}^T = \hat{1}$, is a group denoted by $O(3)$. The subgroup denoted by $SO(3)$ contains **proper rotations**, i.e. orthogonal transformations preserving the orientation ($\det \hat{R} = 1$). This way of describing rotations is tedious: one needs to specify 9 components of a matrix of $\hat{R}$ in a certain basis, while the group $O(3)$ is merely three-dimensional.

A more concise way of representing rotations in $\mathbb{R}^3$ is through quaternions. Quaternions are a generalization of complex numbers where one introduces three "imaginary units" instead of one. I will now show how quaternions describe rotations, and I will also use quaternions to define spinors. (In brief: The four-dimensional space of quaternions is equivalent to the complex two-dimensional spinor space). Let us now briefly explore the properties of quaternions.

The set of **quaternions** $\mathbb{H}$ is a four-dimensional real space spanned by the basis vectors $\mathbf{1}, \mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3$. In other words, $\mathbb{H}$ is the set of linear combinations of the form

$$\mathbf{x} = x_0\mathbf{1} + x_1\mathbf{h}_1 + x_2\mathbf{h}_2 + x_3\mathbf{h}_3,$$

where x_j are real coefficients. The special elements $\mathbf{h}_j$ are called **quaternionic units**[2].

The multiplication of quaternions will be defined once we define the multiplication of the quaternionic units. The quaternionic units obey, by definition, the multiplication rules that can be written concisely as follows:

$$\mathbf{h}_1 * \mathbf{h}_1 = \mathbf{h}_2 * \mathbf{h}_2 = \mathbf{h}_3 * \mathbf{h}_3 = \mathbf{h}_1 * \mathbf{h}_2 * \mathbf{h}_3 = -\mathbf{1}. \tag{7.1}$$

It can be derived from these multiplication rules that

$$\mathbf{h}_j * \mathbf{h}_k = -\delta_{jk}\mathbf{1} + \sum_{l=1}^{3} \varepsilon_{jkl}\mathbf{h}_l, \tag{7.2}$$

where δ is the Kronecker symbol and ε_{jkl} is the coordinate representation of the three-dimensional Levi-Civita symbol, $\varepsilon_{123} = 1$. In particular,

$$\mathbf{h}_1 * \mathbf{h}_2 = \mathbf{h}_3 = -\mathbf{h}_2 * \mathbf{h}_1.$$

With these definitions, the quaternionic space $\mathbb{H}$ becomes a non-commutative algebra.

Statement: Assuming Hamilton's quaternionic multiplication rule (7.1), the associativity and distributivity of quaternionic multiplication, and the property $\mathbf{1} * \mathbf{x} = \mathbf{x} * \mathbf{1} = \mathbf{x}$ for any quaternion $\mathbf{x}$, one can derive Eq. (7.2).

Proof: Consider the expression $\mathbf{h}_1 * \mathbf{h}_2 * \mathbf{h}_3 * \mathbf{h}_3 = -\mathbf{h}_1 * \mathbf{h}_2 = -\mathbf{h}_3$. Similarly, we find $\mathbf{h}_2 * \mathbf{h}_3 = \mathbf{h}_1$. Consider the expression $\mathbf{h}_2 * \mathbf{h}_1 * \mathbf{h}_2 * \mathbf{h}_3 = -\mathbf{h}_2$; after multiplying by $\mathbf{h}_3 * \mathbf{h}_2$ on the right, it follows that

$$\mathbf{h}_2 * \mathbf{h}_1 = -\mathbf{h}_2 * \mathbf{h}_3 * \mathbf{h}_2 = -\mathbf{h}_1 * \mathbf{h}_2 = -\mathbf{h}_3.$$

Consider the expression $\mathbf{h}_3 * \mathbf{h}_1 * \mathbf{h}_2 * \mathbf{h}_3 = -\mathbf{h}_3$; it follows that $\mathbf{h}_3 * \mathbf{h}_1 = \mathbf{h}_2$. All other required relations are obtained in a similar way. ■

It is sometimes useful to visualize a quaternion $\mathbf{x} \equiv x_0\mathbf{1} + \sum_{j=1}^{3} x_j\mathbf{h}_j$ as a formal sum $x_0 + \vec{x}$ of a scalar x_0 and a 3-vector $\vec{x} \equiv (x_1, x_2, x_3)$. Then the quaternionic multiplication law (7.2) is equivalent to the rule

$$(x_0 + \vec{x}) * (y_0 + \vec{y}) = (x_0y_0 - \vec{x} \cdot \vec{y}) + (x_0\vec{y} + y_0\vec{x} + \vec{x} \times \vec{y}). \tag{7.3}$$

The **conjugate** quaternion to $\mathbf{x}$ is defined by

$$\bar{\mathbf{x}} \equiv x_0\mathbf{1} - \sum_{j=1}^{3} x_j\mathbf{h}_j.$$

[1] These two-dimensional objects are also called **Weyl spinors**, in distinction from **Dirac spinors** that have four complex components and are defined using Weyl spinors. The connection between the two will be explained below, and we say "spinor" instead of "Weyl spinor."

[2] A more traditional notation for quaternionic units is $\mathbf{i}, \mathbf{j}, \mathbf{k}$ instead of $\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3$, but I would like to avoid using $\mathbf{i}$ because I will later need to mix quaternions with complex numbers with their own imaginary unit, i.

It is straightforward to verify that the conjugation operation interacts with quaternionic multiplication as[3]

$$\overline{(\mathbf{x} * \mathbf{y})} = \overline{\mathbf{y}} * \overline{\mathbf{x}}.$$

A quaternion $\mathbf{x}$ is called **purely imaginary** if $\bar{\mathbf{x}} = -\mathbf{x}$, which means that the first coefficient vanishes, $x_0 = 0$. The **magnitude** $|\mathbf{x}|$ of a quaternion $\mathbf{x}$ is

$$|\mathbf{x}| \equiv \left(x_0^2 + x_1^2 + x_2^2 + x_3^2\right)^{1/2}.$$

It is easy to see that $\mathbf{x} * \bar{\mathbf{x}} = |\mathbf{x}|^2 \mathbf{1}$ and $|\mathbf{x} * \mathbf{y}| = |\mathbf{x}|\,|\mathbf{y}|$, similarly to the properties of complex numbers.

Calculation: To verify these properties.

Solution: We use the multiplication rule (7.3) in the vector notation. Since $\bar{\mathbf{x}} = x_0\mathbf{1} - \vec{x}$, it follows that $\mathbf{x} * \bar{\mathbf{x}} = |\mathbf{x}|^2 \mathbf{1}$. The property $|\mathbf{x} * \mathbf{y}|^2 = |\mathbf{x}|^2 |\mathbf{y}|^2$ is straightforward to derive, recalling that the vector product $\vec{x} \times \vec{y}$ is orthogonal to both $\vec{x}$ and $\vec{y}$, and using the formula $|\vec{x} \times \vec{y}|^2 = |\vec{x}|^2 |\vec{y}|^2 - |\vec{x} \cdot \vec{y}|^2$.

An important property is that every nonzero quaternion $\mathbf{x}$ has an inverse $\mathbf{x}^{-1}$ such that $\mathbf{x} * \mathbf{x}^{-1} = \mathbf{x}^{-1} * \mathbf{x} = \mathbf{1}$.

Calculation: Find the inverse $\mathbf{x}^{-1}$ to a general quaternion $\mathbf{x} = x_0\mathbf{1} + \sum_{j=1}^{3} x_j \mathbf{h}_j$, assuming $\mathbf{x} \neq 0$. Show that the inverse would not necessarily exist if we admitted complex coefficients x_j.

Solution: Since $\mathbf{x} * \bar{\mathbf{x}} = |\mathbf{x}|^2 \mathbf{1}$, we have

$$\mathbf{x}^{-1} = \frac{\bar{\mathbf{x}}}{|\mathbf{x}|^2} = \frac{1}{|\mathbf{x}|^2}\left(x_0\mathbf{1} - x_1\mathbf{h}_1 - x_2\mathbf{h}_2 - x_3\mathbf{h}_3\right).$$

The inverse exists when $|\mathbf{x}|^2 = \sum_{j=0}^{4} x_j^2 \neq 0$, which is guaranteed for quaternions with real coefficients x_j.

The quaternionic multiplication law also admits a representation in terms of complex 2×2 matrices. We may guess the required representation if we note that the multiplication rule of the **Pauli matrices**,

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

$$\sigma_j \sigma_k = \delta_{jk} + \mathrm{i} \sum_{l=1}^{3} \varepsilon_{jkl} \sigma_l,$$

is very similar to the multiplication rule of the quaternionic units $\mathbf{h}_j$, except for the extra factors of i. Therefore a representation for $\mathbf{h}_j$ by 2×2 complex matrices may be found as $\hat{\mathbf{h}}_j = -\mathrm{i}\sigma_j$, $j = 1, 2, 3$, assuming (naturally) that the identity matrix $\hat{1}$ is used for the quaternion $\mathbf{1}$. The general quaternion $\mathbf{a} = a_0\mathbf{1} + \sum_{j=1}^{3} a_j \mathbf{h}_j$ will be represented by the matrix

$$\hat{\mathbf{a}} = \begin{pmatrix} a_0 - \mathrm{i}a_3 & -\mathrm{i}a_1 - a_2 \\ -\mathrm{i}a_1 + a_2 & a_0 + \mathrm{i}a_3 \end{pmatrix}. \tag{7.4}$$

Note that the determinant of this matrix is equal to $|\mathbf{a}|^2$, and that the Hermitian conjugate matrix corresponds to $\bar{\mathbf{a}}$:

$$\det \hat{\mathbf{a}} = |\mathbf{a}|^2 = a_0^2 + a_1^2 + a_2^2 + a_3^2; \quad \hat{\mathbf{a}}^\dagger = \hat{\bar{\mathbf{a}}}.$$

[3] In mathematical language, the conjugation operator is an anti-automorphism of the algebra of quaternions.

Remark: As an additional motivation, it is helpful to derive a matrix representation of quaternions without guessing the Pauli matrices. The four-dimensional space $\mathbb{H}$ of quaternions can be interpreted as a two-dimensional *complex* vector space if we somehow define what it means to multiply a quaternion by a complex number,

$$(\lambda + \mathrm{i}\mu)\,\mathbf{a} = ???$$

Clearly, it is sufficient to define the multiplication by i, and we need to imitate the property $\mathrm{i}^2 = -1$. Since any of the unit quaternions $\mathbf{h}_j$ satisfies $\mathbf{h}_j^2 = -\mathbf{1}$, we may choose for instance $\mathbf{h}_1$ as the representative of i, so then we define

$$(\lambda + \mathrm{i}\mu)\,\mathbf{a} \equiv \lambda\mathbf{a} + \mu\mathbf{a} * \mathbf{h}_1.$$

This is clearly a linear map in $\mathbb{H}$ (note that we multiply by $\mathbf{h}_1$ from the *right*). We expect that $\mathbb{H}$, considered as a complex vector space, will be two-dimensional, so any quaternion $\mathbf{a} \in \mathbb{H}$ can be represented as a linear combination of two basis quaternions with complex coefficients. We may choose, for instance, $\{\mathbf{1}, \mathbf{h}_2\}$ as the basis, since $\mathbf{h}_1$ is already chosen as the representative of i. (Different such choices will yield different but algebraically equivalent representations.) Then an arbitrary quaternion will be decomposed as

$$\mathbf{a} \equiv a_0\mathbf{1} + \sum_{j=1}^{3} a_j \mathbf{h}_j = (a_0 + \mathrm{i}a_1)\,\mathbf{1} + (a_2 - \mathrm{i}a_3)\,\mathbf{h}_2$$

and so can be represented by an array of two complex numbers $(a_0 + \mathrm{i}a_1, a_2 - \mathrm{i}a_3) \in \mathbb{C}^2$. Since the multiplication by a fixed quaternion $\mathbf{x}$ is a linear transformation of the vector space $\mathbb{C}^2$, it is natural to represent that transformation by a 2×2 complex matrix. For example, the transformation $\mathbf{a} \to \mathbf{h}_3 * \mathbf{a}$ will be equivalent to

$$\begin{aligned} \mathbf{h}_3 * \begin{pmatrix} a_0 + \mathrm{i}a_1 \\ a_2 - \mathrm{i}a_3 \end{pmatrix} &= \mathbf{h}_3 * (a_0\mathbf{1} + a_1\mathbf{h}_1 + a_2\mathbf{h}_2 + a_3\mathbf{h}_3) \\ &= -a_3\mathbf{1} - a_2\mathbf{h}_1 + a_1\mathbf{h}_2 + a_0\mathbf{h}_3 \\ &= \begin{pmatrix} -a_3 - \mathrm{i}a_2 \\ a_1 - \mathrm{i}a_0 \end{pmatrix} \\ &= \begin{pmatrix} 0 & -\mathrm{i} \\ -\mathrm{i} & 0 \end{pmatrix} \begin{pmatrix} a_0 + \mathrm{i}a_1 \\ a_2 - \mathrm{i}a_3 \end{pmatrix}, \end{aligned}$$

therefore $\mathbf{h}_3$ is represented by the matrix $-\mathrm{i}\sigma_1$. Similarly, we find $\mathbf{h}_1 \to i\sigma_3$ and $\mathbf{h}_2 \to -\mathrm{i}\sigma_2$. In this way we find a possible representation of quaternions through Pauli matrices. (This representation is equivalent to the standard one, $\mathbf{h}_j \to -\mathrm{i}\sigma_j$, after a permutation of the unit quaternions.)

7.1.2 Quaternions and rotations

The connection between quaternions and rotations is found by using the following method: We first consider an infinitesimal rotation, represent it through quaternionic multiplication, and then exponentiate the infinitesimal rotation to obtain a finite rotation.

A **rotation** $\hat{R}$ is a linear operator that preserves the scalar product:

$$\forall \mathbf{u}, \mathbf{v} : g(\hat{R}\mathbf{v}, \hat{R}\mathbf{u}) = g(\mathbf{v}, \mathbf{u}). \tag{7.5}$$

To define a rotation, it is actually sufficient to demand that a linear operator preserves the norm of each single vector,

$$\forall \mathbf{x} : g(\hat{R}\mathbf{x}, \hat{R}\mathbf{x}) = g(\mathbf{x}, \mathbf{x}),$$

because the substitution $\mathbf{x} = \mathbf{u} + \mathbf{v}$ allows us to derive Eq. (7.5).

An **infinitesimal rotation** is a linear operator $\hat{S}$ is such that $g(\mathbf{v}, \hat{S}\mathbf{v}) = 0$ for all $\mathbf{v}$ (that is to say, the operator $\hat{S}$ must be antisymmetric with respect to the metric g). The picture behind this is that of transforming a vector $\mathbf{v}_0$ by

$$\mathbf{v}_0 \mapsto \mathbf{v}_0 + \lambda \hat{S}\mathbf{v}_0,$$

where the formal parameter λ is, heuristically, "very small" and describes the "infinitesimal" angle of the rotation. Given an infinitesimal rotation operator $\hat{S}$, one can obtain the corresponding finite rotation with angle λ. This is done by considering the vector $\mathbf{v}$ rotated by angle λ as a vector-valued function $\mathbf{v}(\lambda)$, so that $\mathbf{v}(0) = \mathbf{v}_0$, and then writing the "infinitesimal rotation"

$$\mathbf{v}(\lambda + \delta\lambda) = \mathbf{v}(\lambda) + \delta\lambda\, \hat{S}\mathbf{v}(\lambda)$$

in the limit $\delta\lambda \to 0$, as the differential equation

$$\frac{d}{d\lambda}\mathbf{v}(\lambda) = \hat{S}\mathbf{v}(\lambda), \quad \mathbf{v}(0) = \mathbf{v}_0.$$

The solution is

$$\mathbf{v}(\lambda) = \exp(\lambda\hat{S})\mathbf{v}_0.$$

Infinitesimal rotations can be represented by quaternionic multiplication through the following trick: If $\mathbf{q} = q_0 + \vec{q}$ is a quaternion and $\vec{v}$ is a 3-dimensional vector, understood also as a (purely imaginary) quaternion, then the quaternionic commutator

$$[\mathbf{q}, \vec{v}] \equiv \mathbf{q} * \vec{v} - \vec{v} * \mathbf{q} = 2\vec{q} \times \vec{v}$$

yields a vector that is always orthogonal to $\vec{v}$. Thus the quaternionic commutator with a fixed quaternion $\mathbf{q}$ is an infinitesimal rotation. Geometrically, this is a rotation around the axis given by the vector $\vec{q}$, since the infinitesimal change of $\vec{v}$ is proportional to $\vec{q} \times \vec{v}$. The factor 2 means that an infinitesimal rotation with parameter λ corresponds to a physical rotation by angle 2λ. The scalar part q_0 of the quaternion $\mathbf{q}$ does not play any role in the commutator; therefore, let us suppose that $\mathbf{q}$ is purely imaginary, $\mathbf{q} = \vec{q}$.

It remains to exponentiate this infinitesimal rotation to a finite rotation. This is accomplished with the standard technique for computing the exponential of a commutator: If $\hat{S}\vec{v} = [\mathbf{q}, \vec{v}]$, where the commutator is computed according to the multiplication rules of the quaternion algebra, then

$$\exp(\lambda\hat{S})\vec{v} = \exp(\lambda\mathbf{q}) * \vec{v} * \exp(-\lambda\mathbf{q}), \tag{7.6}$$

where now the exponential $\exp(\lambda\mathbf{x})$ is also to be computed according to the rules of the quaternion algebra. (The exponential of a quaternion $\mathbf{x}$ is defined by the Taylor series,

$$\exp\mathbf{x} = 1 + \mathbf{x} + \frac{1}{2!}\mathbf{x} * \mathbf{x} + \frac{1}{3!}\mathbf{x} * \mathbf{x} * \mathbf{x} + ...,$$

and it can be shown that this series converges for any quaternion $\mathbf{x}$, just like in the case of a matrix exponential.)

Derivation of Eq. (7.6): We introduce the λ-dependent function $\vec{v}(\lambda)$ such that $\vec{v}(0) = \vec{v}_0$ and such that the differential equation holds,

$$\frac{d}{d\lambda}\vec{v}(\lambda) = [\mathbf{q}, \vec{v}(\lambda)].$$

We can verify that Eq. (7.6) is a solution of this differential equation. First, we note that

$$\frac{d}{d\lambda}\exp(\lambda\mathbf{q}) = \mathbf{q} * \exp(\lambda\mathbf{q}) = \exp(\lambda\mathbf{q}) * \mathbf{q};$$

here it is essential that the quaternionic multiplication should be used in defining the exponential. Then we compute

$$\begin{aligned}
&\frac{d}{d\lambda}\left(\exp(\lambda\mathbf{q}) * \vec{v}_0 * \exp(-\lambda\mathbf{q})\right) \\
&\quad = \mathbf{q} * \exp(\lambda\mathbf{q}) * \vec{v}_0 * \exp(-\lambda\mathbf{q}) \\
&\quad + \exp(\lambda\mathbf{q}) * \vec{v}_0 * \exp(-\lambda\mathbf{q}) * (-\mathbf{q}) \\
&\quad = [\mathbf{q}, \exp(\lambda\mathbf{q}) * \vec{v}_0 * \exp(-\lambda\mathbf{q})] \\
&\quad = [\mathbf{q}, \vec{v}(\lambda)].
\end{aligned}$$

Since the solution of the differential equation with a given initial condition is unique, we find that the formula (7.6) is indeed the solution. ■

Remark: The derivation did not essentially use the fact that we are computing in the algebra of quaternions. Therefore, the same derivation will hold in the matrix algebra or in any other algebra. ■

Since we now need quaternionic exponentiation, let us derive an explicit formula for it.

Statement: An explicit formula for the exponential of a purely imaginary quaternion is

$$\exp(\vec{a}) = \mathbf{1}\cos|\vec{a}| + \frac{\vec{a}}{|\vec{a}|}\sin|\vec{a}|,$$

and for the exponential of an arbitrary quaternion is

$$\exp\left(a_0\mathbf{1} + \vec{a}\right) = \mathbf{1}e^{a_0}\cos|\vec{a}| + e^{a_0}\frac{\vec{a}}{|\vec{a}|}\sin|\vec{a}|.$$

It follows that every nonzero quaternion $\mathbf{x}$ is an exponential of some other quaternion $\mathbf{b} \equiv \ln\mathbf{x}$. Also, $\ln\mathbf{x}$ is purely imaginary if $|\mathbf{x}| = 1$.

Proof: If $\vec{a} = \sum_{j=1}^{3} a_j\mathbf{h}_j$ then $\vec{a} * \vec{a} = -|\vec{a}|^2\mathbf{1}$, so we find the same algebraic law as with the complex number $\mathrm{i}|\vec{a}|$. Therefore, the same derivation as for the Euler formula

$$\exp(\mathrm{i}a) = \cos a + \mathrm{i}\sin a$$

goes through. We obtain

$$\exp\vec{a} = \mathbf{1}\cos|\vec{a}| + \frac{\vec{a}}{|\vec{a}|}\sin|\vec{a}|,$$

$$\exp\left(a_0\mathbf{1} + \vec{a}\right) = e^{a_0}\exp\vec{a} = \mathbf{1}e^{a_0}\cos|\vec{a}| + e^{a_0}\frac{\vec{a}}{|\vec{a}|}\sin|\vec{a}|.$$

For an arbitrary $\mathbf{x} = x_0\mathbf{1} + \sum_{j=1}^{3} x_j\mathbf{h}_j$, we set $\mathbf{x} = \exp(a_0\mathbf{1} + \vec{a})$ and find

$$\begin{aligned}
a_0 &= \ln|\mathbf{x}|, \\
\vec{a} &= \frac{x_1\mathbf{h}_1 + x_2\mathbf{h}_2 + x_3\mathbf{h}_3}{\sqrt{x_1^2 + x_2^2 + x_3^2}}\arccos\frac{x_0}{|\mathbf{x}|}.
\end{aligned}$$

In the vector notation,

$$\ln\left(x_0 + \vec{x}\right) = \ln\sqrt{x_0^2 + |\vec{x}|^2} + \frac{\vec{x}}{|\vec{x}|}\arccos\frac{x_0}{\sqrt{x_0^2 + |\vec{x}|^2}}.$$

We note that the direction of the vector part of $\ln \mathbf{x}$ is the same as that of $\mathbf{x}$. ■

We have found that the exponential of an infinitesimal rotation is expressed by the formula

$$\vec{v} \mapsto \mathbf{c} * \vec{v} * \mathbf{c}^{-1},$$

where the quaternion $\mathbf{c}$ is given by

$$\mathbf{c} = \exp(\lambda \vec{q}), \quad \mathbf{c}^{-1} = \exp(-\lambda \vec{q}).$$

The axis of rotation is given by the vector $\vec{q}$, and the angle of rotation is 2λ.

We now consider the linear transformation in the space of quaternions,

$$\hat{R}_{\mathbf{c}} : \mathbf{x} \mapsto \mathbf{c} * \mathbf{x} * \mathbf{c}^{-1},$$

for an *arbitrary* fixed quaternion $\mathbf{c} \neq 0$. We may normalize $|\mathbf{c}| = 1$ without changing $\hat{R}_{\mathbf{c}}$, then

$$\hat{R}_{\mathbf{c}}(\mathbf{x}) = \mathbf{c} * \mathbf{x} * \bar{\mathbf{c}} \tag{7.7}$$

and so $\hat{R}_{\mathbf{c}}(\bar{\mathbf{x}}) = \overline{\hat{R}_{\mathbf{c}}(\mathbf{x})}$. Thus the three-dimensional subspace of purely imaginary quaternions is invariant under $\hat{R}_{\mathbf{c}}$.

Note that the transformation $\hat{R}_{\mathbf{c}}$ preserves the magnitude of quaternions,

$$|\hat{R}_{\mathbf{c}}(\mathbf{x})| = |\mathbf{x}|.$$

A purely imaginary quaternion $\mathbf{x} = -\bar{\mathbf{x}}$, interpreted as a 3-vector, is transformed by $\hat{R}_{\mathbf{c}}$ without changing its Euclidean length; thus $\hat{R}_{\mathbf{c}}$ acts as an *orthogonal* transformation in $\mathbb{R}^3$, i.e. $\hat{R}_{\mathbf{c}} \in O(3)$. Since every quaternion $\mathbf{c}$ such that $|\mathbf{c}| = 1$ is an exponential of some purely imaginary quaternion $\mathbf{b} = \ln \mathbf{c}$, we have

$$\hat{R}_{\mathbf{c}}(\mathbf{x}) = \exp(\mathbf{b}) * \mathbf{x} * \exp(-\mathbf{b}),$$

and thus we can smoothly deform $\hat{R}_{\mathbf{c}}$ into the identity map by decreasing $\mathbf{b}$ to zero. Since the determinant $\det \hat{R} = \pm 1$ for $\hat{R} \in O(3)$, we must have $\det \hat{R}_{\mathbf{c}} = 1$ for all $\mathbf{c}$, and therefore $\hat{R}_{\mathbf{c}} \in SO(3)$. In this way, normalized quaternions represent three-dimensional rotations, and multiplication of quaternions corresponds to composition of rotations: $\hat{R}_{\mathbf{a}} \hat{R}_{\mathbf{b}} = \hat{R}_{\mathbf{a}*\mathbf{b}}$. Since $\hat{R}_{\mathbf{c}}$ is quadratic in $\mathbf{c}$, heuristically one may say that the quaternion $\mathbf{c}$ is a "square root" of the rotation $\hat{R}_{\mathbf{c}}$.

So far, we have shown that quaternions describe rotations, but we have not yet shown that *every* rotation $\hat{R} \in SO(3)$ can be described through some quaternion $\mathbf{c}$ as the operator $\hat{R}_{\mathbf{c}}$. This also turns out to be true. The following theorem summarizes the precise relationship between quaternions and rotations.

Statement 7.1.2.1: (The Euler theorem and the Cayley parametrization of rotations.) An arbitrary proper rotation $\hat{R} \in SO(3)$ can be represented through the operator $\hat{R} = \hat{R}_{\mathbf{c}}$,

$$\hat{R}_{\mathbf{c}}\mathbf{x} \equiv \mathbf{c} * \mathbf{x} * \bar{\mathbf{c}},$$

with some quaternion $\mathbf{c}$ such that $|\mathbf{c}| = 1$. (Here $\mathbf{x} \in \mathbb{R}^3$ is interpreted as a purely imaginary quaternion.) For a given rotation $\hat{R}$, the quaternion $\mathbf{c}$ is unique up to sign; the same rotation is generated also by $(-\mathbf{c})$. The magnitude of the (pure imaginary) quaternion $\vec{u} = \ln \mathbf{c}$ corresponds to a *half* of the angle of the rotation $\hat{R}_{\mathbf{c}}$, and the direction of $\vec{v}$ is the axis of rotation.

Proof: Any quaternion $\mathbf{c} \neq 0$ generates a transformation

$$\hat{R}_{\mathbf{c}} : \mathbf{x} \mapsto \mathbf{c} * \mathbf{x} * \mathbf{c}^{-1}$$

which is a rotation because it is linear and preserves the norm of $\mathbf{x}$:

$$\mathbf{x} * \mathbf{x} = |\mathbf{x}|^2 \mathbf{1} = \mathbf{c} * \mathbf{x} * \mathbf{c}^{-1} * \mathbf{c} * \mathbf{x} * \mathbf{c}^{-1} = (\hat{R}_{\mathbf{c}}\mathbf{x}) * (\hat{R}_{\mathbf{c}}\mathbf{x}).$$

The operator $\hat{R}_{\mathbf{c}}$ remains the same when we multiply $\mathbf{c}$ by a number. Thus, it is possible to restrict to quaternions $\mathbf{c}$ such that $|\mathbf{c}| = 1$ without loss of generality. In that case, $\mathbf{c}^{-1} = \bar{\mathbf{c}}$.

Suppose that $\vec{u}$ is a unit vector. Let us define the quaternion

$$\mathbf{c} = \exp(\lambda \vec{u}) = \cos\lambda + \vec{u}\sin\lambda$$

and write the action of $\hat{R}_{\mathbf{c}}$ on a vector $\vec{x}$ explicitly as follows: Either $\vec{x}$ is parallel to $\vec{u}$, so $\vec{x} = x_1\vec{u}$, and then

$$\hat{R}_{\mathbf{c}}\vec{x} = \mathbf{c} * x_1\vec{u} * \bar{\mathbf{c}} = \vec{x}$$

because $\mathbf{c}$ commutes with $\vec{u}$; or else we can choose a right-handed orthonormal basis $\{\vec{u}, \vec{v}, \vec{w}\}$ such that

$$\vec{x} = x_1\vec{u} + x_2\vec{v},$$

that is, $\vec{x}$ is a linear combination of $\vec{u}$ and $\vec{v}$ but does not contain $\vec{w}$, while $\vec{w} = \vec{u} \times \vec{v}$. Then we compute

$$\begin{aligned}\hat{R}_{\mathbf{c}}\vec{x} &= \mathbf{c} * \vec{x} * \bar{\mathbf{c}} \\ &= x_1\vec{u} + (\cos\lambda + \vec{u}\sin\lambda) * (x_2\vec{v}) * (\cos\lambda - \vec{u}\sin\lambda) \\ &= x_1\vec{u} + (\cos 2\lambda)x_2\vec{v} + (\sin 2\lambda)x_2\vec{w}.\end{aligned}$$

It follows that $\hat{R}_{\mathbf{c}}\vec{x}$ rotates the vector $\vec{x}$ around the axis $\vec{u}$ by angle 2λ. Therefore, a rotation around an axis $\vec{u}$ (where $|\vec{u}| = 1$) by angle α is represented by the quaternion

$$\mathbf{c} = \exp\left(\frac{1}{2}\alpha\vec{u}\right).$$

We also need to show that for *any* given proper rotation $\hat{R}$, a quaternion $\mathbf{c}$ exists such that $\hat{R} = \hat{R}_{\mathbf{c}}$. This is not a trivial statement since we have to be able to analyze an arbitrary rotation $\hat{R}$. We start with the representation of the rotation $\hat{R}$ as a proper orthogonal transformation in $\mathbb{R}^3$ (that is, $\det \hat{R} = 1$ and $\hat{R}^T\hat{R} = \hat{1}$) and will now compute the quaternion $\mathbf{c}$.

First, we note that any linear operator in $\mathbb{R}^3$ has a cubic characteristic polynomial and therefore has at least one real eigenvalue. Since $\hat{R}^T\hat{R} = \hat{1}$, all eigenvalues λ of $\hat{R}$ satisfy $|\lambda|^2 = 1$. Any complex eigenvalues must come in complex conjugate pairs; therefore, there is at most one such pair, $e^{\pm i\alpha}$. If there is such a pair, the remaining eigenvalue must be 1 because $\det \hat{R} = 1$. If there are no complex eigenvalues then all eigenvalues are ± 1 and there are only two possibilities: the eigenvalues of the operator $\hat{R}$ are either $1, -1, -1$ or $1, 1, 1$. In other words, $\hat{R}$ is either equal to the identity operator $\hat{1}$ or else must have a nondegenerate eigenvalue 1 (with algebraic multiplicity 1) and two other eigenvalues of the form $e^{\pm i\alpha}$ with a real α. Therefore, in any case there exists an eigenvector $\vec{v}$ such that $\hat{R}\vec{v} = \vec{v}$. Without loss of generality, we may suppose that $|\vec{v}| = 1$.

Since $\hat{R}$ preserves the scalar product, the 2-dimensional subspace orthogonal to $\vec{v}$ is invariant under $\hat{R}$. In that subspace, $\hat{R}$ acts as a proper orthogonal transformation. In two

dimensions, all proper orthogonal transformations are rotations by some angle α. Therefore, $\hat{R}$ is a rotation around the axis $\vec{v}$ by angle α.

The quaternion representing $\hat{R}$ can now be easily found as $\mathbf{c} = \exp(\frac{1}{2}\alpha\vec{v})$.

$$\mathbf{c} = \exp\left(\frac{1}{2}\alpha\vec{v}\right).$$

Let us now show that, for a given rotation, the quaternion $\mathbf{c}$ is unique up to a sign. If two different quaternions $\mathbf{c}$ and $\mathbf{b}$ generate the same rotation, then

$$\forall \mathbf{x} \in \mathbb{R}^3 \ : \ \mathbf{c} * \mathbf{x} * \bar{\mathbf{c}} = \mathbf{b} * \mathbf{x} * \bar{\mathbf{b}}.$$

Multiply on the left by $\bar{\mathbf{c}}$ and on the right by $\mathbf{b}$:

$$\forall \mathbf{x} \in \mathbb{R}^3 \ : \ \mathbf{x} * \mathbf{a} = \mathbf{a} * \mathbf{x}, \qquad \mathbf{a} \equiv \bar{\mathbf{c}} * \mathbf{b}.$$

Therefore, the quaternion $\mathbf{a}$ commutes with all quaternions. This is possible only if $\mathbf{a}$ is a number. However,

$$|\mathbf{a}|^2 = |\mathbf{c}|^2 |\mathbf{b}|^2 = 1.$$

Thus, $\mathbf{a} = \pm\mathbf{1}$ and so $\mathbf{b} = \pm\mathbf{c}$. ■

Exercise:[4] Suppose that $\hat{R}$ is a given orthogonal transformation in $\mathbb{R}^3$. By the Euler theorem, $\hat{R}$ is a rotation around some axis by some angle. Show that the trace of $\hat{R}$ is equal to $1 + 2\cos\alpha$, where α is the angle of rotation. Show that the operator

$$\hat{S} \equiv \hat{R} - \hat{R}^T$$

is antisymmetric and, if $\hat{S} \neq 0$, can be represented as

$$\hat{S}\vec{x} = \vec{v} \times \vec{x},$$

where $\vec{v} \in \mathbb{R}^3$ is the eigenvector of $\hat{R}$ with eigenvalue 1 (i.e. the axis of rotation). In this way, the angle α and the eigenvector $\vec{v}$ can be easily read off the operator $\hat{R}$ (however, if $\hat{S} \neq 0$ then the vector $\vec{v}$ cannot be determined in this way).

Solution: By the Euler theorem, there exists an axis $\vec{v}$ and the angle α such that $\hat{R}$ is a rotation around $\vec{v}$ by angle α. The matrix representation of $\hat{R}$ in an orthonormal basis $\{\vec{v}, \vec{w}_1, \vec{w}_2\}$ is

$$\hat{R} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{pmatrix}.$$

Therefore, $\operatorname{Tr}\hat{R} = 1 + 2\cos\alpha$.

Consider now the operator $\hat{S}$ defined above. Since $\hat{S}$ is antisymmetric, there exists a vector $\vec{a} \in \mathbb{R}^3$ such that $\hat{S}\vec{x} = \vec{a} \times \vec{x}$ for all $\vec{x} \in \mathbb{R}^3$. It remains to show that $\vec{a}$ is parallel to $\vec{v}$. Indeed,

$$\hat{S}\vec{v} = (\hat{R} - \hat{R}^T)\vec{v} = (\hat{R} - \hat{R}^{-1})\vec{v} = 0.$$

Therefore, $\vec{a} \times \vec{v} = 0$. ■

It is important to note that a quaternion contains only the information about the half-angle of rotation, so a rotation by 2π may be represented by a nontrivial quaternion $\mathbf{c} = -\mathbf{1}$ as well as by $\mathbf{c} = \mathbf{1}$. Imagine that we use a quaternion to keep track of the orientation of a rigid body; then the quaternion will *change sign* after a rotation of the rigid body by 2π and return to the original value only after another rotation by 2π in the same direction. This unusual behavior is the main reason quaternions are useful in the construction of spinors.

[4] This is explained in an online supplementary note to the course Math H110 (UC Berkeley) by W. Kahan, file `Cross.pdf` dated March 2, 2008.

Since $\hat{R}_{\mathbf{c}} = \hat{R}_{-\mathbf{c}}$, every two opposite quaternions correspond to the same three-dimensional rotation. The set of all normalized quaternions $\{\mathbf{c} : |\mathbf{c}| = 1\}$ is a sphere S^3 in the four-dimensional space $\mathbb{H}$. Thus, the set $SO(3)$ is equivalent to a sphere S^3 with identified opposite points. We have thus defined a projection map $S^3 \to SO(3)$ acting as $\mathbf{c} \to \hat{R}_{\mathbf{c}}$, and this map is everywhere 2-to-1. Such maps between manifolds are called **coverings**, so in particular $S^3 \to SO(3)$ is a twofold covering.

Moreover, the sphere S^3 is represented as the set of *unitary* complex matrices with unit determinant via Eq. (7.4); the group of these matrices is denoted by $SU(2)$. Note that this group is at the same time a manifold (S^3), and also the multiplication by a fixed matrix $A \in SU(2)$ is a smooth map $S^3 \to S^3$. Groups that are smooth manifolds are called **Lie groups**.

Note that the multiplication of quaternions corresponds, on the one hand, to the multiplication of matrices, and on the other hand, to multiplication of rotations. Thus the covering $SU(2) \to SO(3)$ is a group homomorphism.

There is no canonical way to define a smooth inverse map $SO(3) \to SU(2)$ because we would need to choose the sign of a quaternion according to "how many 2π rotations were made," and this information cannot be derived from one element of $SO(3)$. If one were given not only an element $\hat{R} \in SO(3)$ but also a continuous path leading from the identity transformation $\hat{1} \in SO(3)$ to $\hat{R}$, then one would be able to choose a unique element of $SU(2)$ corresponding to $\hat{R}$. (Paths in $SO(3)$ corresponding to an odd and an even number of 2π rotations are topologically inequivalent and cannot be continuously deformed into each other.) On the other hand, we have seen that *infinitesimal* transformations can be uniquely mapped 1-to-1 between $SO(3)$ and $SU(2)$.

We have seen that a quaternion can be equivalently viewed as a two-dimensional complex vector $\alpha \in \mathbb{C}^2$ on which quaternions $\mathbf{c}$ act as unitary matrices $\hat{\mathbf{c}} \in SU(2)$. Although the vector α does not belong to the real three-dimensional space $\mathbb{R}^3$, the vector α can be transformed by a rotation $\hat{X} \in SO(3)$ performed in $\mathbb{R}^3$ because we can find a quaternion $\mathbf{c}$ such that $\hat{X} = \hat{R}_{\mathbf{c}}$and transform the vector α with the unitary matrix $\hat{\mathbf{c}}$. (Note that unitary matrices preserve the Hermitian scalar product in $\mathbb{C}^2$.) Transformed in this manner, α represents an unusual kind of a "directed magnitude" that rotates only "half-way."

Thus, the two-dimensional complex vector α is in fact a **spinor** and the auxiliary space $S \equiv \mathbb{C}^2$ is the spinor space. To summarize, spinors are quaternions viewed as elements of $\mathbb{C}^2$.

Remark: A non-proper rotation is an orthogonal transformation $\hat{R}$ such that $\det\hat{R} = -1$. Such transformations can be represented through quaternions as

$$\mathbf{x} \mapsto -\mathbf{c} * \mathbf{x} * \bar{\mathbf{c}}.$$

7.1.3 The Lorentz group

So far, we worked in the Euclidean space $\mathbb{R}^3$ and ignored the time dimension of the spacetime. We shall now extend the considerations to the 3+1-dimensional case. We have seen that proper rotations $\hat{R} \in SO(3)$ are represented by matrices acting in a two-dimensional complex space S of spinors, such that a spinor changes sign under a rotation by 2π. We shall now extend the considerations to the 3+1-dimensional case.

Our goal for now is to define spinors at a point. We shall be working in the Minkowski spacetime, which we interpret as a tangent plane to a point of a 3+1-dimensional spacetime manifold.

Lorentz transformations are linear transformations of the Minkowski spacetime that preserve the metric g.

Statement: Show that *any* transformation $\hat{L}$ preserving the metric, $g(\hat{L}\mathbf{x}, \hat{L}\mathbf{y}) = g(\mathbf{x}, \mathbf{y})$, must be a linear transformation.

Hint: Consider an image of an orthogonal basis under $\hat{L}$.

Solution: An orthogonal basis $\{\mathbf{e}_j\}$ must be transformed into another orthogonal basis $\{\hat{L}\mathbf{e}_j\}$. Then, we have

$$\begin{aligned} g(\hat{L}(\mathbf{x}+\lambda\mathbf{y}), \hat{L}\mathbf{z}) &= g(\mathbf{x}+\lambda\mathbf{y}, \mathbf{z}) \\ &= g(\hat{L}\mathbf{x}, \hat{L}\mathbf{z}) + \lambda g(\hat{L}\mathbf{y}, \hat{L}\mathbf{z}). \end{aligned}$$

Choosing $\mathbf{z} = \mathbf{e}_j$ for $j = 0...3$, we obtain the required linearity property.

All the Lorentz transformations $\hat{L}$ form a group $O(3,1)$, with a subgroup $SO(3,1)$ of orientation-preserving transformations ($\det \hat{L} = 1$). Below we shall always consider only the group $SO(3,1)$ which we shall call the **Lorentz group**. Lorentz transformations include proper rotations, $SO(3) \subset SO(3,1)$, and we already know how spinors transform under $SO(3)$. Now we need to investigate how spinors may transform under other elements the Lorentz group that are not purely rotations. These other elements are the **boosts**, i.e. a transformation between reference frames moving with respect to one another. (A proper rotation transforms between frames that are at rest with respect to one another.)[5] We may identify a proper rotation in a geometric way as a Lorentz transformation that has an invariant timelike vector $\mathbf{v}_0$ and an invariant spacelike vector $\mathbf{s}_0$, such that $g(\mathbf{v}_0, \mathbf{s}_0) = 0$. A transformation $\hat{L}$ such that $\hat{L}\mathbf{v}_0 = \mathbf{v}_0$, $\hat{L}\mathbf{s}_0 = \mathbf{s}_0$, and $g(\mathbf{v}_0, \mathbf{s}_0) = 0$, is interpreted as a spatial rotation in the rest frame of an observer with 4-velocity $\mathbf{v}_0$ around the axis $\mathbf{s}_0$. A boost cannot have invariant timelike vectors, but instead it has *two* invariant spacelike vectors $\mathbf{s}_1$ and $\mathbf{s}_2$, which are orthogonal to the direction of the boost "velocity." Let us describe boosts more explicitly.

Statement: Show by deriving an explicit formula that there exists a boost $\hat{L} \in SO(3,1)$ transforming a timelike vector $\mathbf{u}$ into another timelike vector $\mathbf{v} \neq \mathbf{u}$, and leaving invariant two spacelike vectors $\mathbf{s}_{1,2}$, orthogonal to $\mathbf{u}$.

Solution: The required Lorentz transformation $\hat{L}$ should act in the 2-plane spanned by $\{\mathbf{u}, \mathbf{v}\}$. Such a transformation will leave invariant every spacelike vector orthogonal to $\mathbf{u}$ and $\mathbf{v}$. The subspace $(\mathbf{u}, \mathbf{v})^\perp$ is two-dimensional, therefore there exist two spacelike vectors orthogonal to $\mathbf{u}$ which will be left invariant. An explicit formula for $\hat{L}$ can be derived e.g. by determining the coefficients λ_{ij} in the general expression for a linear transformation in the $\{\mathbf{u}, \mathbf{v}\}$ plane,

$$\hat{L}\mathbf{x} = \lambda_{11}\mathbf{u}g(\mathbf{x}, \mathbf{u}) + \lambda_{12}\mathbf{u}g(\mathbf{x}, \mathbf{v}) + \lambda_{21}\mathbf{v}g(\mathbf{x}, \mathbf{u}) + \lambda_{22}\mathbf{v}g(\mathbf{x}, \mathbf{v}).$$

Assuming $g(\mathbf{u}, \mathbf{u}) = g(\mathbf{v}, \mathbf{v}) = 1$, the result is

$$\hat{L}\mathbf{x} = \mathbf{v}g(\mathbf{x}, \mathbf{u}) + \frac{\gamma\mathbf{v} - \mathbf{u}}{\gamma^2 - 1} g(\mathbf{x}, \gamma\mathbf{u} - \mathbf{v}),$$

where $\gamma \equiv g(\mathbf{u}, \mathbf{v}) > 1$ is the "boost factor." Note that for $\mathbf{v} \to \mathbf{u}$, $\gamma \to 1$ the result is well-behaved.

The relationship between boosts and rotations is the following.

Statement 1: Every Lorentz transformation can be represented as a product of a boost and a proper rotation.

Proof: Suppose a Lorentz transformation $\hat{L} \in SO(3,1)$ brings a timelike vector $\mathbf{u}$ into a (timelike) vector $\mathbf{v}$. If $\mathbf{v} = \mathbf{u}$ then $\hat{L}$ is itself a proper rotation, so we can still say that $\hat{L}$ is a product of a "trivial" boost and a rotation. If $\mathbf{v} \neq \mathbf{u}$, we can find a boost $\hat{L}_1$ that brings $\mathbf{u}$ into $\mathbf{v}$. After performing the boost, the 3-plane orthogonal to $\mathbf{u}$ is brought into the 3-plane $\mathbf{v}^\perp$ orthogonal to $\mathbf{v}$. Now we look for the remaining transformation $\hat{L}_2 \in SO(3,1)$ such that $\hat{L} = \hat{L}_2\hat{L}_1$. Since $\hat{L}_2$ should leave $\mathbf{v}$ invariant, we have $\hat{L}_2(\mathbf{v}^\perp) \subset \mathbf{v}^\perp$, and thus $\hat{L}_2$ is a proper rotation. ■

This statement allows us to understand the topological structure of the group $SO(3,1)$.

Statement 2: The Lorentz group $SO(3,1)$ has the topology $SO(3) \times \mathbb{R}^3$.

Proof: If a boost changes a timelike vector $\mathbf{u}$ to $\mathbf{v}$ then the spacelike vector

$$\vec{v}_{rel} \equiv \frac{\mathbf{v} - \gamma\mathbf{u}}{\gamma},$$

such that $g(\mathbf{u}, \vec{v}_{rel}) = 0$, represents the relative velocity of $\mathbf{v}$ in the frame $\mathbf{u}$. Note that $\gamma^{-2} = 1 - |\vec{v}_{rel}|^2$. If we consider the images of $\mathbf{u}$ under every possible boost, we find that each 3-vector $\vec{v}_{rel} \in \mathbb{R}^3$ uniquely corresponds to a boost. Therefore the space of all boosts has the topology $\mathbb{R}^3$. By Statement 1, every $\hat{L} \in SO(3,1)$ can be decomposed into a product of a boost and a proper rotation. Thus, the manifold $SO(3,1)$ can be mapped one-to-one into the manifold $SO(3) \times \mathbb{R}^3$.

7.1.4 Lorentz transformations of spinors

To determine how spinors transform under a boost, we need to extend the quaternionic picture of proper rotations to boosts. Once each Lorentz transformation is mapped to a quaternion, the representation (7.4) will yield the corresponding spinor transformation.

We can arrive at a quaternionic representation of Lorentz transformations by analogy with Eq. (7.7) if we consider quaternions with *complex* coefficients. Denote by $\mathbb{H} \otimes \mathbb{C}$ the space of linear combinations $\mathbf{c} = c_0\mathbf{1} + \sum_{j=1}^3 c_j\mathbf{h}_j$, where c_j are complex coefficients. By analogy with Eq. (7.7), let us consider the quaternionic transformation[6]

$$\hat{L}_\mathbf{c} : \mathbf{x} \to \mathbf{c} * \mathbf{x} * \bar{\mathbf{c}}, \tag{7.8}$$

where $\mathbf{c}$ is a quaternion satisfying

$$|\mathbf{c}|^2 \equiv c_0^2 + c_1^2 + c_2^2 + c_3^2 = 1,$$

and the quaternion $\bar{\mathbf{c}}$ is at once the quaternionic and the complex conjugate of $\mathbf{c}$,

$$\bar{\mathbf{c}} \equiv \bar{c}_0\mathbf{1} - \sum_{j=1}^3 \bar{c}_j\mathbf{h}_j.$$

[5] In the geometric approach adopted in this text, coordinate systems are not used, so Lorentz transformations *change vectors* rather than coordinate systems. In the coordinate-based approach, such transformations are called **active**, while transformations that merely change coordinate systems are called **passive**.

[6] Note that the formula $\mathbf{x} \to \mathbf{c} * \mathbf{x} * \mathbf{c}^{-1}$ does not generate Lorentz transformations!

"Complex-valued quaternions" do not necessarily satisfy $|\mathbf{c}|^2 \geq 0$, but it is easy to check that the algebraic property $|\mathbf{x} * \mathbf{y}|^2 = |\mathbf{x}|^2 |\mathbf{y}|^2$ still holds (its derivation is independent of the assumption that c_j are real-valued). Thus we have

$$|\hat{L}_{\mathbf{c}}\mathbf{x}|^2 = |\mathbf{x}|^2$$

as before.

It is easy to see that $\overline{\mathbf{x} * \mathbf{y}} = \bar{\mathbf{y}} * \bar{\mathbf{x}}$, and so the transformation $\hat{L}_{\mathbf{c}}$ acting in the complex quaternionic space $\mathbb{H} \otimes \mathbb{C}$ preserves the subspace of "self-adjoint" or "Hermitian" quaternions $\mathbf{x}$ such that $\bar{\mathbf{x}} = \mathbf{x}$:

$$\overline{\hat{L}_{\mathbf{c}}\mathbf{x}} \equiv \overline{\mathbf{c} * \mathbf{x} * \bar{\mathbf{c}}} = \mathbf{c} * \bar{\mathbf{x}} * \bar{\mathbf{c}} = \mathbf{c} * \mathbf{x} * \bar{\mathbf{c}}.$$

Note that a "Hermitian" quaternion must have the form $\mathbf{x} = x_0\mathbf{1} + i\sum_{j=1}^{3} x_j \mathbf{h}_j$, where x_j are *real*. Hence, $\mathbf{x}$ can be viewed as a real 4-vector with components (x_0, x_1, x_2, x_3). Then we immediately find that the quantity $|\mathbf{x}|^2 = x_0^2 - x_1^2 - x_2^2 - x_3^2$ is conserved by transformations $\hat{L}_{\mathbf{c}}$. Therefore $\hat{L}_{\mathbf{c}}$ act as Lorentz transformations in the 4-dimensional space of Hermitian quaternions $\mathbf{x}$. (The same result holds for anti-Hermitian quaternions $\mathbf{x}$ such that $\bar{\mathbf{x}} = -\mathbf{x}$.)

Hence, we have found a correspondence between quaternions $\mathbf{c}$ and Lorentz transformations $\hat{L}_{\mathbf{c}}$. The following simple considerations show that the inverse map, $SO(3,1) \to \mathbb{H} \otimes \mathbb{C}$, also exists. It follows from previous calculations that every $\mathbf{c}$ such that $|\mathbf{c}|^2 = 1$ is equal to an exponential of a "purely imaginary" quaternion,

$$\mathbf{c} = \exp \mathbf{b}, \quad \mathbf{b} = b_1\mathbf{h}_1 + b_2\mathbf{h}_2 + b_3\mathbf{h}_3,$$

where now b_j are also complex-valued. As before, only orientation-preserving transformations are generated by quaternions. Transformations $\hat{L}_{\exp \mathbf{b}}$ with all real b_j correspond to proper rotations that preserve the 4-vector $\mathbf{u} \equiv (1,0,0,0)$. These rotations form the subgroup $SO(3)$. Transformations with imaginary b_j are boosts that preserve two spacelike vectors orthogonal to $\mathbf{u}$ (see Calculation below). Since every Lorentz transformation is a product of a proper rotation and a boost, and since the product of quaternions corresponds to the product of transformations, it follows that for any $\hat{L} \in SO(3,1)$ there exists a complex-valued quaternion $\mathbf{c}$ such that $\hat{L} = \hat{L}_{\mathbf{c}}$.

Calculation: Derive the explicit formula for the transformation of the coefficients x_j of the 4-vector $\mathbf{x}$ under $\mathbf{x} \to \hat{L}_{\mathbf{c}}\mathbf{x}$, where $\mathbf{c} = \exp(i\alpha\mathbf{h}_1)$, and show that this is a boost in the direction x_1.

Now let us use the matrix representation (7.4) which maps quaternions $\mathbf{c}$ into 2×2 complex matrices $\hat{\mathbf{c}}$ acting in $S \equiv \mathbb{C}^2$. The condition $|\mathbf{c}|^2 = 1$ is equivalent to $\det \hat{\mathbf{c}} = 1$, therefore the set of all the admissible quaternions is the same as the set of all 2×2 complex matrices with unit determinant. This set is denoted by $SL(2,\mathbb{C})$ and is called the **special linear group** in two complex dimensions. It is clear that we have built a group homomorphism $SL(2,\mathbb{C}) \to SO(3,1)$. As before, quaternions $\mathbf{c}$ and $-\mathbf{c}$ generate the same Lorentz transformation $\hat{L}_{\mathbf{c}}$. Thus, the map $SL(2,\mathbb{C}) \to SO(3,1)$ is a 2-to-1 covering.

It is easy to see that the quaternionic conjugation $\mathbf{c} \to \bar{\mathbf{c}}$ corresponds to the Hermitian conjugation of matrices, $\hat{\mathbf{c}} \to \hat{\mathbf{c}}^\dagger$. Therefore the matrix formula describing the transformation (7.8) is

$$\hat{L}_{\mathbf{c}}\mathbf{x} = \hat{\mathbf{c}}\hat{\mathbf{x}}\hat{\mathbf{c}}^\dagger,$$

where $\hat{\mathbf{c}} \in SL(2,\mathbb{C})$ and the matrix $\hat{\mathbf{x}}$ describing the 4-vector $\mathbf{x}$ is

$$\hat{\mathbf{x}} \equiv \begin{pmatrix} x_0 + x_3 & x_1 - ix_2 \\ x_1 + ix_2 & x_0 - x_3 \end{pmatrix}. \tag{7.9}$$

Note that this matrix is Hermitian, $\hat{\mathbf{x}}^\dagger = \hat{\mathbf{x}}$; thus the subspace of "Hermitian" quaternions is equivalent to the subspace of Hermitian 2×2 matrices. A concise expression for the matrix $\hat{\mathbf{x}}$ is

$$\hat{\mathbf{x}} = \sum_{j=0}^{3} \sigma_j x_j,$$

where σ_j, $j = 1,2,3$ are the Pauli matrices, and we have set $\sigma_0 \equiv \hat{1}$. As before, we have

$$\det \hat{\mathbf{x}} = g(\mathbf{x},\mathbf{x}) = x_0^2 - x_1^2 - x_2^2 - x_3^2.$$

In this way, the Pauli matrices σ_j provide a 1-to-1 map between the Minkowski spacetime $\mathbb{R}^4$ and the space of all 2×2 Hermitian matrices.[7] Below we shall denote this map by $\hat{\sigma}$.

Thus we have built the **spinor representation** of the Lorentz group in the spinor space $S = \mathbb{C}^2$. The procedure can be summarized like this: Every Lorentz transformation $\hat{L} \in SO(3,1)$ yields a quaternion $\mathbf{c}$ such that $\hat{L} = \hat{L}_{\mathbf{c}}$, and then we compute the corresponding matrix $\hat{\mathbf{c}}$ acting in S. The product of Lorentz transformations $\hat{L}$ corresponds to the product of matrices $\hat{\mathbf{c}}$.

Given a Lorentz transformation $\hat{L} \in SO(3,1)$, the quaternion $\mathbf{c}$ (and thus the matrix $\hat{\mathbf{c}}$) is defined only up to a sign. Hence, it is necessary to consider spinors $\alpha \in \mathbb{C}^2$ as quantities that are defined only up to a sign. (In quantum mechanics, this would be a natural conclusion if the spinor were considered as a "wave function" which is defined only up to a phase.) As before, a spinor changes sign after a rotation by 2π. However, now we know how to rotate the spinor in *any* reference frame since we derived the action of an arbitrary Lorentz transformation on a spinor.

7.2 * Rotations in higher dimensions

The construction of spinors depends on the description of rotations in $\mathbb{R}^3$ through quaternions. Quaternions work well for $\mathbb{R}^3$ and even for $\mathbb{R}^4$ if one considers complex-valued quaternions; however, quaternions cannot be straightforwardly generalized to describe higher-dimensional rotations. Instead, one can use the construction based on Clifford algebras.

Generally, an **algebra** is a vector space with a bilinear multiplication (any two vectors can be multiplied to yield another vector in the same space). So, to define an algebra, we need to specify a vector space and define the multiplication rule.

A formal definition of a Clifford algebra will be given in the next subsection. A motivation for introducing the Clifford algebra is found by recalling the construction of rotations in $\mathbb{R}^3$ through quaternions. We started with infinitesimal rotations and then exponentiated them to obtain finite rotations. An arbitrary infinitesimal rotation $\hat{S}$ in $\mathbb{R}^3$ can be represented by the cross product $\hat{S}\vec{x} = \vec{v} \times \vec{x}$, which is then replaced by a commutator $[\vec{v},\vec{x}]$ computed in the quaternion algebra. A finite rotation is found by exponentiating the commutator, which yields the formula $e^{\vec{v}} * \vec{x} * e^{-\vec{v}}$. We would like to obtain a similar description for vectors from an arbitrary vector space V where, for instance, the vector "cross" product is not defined. Thus,

[7] This map is called the **Infeld-van der Waerden map**.

our goal is to find an algebra where the commutator corresponds to infinitesimal rotations in the vector space V. An arbitrary infinitesimal rotation $\hat{S}$ is an antisymmetric tensor and can be identified with an element of $\wedge^2 V$, the space of bivectors, i.e. of linear combinations of the form $\mathbf{a}\wedge\mathbf{b}+\mathbf{c}\wedge\mathbf{d}+...$, where $\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d},...\in V$ and the "wedge" symbol $\wedge$ represents the antisymmetric (exterior) tensor product. Thus it appears that we can obtain the required algebra only if the vector space of the algebra includes not only the vector space V but also, at least, the space of bivectors. The enlargement of the vector space did not appear in the case of $\mathbb{R}^3$ because the space of bivectors over $\mathbb{R}^3$ is three-dimensional and could be identified with $\mathbb{R}^3$ itself. In higher dimensions, this identification is not possible any more.

In order to guess the new multiplication law, let us replace quaternions by bivectors in the case of $\mathbb{R}^3$. We can choose an orthonormal basis $\{\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3\}$ in $\mathbb{R}^3$ and note that the ordinary right-hand rule identifies the bivector $\mathbf{a}\wedge\mathbf{b}$ with the cross product $\mathbf{a}\times\mathbf{b}$; for instance, the bivector $\mathbf{e}_1\wedge\mathbf{e}_2$ is identified with the vector $\mathbf{e}_3$. At this point, we would like to avoid performing this identification. So we are motivated to replace the quaternionic units $\mathbf{h}_1$, $\mathbf{h}_2$, $\mathbf{h}_3$ by bivectors $\mathbf{e}_2\wedge\mathbf{e}_3$, $\mathbf{e}_3\wedge\mathbf{e}_1$, $\mathbf{e}_1\wedge\mathbf{e}_2$, and also to replace cross products by bivectors where appropriate. In order to make contact with the standard definition of the Clifford algebra, we will choose the signs as follows:

$$\mathbf{h}_1\mapsto-\mathbf{e}_2\wedge\mathbf{e}_3,\quad \mathbf{h}_2\mapsto-\mathbf{e}_3\wedge\mathbf{e}_1,\mathbf{h}_3\mapsto-\mathbf{e}_1\wedge\mathbf{e}_2.$$

The rules of quaternionic multiplication can now be written as rules for multiplication of vectors and bivectors:

- We distinguish vectors $\vec{a}\in\mathbb{R}^3$ from quaternionic units such as $\mathbf{h}_1\equiv\mathbf{e}_3\wedge\mathbf{e}_2\in\wedge^2\mathbb{R}^3$.
- Multiplication of two quaternions: Instead of $\mathbf{h}_3*\mathbf{h}_3=-\mathbf{1}$ we write $(\mathbf{e}_2\wedge\mathbf{e}_1)*(\mathbf{e}_2\wedge\mathbf{e}_1)=-1$, and similarly for other quaternionic units.
- Multiplication of a vector and a quaternion: Instead of

  $$\mathbf{h}_1*\vec{e}_2=-\vec{e}_2*\mathbf{h}_1=\vec{e}_3$$

  we write

  $$(\mathbf{e}_3\wedge\mathbf{e}_2)*\mathbf{e}_2=-\mathbf{e}_2*(\mathbf{e}_3\wedge\mathbf{e}_2)=\mathbf{e}_3.$$

These rules of multiplication can be summarized as follows: if vectors $\mathbf{x}$ and $\mathbf{y}$ are unit vectors such that $g(\mathbf{x},\mathbf{y})=0$ then

$$\mathbf{x}*\mathbf{y}=\mathbf{x}\wedge\mathbf{y},\quad \mathbf{x}*\mathbf{x}=\mathbf{1},$$
$$\mathbf{x}*(\mathbf{x}\wedge\mathbf{y})=\mathbf{y},\quad (\mathbf{x}\wedge\mathbf{y})*\mathbf{y}=\mathbf{x}.$$

We find that the algebra of quaternions is equivalent to a larger algebra containing numbers, vectors, and bivectors. We have not yet completely defined the multiplication in the larger algebra; for instance, we have not yet defined $(\mathbf{e}_1\wedge\mathbf{e}_2)*\mathbf{e}_3$. According to the quaternion algebra, this is $-\mathbf{h}_3*\vec{e}_3=\mathbf{1}$, so it is natural to define

$$(\mathbf{e}_3\wedge\mathbf{e}_2)*\mathbf{e}_1=\mathbf{e}_3\wedge\mathbf{e}_2\wedge\mathbf{e}_1=-\mathbf{e}_1\wedge\mathbf{e}_2\wedge\mathbf{e}_3.$$

Note that this is not a number but an antisymmetric tensor of rank 3. Since the space of antisymmetric 3-tensors over $\mathbb{R}^3$ is one-dimensional, we again find a natural correspondence with the quaternion algebra.

Since we are interested in representing rotations through an algebra, the rule for multiplication of a vector and a bivector must be such that all infinitesimal rotations are represented by commutators of vectors with some bivectors. For instance, an infinitesimal rotation around the axis $\mathbf{e}_3$ by an (infinitesimal) angle λ was previously represented by the commutator

$$\vec{x}\mapsto\vec{x}+\lambda\vec{e}_3\times\vec{x}=\vec{x}+\frac{\lambda}{2}\left(\mathbf{h}_3*\vec{x}-\vec{x}*\mathbf{h}_3\right).$$

We can easily check that this holds with the above multiplication rules if we use the bivector $\mathbf{e}_2\wedge\mathbf{e}_1$ instead of the quaternion $\mathbf{h}_3$.

The multiplication operator $*$ applied to orthonormal unit vectors can be summarized as follows. When all vectors in a product term are different, the term yields the exterior multiplication, e.g. $\mathbf{x}*\mathbf{y}=\mathbf{x}\wedge\mathbf{y}$ and $\mathbf{x}*\mathbf{y}*\mathbf{z}=\mathbf{x}\wedge\mathbf{y}\wedge\mathbf{z}$. When not all vectors are different, any equal unit vectors can be brought together using antisymmetry, e.g.

$$...*\mathbf{a}*\mathbf{x}*\mathbf{b}*\mathbf{x}*...=-...*\mathbf{a}*\mathbf{b}*\mathbf{x}*\mathbf{x}*...$$

and then the product $\mathbf{x}*\mathbf{x}\equiv1$ can be omitted, e.g.

$$...*\mathbf{a}*\mathbf{x}*\mathbf{x}*\mathbf{b}*...=...*\mathbf{a}*\mathbf{b}*...$$

This multiplication rule is associative but not commutative.

7.2.1 Clifford algebra

This construction can be now generalized from $\mathbb{R}^3$ to an arbitrary N-dimensional vector space V. The **Clifford algebra** $\mathrm{Cl}(V)$ based on a vector space V is defined as follows.

The vector space of the Clifford algebra $\mathrm{Cl}(V)$ is the same as the vector space of the exterior algebra of V, that is, $\mathrm{Cl}(V)$ consists of a direct sum of the spaces of all the possible antisymmetric tensors built out of vectors from V. Thus, an element of $\mathrm{Cl}(V)$ could be of the form

$$\alpha_0\mathbf{1}+\mathbf{x}_1+\mathbf{x}_2\wedge\mathbf{x}_3+\mathbf{x}_4\wedge\mathbf{x}_5\wedge\mathbf{x}_6+...,$$

where α_0 is a number. The above is just an example; there could be of course more terms of each tensorial rank. (The symbol $\mathbf{1}$ denotes the scalar unit.)

In the exterior algebra, the multiplication is denoted by the "wedge" symbol $\wedge$. In the Clifford algebra, a different multiplication rule is defined, which we will denote by the symbol $*$ like the quaternion multiplication. The Clifford multiplication can be defined most concisely as follows.[8] Scalars are multiplied with other terms in the natural way: for example,

$$\sqrt{2}\mathbf{1}*(\mathbf{x}\wedge\mathbf{y})=\sqrt{2}\mathbf{x}\wedge\mathbf{y}.$$

Scalars commute with all other terms. In order to define the Clifford multiplication of arbitrary tensor terms, it is sufficient to define the multiplication of the form, say,

$$(\mathbf{e}_1\wedge\mathbf{e}_2)*(\mathbf{e}_3\wedge\mathbf{e}_4\wedge\mathbf{e}_5),$$

where $\mathbf{e}_j$ are vectors from an *orthonormal* basis in V.

The multiplication of basis tensors is defined as follows. Suppose $\{\mathbf{e}_1,...,\mathbf{e}_N\}$ is an orthonormal basis and $\psi_1=\mathbf{e}_{i_1}\wedge...\wedge\mathbf{e}_{i_m}$ and $\psi_2=\mathbf{e}_{j_1}\wedge...\wedge\mathbf{e}_{j_n}$ are two antisymmetric tensors built out of the vectors of this basis. We would like to define

[8] This definition follows Rashevskij P.K. Teorija spinorov, URSS, 2006 (in Russian).

the Clifford product $\psi_1 * \psi_2$. We first compute the exterior product $\psi_1 \wedge \psi_2$, and we must find that either $\psi_1 \wedge \psi_2 \neq 0$ or $\psi_1 \wedge \psi_2 = 0$. In the first case, we define

$$\psi_1 * \psi_2 \equiv \psi_1 \wedge \psi_2 \quad \text{if } \psi_1 \wedge \psi_2 \neq 0.$$

Therefore, the exterior multiplication can be replaced by Clifford multiplication as long as all vectors are orthogonal to each other,

$$\psi_1 = \mathbf{e}_{i_1} \wedge ... \wedge \mathbf{e}_{i_m} = \mathbf{e}_{i_1} * ... * \mathbf{e}_{i_m},$$

and similarly for ψ_2. It also follows that

$$\mathbf{a} * \mathbf{b} = -\mathbf{b} * \mathbf{a} \quad \text{if } g(\mathbf{a}, \mathbf{b}) = 0.$$

In the second case ($\psi_1 \wedge \psi_2 = 0$), the resulting product $\psi_1 * \psi_2$ will be a tensor of a lower rank.[9] There must exist at least one basis vector $\mathbf{e}_s$ that enters both ψ_1 and ψ_2. Now there are two further cases: either $\psi_1 = \psi_2 = \mathbf{e}_s$, or ψ_1 and ψ_2 have a higher rank. In the first case, we define

$$\mathbf{e}_s * \mathbf{e}_s \equiv \mathbf{1}.$$

In the second case, we use the antisymmetry of the exterior product to reorder the vectors in ψ_1 and ψ_2 such that these tensors have the form

$$\psi_1 = \pm \mathbf{e}_{i_1} \wedge ... \wedge \mathbf{e}_{i_{m-1}} \wedge \mathbf{e}_s, \quad \psi_2 = \pm \mathbf{e}_s \wedge \mathbf{e}_{j_2} \wedge ... \wedge \mathbf{e}_{j_n}.$$

Then we define (with the appropriate sign)

$$\psi_1 * \psi_2 = \pm \left(\mathbf{e}_{i_1} \wedge ... \wedge \mathbf{e}_{i_{m-1}} \right) * \left(\mathbf{e}_{j_2} \wedge ... \wedge \mathbf{e}_{j_n} \right).$$

Now we have a Clifford product of two tensors of a lower rank; we use again the same rule until all Clifford products are simplified. Since the multiplication is defined by moving vectors together towards the symbol $*$, the rule is automatically associative.

Once the multiplication of basis tensors is defined, we use linearity to define the Clifford multiplication of arbitrary terms.

Exercise 1: If $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are any vectors from V, show that

$$\mathbf{a} * \mathbf{b} = g(\mathbf{a}, \mathbf{b})\mathbf{1} + \mathbf{a} \wedge \mathbf{b}.$$

In particular,

$$\mathbf{x} * \mathbf{x} = g(\mathbf{x}, \mathbf{x})\mathbf{1}.$$

Solution: We may choose a basis $\{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_N\}$ such that (for simplicity)

$$\mathbf{a} = a\mathbf{e}_1, \quad \mathbf{b} = b_1\mathbf{e}_1 + b_2\mathbf{e}_2.$$

Then

$$\begin{aligned} \mathbf{a} * \mathbf{b} &= ab_1 \mathbf{e}_1 * \mathbf{e}_1 + ab_2 \mathbf{e}_1 * \mathbf{e}_2 \\ &= ab_1 \mathbf{1} + ab_2 \mathbf{e}_1 \wedge \mathbf{e}_2 \\ &= g(\mathbf{a}, \mathbf{b})\mathbf{1} + \mathbf{a} \wedge \mathbf{b}. \end{aligned}$$

■

[9] The two cases of the Clifford product rule unite the "exterior" and the "interior" multiplication as first defined by H. Grassmann.

Exercise 2: If $\mathbf{a}$ and $\mathbf{b}$ are vectors orthogonal to each other and $\mathbf{c}$ is any vector, show that

$$\begin{aligned} [\mathbf{a} * \mathbf{b}, \mathbf{c}] &= (\mathbf{a} \wedge \mathbf{b}) * \mathbf{c} - \mathbf{c} * (\mathbf{a} \wedge \mathbf{b}) \\ &= 2\mathbf{a}g(\mathbf{b}, \mathbf{c}) - 2\mathbf{b}g(\mathbf{a}, \mathbf{c}). \end{aligned}$$

Note that the final formula resembles the double cross product in $\mathbb{R}^3$,

$$(\vec{a} \times \vec{b}) \times \vec{c} = \vec{b}g(\vec{a}, \vec{c}) - \vec{a}g(\vec{b}, \vec{c}),$$

up to a constant factor.

Hint: Choose an orthonormal basis $\{\mathbf{e}_j\}$ such that $\mathbf{a}$ is parallel to $\mathbf{e}_1$, $\mathbf{b}$ is parallel to $\mathbf{e}_2$, and $\mathbf{c}$ is a linear combination of $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$. ■

Remark: The Clifford multiplication rule can be summarized more concisely (but less transparently) by the formula

$$\mathbf{a} * \mathbf{b} + \mathbf{b} * \mathbf{a} = 2g(\mathbf{a}, \mathbf{b})\mathbf{1}.$$

Starting with this formula, one can (after some work) derive the more explicit rules that we have presented earlier. We will not use this formula directly. ■

7.2.2 Representing rotations

Let us now use the Clifford algebra $\mathrm{Cl}(V)$ in the manner quite analogous to the use of quaternions, for representing rotations in an arbitrary vector space V with a given metric g.

As before, an arbitrary infinitesimal rotation $\hat{S}$ is an antisymmetric operator. The space of antisymmetric operators is isomorphic to the space of bivectors. Explicitly, a bivector such as $\mathbf{a} \wedge \mathbf{b}$ is mapped to the antisymmetric operator $\hat{S}_{\mathbf{a} \wedge \mathbf{b}}$ that acts on vectors $\mathbf{x}$ as

$$\hat{S}_{\mathbf{a} \wedge \mathbf{b}}(\mathbf{x}) \equiv \mathbf{a}g(\mathbf{b}, \mathbf{x}) - \mathbf{b}g(\mathbf{a}, \mathbf{x}).$$

As shown in Exercise 2, this is equivalent to the Clifford commutator

$$\hat{S}_{\mathbf{a} \wedge \mathbf{b}}(\mathbf{x}) = \frac{1}{2}[\mathbf{a} \wedge \mathbf{b}, \mathbf{x}].$$

A general antisymmetric operator $\hat{S}$ may require more than one exterior product term, e.g. for a certain given operator $\hat{S}$ we may need more vectors $\mathbf{a}_1$, $\mathbf{b}_1$, $\mathbf{a}_2$, $\mathbf{b}_2$, etc., so that

$$\begin{aligned} \hat{S}(\mathbf{x}) = \hat{S}_{\boldsymbol{\psi}}(\mathbf{x}) &\equiv [\boldsymbol{\psi}, \mathbf{x}], \\ \boldsymbol{\psi} &\equiv \frac{1}{2}\left(\mathbf{a}_1 \wedge \mathbf{b}_1 + \mathbf{a}_2 \wedge \mathbf{b}_2 + ... + \mathbf{a}_n \wedge \mathbf{b}_n \right). \end{aligned}$$

In $\mathbb{R}^3$, this representation is always possible with just one term. One can show[10] that no more than $N/2$ terms suffice in $\mathbb{R}^N$, and that all the vectors $\mathbf{a}_i$, $\mathbf{b}_i$ can be chosen mutually orthogonal.

A finite rotation $\hat{R}_{\boldsymbol{\psi}}$ is now obtained by exponentiating an infinitesimal rotation $\hat{S}_{\boldsymbol{\psi}}$:

$$\hat{R}_{\boldsymbol{\psi}}(\mathbf{x}) \equiv \exp(\boldsymbol{\psi}) * \mathbf{x} * \exp(-\boldsymbol{\psi}).$$

As in the case of quaternions, one can easily show that $\hat{R}_{\boldsymbol{\psi}}$ is an orthogonal transformation in V. Indeed, since

$$e^{\boldsymbol{\psi}} * e^{-\boldsymbol{\psi}} = \mathbf{1},$$

[10] Statement 2 and Exercise 7 in Section 5.7 of S. Winitzki, *Linear Algebra via Exterior Products*, lulu.com, 2010.

we have

$$\begin{aligned} g(\hat{R}_{\boldsymbol{\psi}}\mathbf{x}, \hat{R}_{\boldsymbol{\psi}}\mathbf{x})\mathbf{1} &= (\hat{R}_{\boldsymbol{\psi}}\mathbf{x}) * (\hat{R}_{\boldsymbol{\psi}}\mathbf{x}) \\ &= e^{\boldsymbol{\psi}} * \mathbf{x} * e^{-\boldsymbol{\psi}} * e^{\boldsymbol{\psi}} * \mathbf{x} * e^{-\boldsymbol{\psi}} \\ &= e^{\boldsymbol{\psi}} * \mathbf{x} * \mathbf{x} * e^{-\boldsymbol{\psi}} \\ &= e^{\boldsymbol{\psi}} * g(\mathbf{x}, \mathbf{x})\mathbf{1} * e^{-\boldsymbol{\psi}} \\ &= g(\mathbf{x}, \mathbf{x})\mathbf{1}. \end{aligned}$$

Given an explicit formula for $\boldsymbol{\psi}$, one can compute $e^{\boldsymbol{\psi}}$ as follows. Assume, without loss of generality, that

$$\boldsymbol{\psi} = \sum_{i=1}^{n} \lambda_i \mathbf{a}_i \wedge \mathbf{b}_i,$$

where all the vectors $\{\mathbf{a}_i, \mathbf{b}_i\}$ are mutually orthogonal unit vectors, while λ_i are some numbers. We first note that a term $\mathbf{a}_i \wedge \mathbf{b}_i$ commutes with any other term $\mathbf{a}_j \wedge \mathbf{b}_j$ in the Clifford algebra: If $i \neq j$ then the Clifford product coincides with the exterior product, so

$$\begin{aligned} (\mathbf{a}_i \wedge \mathbf{b}_i) * (\mathbf{a}_j \wedge \mathbf{b}_j) &= \mathbf{a}_i \wedge \mathbf{b}_i \wedge \mathbf{a}_j \wedge \mathbf{b}_j \\ &= \mathbf{a}_j \wedge \mathbf{b}_j \wedge \mathbf{a}_i \wedge \mathbf{b}_i = (\mathbf{a}_j \wedge \mathbf{b}_j) * (\mathbf{a}_i \wedge \mathbf{b}_i). \end{aligned}$$

Therefore the exponential $e^{\boldsymbol{\psi}}$ can be factorized as

$$\exp \boldsymbol{\psi} = e^{\lambda_1 \mathbf{a}_1 \wedge \mathbf{b}_1} * ... * e^{\lambda_n \mathbf{a}_n \wedge \mathbf{b}_n}.$$

Each single-term exponential is computed just as in the case of quaternions: since

$$(\mathbf{a} \wedge \mathbf{b}) * (\mathbf{a} \wedge \mathbf{b}) = -(\mathbf{b} \wedge \mathbf{a}) * (\mathbf{a} \wedge \mathbf{b}) = -\mathbf{1},$$

we have

$$\exp(\lambda \mathbf{a} \wedge \mathbf{b}) = \mathbf{1} \cos\lambda + \mathbf{a} \wedge \mathbf{b} \sin\lambda.$$

Therefore

$$\begin{aligned} \exp \boldsymbol{\psi} = (\mathbf{1} \cos\lambda_1 + \mathbf{a}_1 \wedge \mathbf{b}_1 \sin\lambda_1) * ... \\ * (\mathbf{1} \cos\lambda_n + \mathbf{a}_n \wedge \mathbf{b}_n \sin\lambda_n). \end{aligned}$$

Expanding the brackets, we would obtain an expression of the form

$$\begin{aligned} \exp \boldsymbol{\psi} = \cos\lambda_1 ... \cos\lambda_n \left(\mathbf{1} + \sum_{i=1}^{n} \mathbf{a}_i \wedge \mathbf{b}_i \tan\lambda_i + ... \right) \\ + \mathbf{a}_1 \wedge \mathbf{b}_1 \wedge ... \wedge \mathbf{a}_n \wedge \mathbf{b}_n \sin\lambda_1 ... \sin\lambda_n. \end{aligned}$$

We can consider all elements of the Clifford algebra that are generated as exponentials of some bivectors $\boldsymbol{\psi}$, or Clifford products of such exponentials. It is clear that a Clifford product $e^{\boldsymbol{\psi}_1} * e^{\boldsymbol{\psi}_2}$ generates the composition of two consecutive rotations $\hat{R}_{\boldsymbol{\psi}_1}$ and $\hat{R}_{\boldsymbol{\psi}_2}$, even though generally

$$e^{\boldsymbol{\psi}_1} * e^{\boldsymbol{\psi}_2} \neq e^{\boldsymbol{\psi}_1 + \boldsymbol{\psi}_2}.$$

The set of all such elements (allowing products of arbitrary many single exponentials) is called the **Clifford group**.

We have shown that every element of the Clifford group generates a rotation in V. It remains to show that *every* proper rotation in V can be represented by a certain element of the Clifford group.

Statement: For any proper orthogonal transformation $\hat{R}$ of the space $V = \mathbb{R}^N$, there exists a set of $n \leq N-1$ bivectors $\boldsymbol{\psi}_i \equiv \mathbf{a}_i \wedge \mathbf{b}_i$ (where $i = 1, ..., n$) from the Clifford algebra $\mathrm{Cl}(V)$, such that

$$\begin{aligned} \hat{R}\mathbf{x} &= \hat{R}_{\boldsymbol{\psi}_1} \circ ... \circ \hat{R}_{\boldsymbol{\psi}_n} \mathbf{x} \\ &= e^{\boldsymbol{\psi}_1} * ... * e^{\boldsymbol{\psi}_n} * \mathbf{x} * e^{-\boldsymbol{\psi}_n} * ... * e^{-\boldsymbol{\psi}_1} \end{aligned}$$

for all $\mathbf{x} \in V$.

Proof: We will show this by induction in the number of dimensions. In $\mathbb{R}^1$, the statement is true because every proper rotation is an identity transformation. Now we suppose that the statement is proved in $\mathbb{R}^{N-1}$, and we will prove it for $\mathbb{R}^N$. An arbitrary rotation $\hat{R}$ in $\mathbb{R}^N$ brings an orthonormal basis $\{\mathbf{e}_1, ..., \mathbf{e}_N\}$ into another orthonormal basis $\{\mathbf{f}_1, ..., \mathbf{f}_N\}$, where $\mathbf{f}_i \equiv \hat{R}\mathbf{e}_i$. We can now compute a bivector $\boldsymbol{\psi}_1$ such that the rotation $\hat{R}_{\boldsymbol{\psi}_1}$ maps $\mathbf{e}_1$ to $\mathbf{f}_1$. Indeed, we have either $\mathbf{e}_1 = \mathbf{f}_1$ or $\mathbf{e}_1 \neq \mathbf{f}_1$. If $\mathbf{e}_1 = \mathbf{f}_1$, we choose $\boldsymbol{\psi}_1 = 0$. If $\mathbf{e}_1 \neq \mathbf{f}_1$, we choose

$$\boldsymbol{\psi}_1 = \lambda \mathbf{f}_1 \wedge \mathbf{e}_1$$

with an appropriate value of the number λ. To compute λ, we denote $g(\mathbf{f}_1, \mathbf{e}_1) \equiv \cos\phi$ and write

$$\mathbf{f}_1 \wedge \mathbf{e}_1 = (\mathbf{f}_1 - \mathbf{e}_1 g(\mathbf{f}_1, \mathbf{e}_1)) \wedge \mathbf{e}_1 = (\mathbf{f}_1 - \mathbf{e}_1 \cos\phi) * \mathbf{e}_1;$$

$$(\mathbf{f}_1 \wedge \mathbf{e}_1) * (\mathbf{f}_1 \wedge \mathbf{e}_1) = -|\mathbf{f}_1 - \mathbf{e}_1 \cos\phi|^2 = -\sin^2\phi;$$

$$\mathbf{f}_1 \wedge \mathbf{e}_1 = \mathbf{b}_1 \wedge \mathbf{e}_1 \sin\phi, \quad \mathbf{b}_1 \equiv \frac{1}{\sin\phi}(\mathbf{f}_1 - \mathbf{e}_1 \cos\phi);$$

$$\exp(\boldsymbol{\psi}_1) = \mathbf{1} \cos(\lambda \sin\phi) + \mathbf{b}_1 \wedge \mathbf{e}_1 \sin(\lambda \sin\phi),$$

and then obtain

$$\begin{aligned} \hat{R}_{\boldsymbol{\psi}} \mathbf{e}_1 &= \exp(\boldsymbol{\psi}_1) * \mathbf{e}_1 * \exp(-\boldsymbol{\psi}_1) \\ &= (\mathbf{1} \cos(\lambda \sin\phi) + \mathbf{b}_1 \wedge \mathbf{e}_1 \sin(\lambda \sin\phi)) * \mathbf{e}_1 \\ &\quad * (\mathbf{1} \cos(\lambda \sin\phi) - \mathbf{b}_1 \wedge \mathbf{e}_1 \sin(\lambda \sin\phi)) \\ &= \mathbf{e}_1 \cos(2\lambda \sin\phi) + \mathbf{b}_1 \sin(2\lambda \sin\phi). \end{aligned}$$

This expression will be equal to $\mathbf{f}_1$ if

$$\cos(2\lambda \sin\phi) = g(\mathbf{e}_1, \mathbf{f}_1) = \cos\phi.$$

This determines the required value of λ as

$$\lambda = \frac{1}{2} \frac{\phi}{\sin\phi}.$$

Therefore, the required rotation $\hat{R}_{\boldsymbol{\psi}_1}$ exists. The rotation $\hat{R}_{\boldsymbol{\psi}_1}$ brings $\mathbf{e}_1$ to $\mathbf{f}_1$ and all other basis vectors $\mathbf{e}_2$, ..., $\mathbf{e}_N$ to some vectors that all belong to the orthogonal complement of $\mathbf{f}_1$. The original rotation $\hat{R}$ can be then represented as a composition of $\hat{R}_{\boldsymbol{\psi}_1}$ and an auxiliary rotation $\hat{R}'$ that does not change $\mathbf{f}_1$ but operates only within the $(N-1)$-dimensional subspace orthogonal to $\mathbf{f}_1$:

$$\hat{R} = \hat{R}' \circ \hat{R}_{\boldsymbol{\psi}_1}.$$

By the inductive hypothesis, the rotation $\hat{R}'$ can be represented as a product of not more than $(N-2)$ rotations of the form $\hat{R}_{\boldsymbol{\psi}_2}$, ..., $\hat{R}_{\boldsymbol{\psi}_n}$, where now $n \leq N-1$. Therefore, it is possible to express $\hat{R}$ by

$$\hat{R} = \hat{R}_{\boldsymbol{\psi}_2} \circ ... \circ \hat{R}_{\boldsymbol{\psi}_n} \circ \hat{R}_{\boldsymbol{\psi}_1}$$

as required. ■

We note that an arbitrary rotation in $\mathbb{R}^N$ is characterized by up to $N/2$ "axes" of rotation and the corresponding $N/2$

angles λ_i, where each "axis of rotation" is a bivector of the form $\mathbf{a} \wedge \mathbf{b}$ with unit vectors $\mathbf{a}$, $\mathbf{b}$ and can be visualized as a rotation in the plane spanned by $\mathbf{a}$, $\mathbf{b}$ by angle $2\lambda_i$.

If we apply this description to Lorentz transformations in $\mathbb{R}^4$, we find that an arbitrary Lorentz transformation can be characterized by two bivectors, say $\mathbf{a} \wedge \mathbf{b}$ and $\mathbf{c} \wedge \mathbf{d}$, where $\{\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}\}$ is an orthonormal basis, and the corresponding two "angles" α and β. Without loss of generality, we may suppose that $\mathbf{a}$ is time-like while $\mathbf{b}, \mathbf{c}, \mathbf{d}$ are spacelike. In this way, we can describe an *arbitrary* Lorentz transformation as a boost in the spacelike direction $\mathbf{b}$ with boost parameter α combined with a spatial rotation around the *same* axis $\mathbf{b}$ by angle β. Once the vectors $\mathbf{a}, \mathbf{b}$ are fixed, the bivector $\mathbf{c} \wedge \mathbf{d}$ can be determined as $*(\mathbf{a} \wedge \mathbf{b})$. Thus, a Lorentz transformation is fully characterized by two numbers, a time-like unit vector $\mathbf{a}$, and a spacelike unit vector $\mathbf{b}$. (The choice of unit vectors $\mathbf{a}, \mathbf{b}$ is arbitrary as long as the bivector $\mathbf{a} \wedge \mathbf{b}$ remains constant.)

7.3 Spinor algebra

The role of quaternions in the construction of spinors is illustrative but purely auxiliary. Quaternions provide a visual derivation of the concept of spinor spaces adapted to 3+1-dimensional spacetimes, as well as a convenient tools for computations with rotations and Lorentz transformations.[11] From now on, we may forget quaternions and simply assume that spinors are elements of a complex vector space $S = \mathbb{C}^2$ in which two structures are defined:

1. A representation "up to a sign" of the Lorentz group $SO(3,1)$ which "rotates by half-angle" (the spinor representation).

2. A fixed 1-to-1 correspondence $\hat{\sigma}$ between 2×2 Hermitian matrices and vectors from the Minkowski spacetime $\mathbb{R}^4$.

Using these properties as building blocks, we shall now study the tensor algebra generated by the space S and its relation to the usual tensor algebra in the Minkowski spacetime. Since quaternions will not be used any more, we shall denote the spinorial form of Lorentz transformations simply by $\hat{L}_s$ rather than by $\hat{\mathbf{c}}$. Also, we shall not distinguish between $SO(3,1)$ and $SL(2,\mathbb{C})$ since this distinction will play no role in our calculations.

In a real vector space V, tensors of rank (p, q) are defined as tensor products of p copies of V and q copies of the dual space V^*. To build tensors from the space S, we need to consider the dual space to S. We define the dual space S^* as the space of linear functions on S, i.e. functions $f : S \to \mathbb{C}$ such that

$$f(\mathbf{b} + \lambda \mathbf{c}) = f(\mathbf{b}) + \lambda f(\mathbf{c}), \quad \mathbf{b}, \mathbf{c} \in S,\ \lambda \in \mathbb{C}.$$

Additionally, we can define the *complex conjugate* dual space $\bar{S}^*$ as the space of **anti-linear functions** on S, i.e. functions $f : S \to \mathbb{C}$ such that

$$f(\mathbf{b} + \lambda \mathbf{c}) = f(\mathbf{b}) + \bar{\lambda} f(\mathbf{c}), \quad \mathbf{b}, \mathbf{c} \in S,\ \lambda \in \mathbb{C}.$$

Finally, we define the space dual to $\bar{S}^*$ as the space $\bar{S}$ of "complex conjugate" spinors.

All the four spaces S, $\bar{S}$, S^*, $\bar{S}^*$ are two-dimensional, and a basis $\{\mathbf{a}, \mathbf{b}\}$ in S naturally generates a dual basis $\{\mathbf{a}^*, \mathbf{b}^*\}$ in S^*, a conjugate dual basis $\{\bar{\mathbf{a}}^*, \bar{\mathbf{b}}^*\}$ in $\bar{S}^*$, and a conjugate basis $\{\bar{\mathbf{a}}, \bar{\mathbf{b}}\}$ in $\bar{S}$. There is a natural antilinear map from S^* to $\bar{S}^*$: a linear function $f(\mathbf{x})$ generates an anti-linear function whose values are $\bar{f}(\mathbf{x})$. The corresponding map $S \to \bar{S}$ is also denoted by the overbar, so for instance

$$\overline{\mathbf{a} + \lambda \mathbf{b}} = \bar{\mathbf{a}} + \bar{\lambda} \bar{\mathbf{b}} \in \bar{S} \quad \text{if } \mathbf{a}, \mathbf{b} \in S,\ \lambda \in \mathbb{C}.$$

Lorentz transformations act on dual spinors by inverse matrices $\hat{L}_s^{-1}$, and on conjugate spinors by Hermitian conjugate matrices $\hat{L}_s^\dagger$.

Spinorial tensors can be built by taking tensor products of the spinor space S and its three conjugates. The rank of a spinorial tensor contains four numbers, for instance, $(1\bar{1}, 2\bar{2})$, showing how many copies of each of S, $\bar{S}$, S^*, $\bar{S}^*$ participate in the tensor product. For instance, Hermitian bilinear forms are $(0, 1\bar{1})$-tensors on S. Complex conjugation is then a naturally defined map from $(j\bar{k}, l\bar{m})$-tensors to $(k\bar{j}, m\bar{l})$-tensors.

Since we have many different types of tensors, it is inevitable that we must eventually use the (abstract) index notation. It is traditional to denote indices in spinor spaces by capital letters, and indices in conjugate spaces by a prime, thus $T_{AB'}$ denotes a rank $(0, 1\bar{1})$ spinorial tensor. Note that the order of primed and unprimed indices is insignificant: $T_{AB'}$ and $T_{B'A}$ is the same object.

7.3.1 The fundamental 2-form

Since the spinor space S is two-dimensional, every 2-form on S is proportional to an arbitrarily selected one. There is in fact a naturally selected 2-form, which we shall call the **fundamental 2-form** and denote by ε. We shall now define the 2-form ε by using the map $\hat{\sigma}$ and the Minkowski metric g.

The map $\hat{\sigma}$ transforms "Hermitian matrices" into 4-vectors, and we need to be more specific about the space of such "matrices." Suppose h is a Hermitian matrix and $\hat{\sigma} h \in \mathbb{R}^4$ is the corresponding 4-vector. Since the formula for the Lorentz transformation is

$$\hat{L}\,(\hat{\sigma} h) = \hat{\sigma}\left(\hat{L}_s h \hat{L}_s^\dagger\right),$$

it is clear that h is a spinorial tensor of rank $(1\bar{1}, 0)$, i.e. $h \in S \otimes \bar{S}$. In the index notation, we may write the application of the map $\hat{\sigma}$ to such a tensor as

$$(\hat{\sigma} h)^\mu = \sigma^\mu_{AA'} h^{AA'}.$$

Moreover, the property of being Hermitian means that $\bar{h} = h$. Hence, the Minkowski metric g defines a bilinear form in the space $S \otimes \bar{S}$. We shall temporarily denote this bilinear form by angular brackets,

$$\langle h_1, h_2 \rangle \equiv g\,(\hat{\sigma} h_1, \hat{\sigma} h_2)\,.$$

On the other hand, we have (*with some choice of basis* in S) the explicit formula (7.9), from which it follows that the determinant of a Hermitian matrix h is equal to $g(\hat{\sigma} h, \hat{\sigma} h)$. The determinant of h is a quadratic form in the space $S \otimes \bar{S}$, and it is clear from the above equation that this quadratic form generates the bilinear form $\langle h_1, h_2 \rangle$. However, the determinant can be expressed through the Levi-Civita symbol ε as

$$\det h = \frac{1}{2} \varepsilon_{AB} \varepsilon_{A'B'} h^{AA'} h^{BB'}.$$

[11] A definition of spinors for higher-dimensional spacetimes can be formulated using a suitable generalization of quaternions, called **Clifford algebras**. A primer on Clifford algebras is found in Sec. 7.2.

It follows that, *in the same basis* where Eq. (7.9) holds,

$$g_{\mu\nu}\sigma^{\mu}_{AA'}\sigma^{\nu}_{BB'} = \frac{1}{2}\varepsilon_{AB}\varepsilon_{A'B'}. \tag{7.10}$$

Although the above formula is written in a specific basis, it can be reinterpreted as the definition of the fundamental 2-form ε. The form ε is chosen as follows: First we take any nonzero 2-form $\omega \in S^* \wedge S^*$; then we multiply ω by a suitable constant λ so that $\varepsilon = \lambda\omega$ satisfies Eq. (7.10).

Once the 2-form ε is defined on $S \wedge S$, we naturally define the map $\hat{\varepsilon} : S \to S^*$ and the inverse form ε^{-1} on $S^* \wedge S^*$, as well as the corresponding conjugate forms. Note, however, that the 2-form ε is antisymmetric and hence there is a choice of sign in defining the inverse. For instance, we might debate whether the vector $\varepsilon^{-1}\mathbf{a}^*$ should be defined by

$$\varepsilon(\varepsilon^{-1}\mathbf{a}^*, \mathbf{b}) \equiv \mathbf{a}^* \circ \mathbf{b}$$

or by $\varepsilon(\varepsilon^{-1}\mathbf{a}^*, \mathbf{b}) \equiv \mathbf{a}^* \circ \mathbf{b}$? The convention is to use the *first* argument of ε (as shown above) but the *second* argument of ε^{-1}. In the index notation, this corresponds to the following definitions,

$$a_A\varepsilon^{BA} = a^B, \quad b^B\varepsilon_{BA} = b_A, \quad \varepsilon^{AB}\varepsilon_{BC} = -\delta^A{}_C \equiv \varepsilon^A_C.$$

In the index notation, additional trouble arises because $\delta^A{}_C = -\delta_C{}^A$ if we raise and lower indices using ε_{AB}, so one uses the unambiguous notation $\varepsilon^A{}_C = \varepsilon_C{}^A$. The same conventions apply to the conjugate forms $\bar{\varepsilon}_{A'B'}$ and $\bar{\varepsilon}^{A'B'}$.

Statement: Show that the fundamental 2-form ε is invariant under Lorentz transformations.

Hint: $\det \hat{L}_s = 1$.

Solution: A Lorentz transformation of ε is the 2-form $\hat{L}\varepsilon$ defined by

$$\begin{aligned}\hat{L}\varepsilon(\mathbf{a}, \mathbf{b}) &\equiv \varepsilon(\hat{L}_s\mathbf{a}, \hat{L}_s\mathbf{b}) = \varepsilon(\hat{L}_s\mathbf{a} \wedge \hat{L}_s\mathbf{b}) \\ &= (\det \hat{L}_s)\, \varepsilon(\mathbf{a} \wedge \mathbf{b}) = \varepsilon(\mathbf{a}, \mathbf{b}).\end{aligned}$$

In the index notation: The transformed ε_{AB} is

$$\tilde{\varepsilon}_{AB} = L^C_A L^D_B \varepsilon_{CD}$$

which is antisymmetric in A, B and thus should be proportional to ε_{AB}. Contracting with ε^{AB}, we find the coefficient of proportionality to be equal to $(\det L)^2 = 1$.

7.3.2 Relationship of spinors and vectors

Given the fundamental 2-form ε, we may choose a preferred basis adapted to the form. The traditional notation for the basis is $\{o, \iota\}$, where $o, \iota \in S$ and $\varepsilon(o, \iota) = 1$. This is as close as possible to an orthogonal basis, given that ε is antisymmetric. The basis $\{o, \iota\}$ is called the **spin basis** or the **spin frame**. The basis vectors o, ι naturally induce to spin frames in the dual, conjugate, and dual conjugate spaces. It follows that the fundamental 2-form is written as

$$\varepsilon_{AB} = o_A\iota_B - \iota_A o_B; \quad \varepsilon^{-1} = \iota \wedge o = \iota \otimes o - o \otimes \iota.$$

Using the spin frame, we obtain a natural basis in the (four-dimensional) space of $(1\bar{1}, 0)$-tensors:

$$\begin{aligned}\left\{l^{AA'}, m^{AA'}, \bar{m}^{AA'}, n^{AA'}\right\} &\equiv \left\{o^A\bar{o}^{A'}, o^A\bar{\iota}^{A'}, \iota^A\bar{o}^{A'}, \iota^A\bar{\iota}^{A'}\right\} \\ &\equiv \left\{o \otimes \bar{o}, o \otimes \bar{\iota}, \iota \otimes \bar{o}, \iota \otimes \bar{\iota}\right\}.\end{aligned}$$

Transforming these basis tensors by the map $\hat{\sigma}$, we obtain the 4-vectors $\{\mathbf{l}, \mathbf{m}, \bar{\mathbf{m}}, \mathbf{n}\}$. The vectors $\mathbf{l}$ and $\mathbf{n}$ are real-valued but $\mathbf{m}$ and $\bar{\mathbf{m}}$ are complex-valued, because the corresponding basis elements are not Hermitian. It is easy to see, however, that each of the vectors $\{\mathbf{l}, \mathbf{m}, \bar{\mathbf{m}}, \mathbf{n}\}$ is null. In fact, these vectors are a Newman-Penrose null tetrad (see Sec. 2.5).

Statement: Show that $\hat{\sigma}(a^A\bar{b}^{A'})$ is a null vector when $a_A \in S$, $\bar{b}_{B'} \in \bar{S}$ are two arbitrary spinors. Show that the tetrad $\{\mathbf{l}, \mathbf{m}, \bar{\mathbf{m}}, \mathbf{n}\}$ defined above satisfies the required properties of a NP null tetrad: $g(\mathbf{l}, \mathbf{n}) = 1$, $g(\mathbf{m}, \bar{\mathbf{m}}) = -1$, $g(\mathbf{n}, \mathbf{n}) = 0$, etc. Find a linear transformation of the spin basis $\{o, \iota\}$ that preserves $\varepsilon^{-1} = \iota \wedge o$ and induces the tetrad transformation (2.21)-(2.22).

Hint: Use Eq. (7.10).

Answer: The spin basis can be transformed as $\iota \to \iota e^{-\mathrm{i}\phi/2}$, $o \to o e^{\mathrm{i}\phi/2} + A\iota e^{-\mathrm{i}\phi/2}$, where ϕ and A are the parameters of Eqs. (2.21)-(2.22).

Since the $\hat{\sigma}$ map is fixed, we can identify 4-vectors and $(1\bar{1}, 0)$ spinorial tensors. Let us now consider how we could construct some tensor objects out of spinors and in this way obtain a geometric interpretation for spinors.

Firstly, a spinor $\mathbf{c} \in S$ defines a null vector $\mathbf{n}_\mathbf{c} = \hat{\sigma}(\mathbf{c} \otimes \bar{\mathbf{c}})$. In components, this is written as $n^\mu = \sigma^\mu_{AA'}c^A\bar{c}^{A'}$ or more concisely $n^{AA'} = c^A\bar{c}^{A'}$. (The map $\hat{\sigma}$ is usually omitted for brevity since it is a fixed structure relating the spinor and the vector spaces.) One may say figuratively that "a spinor is a square root of a null vector." The null direction $\mathbf{n}_\mathbf{c}$ is the principal geometric information contained in a spinor $\mathbf{c}$. However, spinors differing by a phase, $\mathbf{c}e^{i\alpha}$, define the same null vector $\mathbf{n}_\mathbf{c}$. The phase information can be extracted if we define the **flag bivector**

$$\begin{aligned}F_{(\mathbf{c})} &\equiv \mathbf{c} \otimes \mathbf{c} \otimes \bar{\varepsilon} + \varepsilon \otimes \bar{\mathbf{c}} \otimes \bar{\mathbf{c}}, \\ F^{AA'BB'} &\equiv c^A c^B \bar{\varepsilon}^{A'B'} + \varepsilon^{AB}\bar{c}^{A'}\bar{c}^{B'}.\end{aligned} \tag{7.11}$$

The spinorial tensor $F^{AA'BB'}$ is equivalent to an antisymmetric bivector $F^{\mu\nu}$ satisfying

$$F^{\mu\nu}n_\nu = 0, \quad F^{\mu\nu}F_{\mu\nu} = 0. \tag{7.12}$$

It follows that there exists a ($\mathbf{c}$-dependent) vector $\mathbf{k}_\mathbf{c}$ such that

$$F_\mathbf{c} = \mathbf{n}_\mathbf{c} \wedge \mathbf{k}, \quad F^{\mu\nu} = n^\mu k^\nu - n^\nu k^\mu, \quad g(\mathbf{n}_\mathbf{c}, \mathbf{k}_\mathbf{c}) = 0.$$

The choice of the vector $\mathbf{k}_\mathbf{c}$ is up to a multiple of $\mathbf{n}_\mathbf{c}$, so the pair $(\mathbf{n}_\mathbf{c}, F_\mathbf{c})$ only defines a null *plane* containing the null direction $\mathbf{n}_\mathbf{c}$. This geometric configuration is similar to a "flag" that has a "flagpole" $\mathbf{n}_\mathbf{c}$ and the "flag plane" $\{\mathbf{k}_\mathbf{c}, \mathbf{n}_\mathbf{c}\}$; so we call a pair $(\mathbf{n}_\mathbf{c}, F_\mathbf{c})$ a **flag** corresponding to a spinor $\mathbf{c}$. Spinors $\mathbf{c}$ and $\mathbf{c}'$ differing by sign give the same flag, and conversely, a pair $(\mathbf{n}, F)$ determines a spinor $\mathbf{c}$ up to a sign.[12]

Calculation: Show that a phase rotation, $\mathbf{c} \to e^{\mathrm{i}\alpha}\mathbf{c}$, rotates the flag plane by the angle 2α.

These properties offer a possible geometric interpretation of spinors and illustrate the fact that spinors are related to null directions. Since the flag bivector carries the entire physical information in a spinor (the sign of a spinor is not observable), one could in principle forego spinors and reformulate every spinor equation in terms of unambiguously defined flag bivectors. However, flags are quadratic in spinors, and thus

[12] This can be shown following [36], Eqs. (13.1.47)-(13.1.50), and deriving an explicit formula for $\mathbf{c}$ given a pair $(\mathbf{n}, F)$.

linear equations of motion for spinor fields will become nonlinear tensor equations if written in terms of flags. Spinors provide a simpler formulation of many equations, especially those with massless fields for which null directions are especially important.

7.3.3 Simplification of spinorial tensors

Since the spinor space S is two-dimensional, its properties are significantly simpler than those of a higher-dimensional space. For instance, every antisymmetric tensor of rank $(0,2)$ is proportional to ε. Hence, an arbitrary rank $(0,2)$ spinorial tensor T_{AB} can be simplified into a symmetric tensor and a multiple of ε,

$$\begin{aligned} T_{AB} &= \frac{1}{2}\left(T_{AB}+T_{BA}\right)+\frac{1}{2}\left(T_{AB}-T_{BA}\right) \\ &= T_{(AB)}+\frac{1}{2}\left(T_{CD}\varepsilon^{CD}\right)\varepsilon_{AB}. \end{aligned}$$

Moreover, every spinorial tensor can be reduced to a combination of products of the 2-form ε and some totally symmetric tensors.[13] For example, a rank $(0,2\bar{2})$ tensor can be decomposed as

$$T_{ABA'B'} = T_{(AB)A'B'}+\frac{1}{2}T_{CDA'B'}\varepsilon^{CD}\varepsilon_{AB}.$$

The same decomposition can be applied to the primed indices.

Now let us consider two special cases. Suppose the tensor $T_{ABA'B'}$ comes from an *antisymmetric* 4-tensor $T_{\mu\nu}$, we have $T_{ABA'B'} = -T_{BAB'A'}$, and then

$$T_{ABA'B'} = \phi_{AB}\bar{\varepsilon}_{A'B'}+\varepsilon_{AB}\bar{\phi}_{A'B'},$$

where the symmetric spinorial tensor ϕ_{AB} of rank $(0,2)$ is defined by

$$\phi_{AB} \equiv \frac{1}{2}T_{ABA'B'}\bar{\varepsilon}^{A'B'}.$$

If, on the other hand, $T_{ABA'B'}$ comes from a *symmetric* 4-tensor $T_{\mu\nu}$, then

$$T_{ABA'B'} = T_{(AB)(A'B')}+\frac{1}{4}\varepsilon_{AB}\bar{\varepsilon}_{A'B'}\left(T_{CDC'D'}\varepsilon^{CD}\bar{\varepsilon}^{C'D'}\right). \tag{7.13}$$

Thus the only nontrivial spinorial tensors we need to consider are totally symmetric ones. For those, the following decomposition property holds.

Theorem: Every totally symmetric spinorial tensor can be decomposed into a symmetric product of spinors:

$$T_{(AB)} = a_{(A}b_{B)}, \quad T_{(ABC)} = a_{(A}b_{B}c_{C)},$$

etc. The null flagpoles $\mathbf{n_a}$, $\mathbf{n_b}$, etc., corresponding to the spinors $\mathbf{a}$, $\mathbf{b}$, etc., are called the **principal null directions** of the spinorial tensor T.

Proof: A totally symmetric tensor of rank n is completely determined by its values on a single vector; e.g. $T(\mathbf{x},\mathbf{y},\mathbf{z})$ can be restored if we know $T(\mathbf{x},\mathbf{x},\mathbf{x})$ for every $\mathbf{x}$. Setting $\mathbf{x} = o+\iota z$, where o and ι are two spinors building a spin frame, we find that $T(\mathbf{x},\mathbf{x},\mathbf{x})$ is a cubic polynomial in z. A cubic polynomial can be factorized as

$$T(\mathbf{x},\mathbf{x},\mathbf{x}) = (a_0+a_1z)(b_0+b_1z)(c_0+c_1z),$$

where $a_0, a_1, ..., c_1$ are suitable complex numbers (not uniquely determined). We can interpret these numbers as the spin frame components a_A, b_A, c_A of three spinors $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$. Since T_{ABC} is totally symmetric, we have obtained a decomposition $T_{(ABC)} = a_{(A}b_Bc_{C)}$. The general case of a spinorial tensor of rank n is treated analogously.

Remark: Classification of 4-tensors. Since every Minkowski 4-tensor can be mapped into a spinorial tensor, we may reduce every 4-tensor first to totally symmetric spinorial tensors, and then to products of individual spinors. In this way, every 4-tensor gives rise to a set of principal null directions. This construction provides a classification of 4-tensors based on how many of their principal null directions coincide.

Calculation: Derive the spinorial representation of the Levi-Civita symbol $\varepsilon_{\kappa\lambda\mu\nu}$ by mapping it to spinorial tensors and reducing to products of the fundamental 2-form ε_{AB}.

Answer:

$$\varepsilon_{AA'BB'CC'DD'} = \mathrm{i}(\varepsilon_{AB}\varepsilon_{CD}\bar{\varepsilon}_{A'C'}\bar{\varepsilon}_{B'D'}-\bar{\varepsilon}_{A'B'}\bar{\varepsilon}_{C'D'}\varepsilon_{AC}\varepsilon_{BD}). \tag{7.14}$$

Statement: Show that the flag bivector $F^{\mu\nu}$ given by Eq. (7.11) satisfies

$$\varepsilon_{\kappa\lambda\mu\nu}F^{\kappa\lambda}F^{\mu\nu} = 0. \tag{7.15}$$

This is an additional property of flags, besides the properties given by Eq. (7.12).

7.4 Equations for spinor fields

We have worked in the Minkowski spacetime $\mathbb{R}^4$, interpreted as the tangent space at a single point of a spacetime manifold. Now we shall consider a (curved) spacetime where a spinor field should be defined at every point.

7.4.1 Spinors in curved spacetime

A **spinor field** is, roughly speaking, a spinor-valued function on a manifold $\mathcal{M}$; thus we would like to write $\psi(p) \in S$ to denote the value of the spinor field at a point p. However, spinors are tied to the Lorentz transformations in a tangent space $T_p\mathcal{M}$, and tangent spaces differ at different points of the manifold. Therefore, we cannot directly compare the values of spinors at different points.

A precise picture of a spinor field is provided by the construction of a vector bundle (see Sec. 6.3). We consider a **spinor bundle** with the spacetime $\mathcal{M}$ as the base and the spinor space S as the fiber. Thus, we would like to have a separate copy of the spinor space S at each point $p \in \mathcal{M}$. A spinor field is defined[14] as a (smooth) section of the spinor bundle; thus, the value $\psi(p)$ at a point p is a spinor belonging to the copy $S(p)$ of the spinor space.

In the previous section, we constructed the spinor space S through the Minkowski metric, $\eta \equiv \mathrm{diag}(1,-1,-1,-1)$. However, we now need to assign a spinor space $S(p)$ to each point p in a curved spacetime where the Minkowski metric is replaced by a general p-dependent metric tensor $g(p)$. Therefore, we need to modify the construction of the spinor space,

[13] See [27] or [32] for a complete proof.

[14] We ignore possible topological difficulties with the construction of a spinor bundle for a given base manifold $\mathcal{M}$. The spinor bundle exists for all physically relevant spacetimes $\mathcal{M}$.

so that the map $\hat{\sigma}$ and the antisymmetric tensor ε_{AB} are compatible with the metric $g(p)$ at each point. This is achieved by the following construction.

Consider a set of four vectors $\{\mathbf{e}_j\}$, $j = 0, 1, 2, 3$ such that

$$g(\mathbf{e}_j, \mathbf{e}_k) \equiv g_{\mu\nu} e_j^\mu e_k^\nu = \eta_{jk}, \tag{7.16}$$

where η is defined above as a standard, fixed matrix representing the "fiducial" Minkowski metric. These four vectors obviously form a basis in the tangent space $T_p\mathcal{M}$, and such a basis is called an **orthonormal tetrad**. We can always choose such a tetrad $\{\mathbf{e}_j(p)\}$ in some neighborhood of any point p. The tetrad $\{\mathbf{e}_j\}$ can also be viewed as a map from an auxiliary Minkowski space $\mathbb{R}^4$ to the tangent space $T_p\mathcal{M}$, whereby a 4-vector $(x_0, x_1, x_2, x_3) \in \mathbb{R}^4$ is mapped into the spacetime vector $\mathbf{x} \equiv \sum_{j=0}^3 \mathbf{e}_j x_j \in T_p\mathcal{M}$. To contrast 4-vectors and 4-tensors in the fiducial Minkowski space $\mathbb{R}^4$ with vectors and tensors defined through tangent spaces $T_p\mathcal{M}$ in the actual curved spacetime, the latter are sometimes called **world vectors** and **world tensors**. Since $\{\mathbf{e}_j(p)\}$ is a basis, we obtain a 1-to-1 correspondence between $\mathbb{R}^4$ and $T_p\mathcal{M}$; it is easy to see that the inverse map $T_p\mathcal{M} \to \mathbb{R}^4$ is defined as $\mathbf{x} \to (x_j)$, where $x_j \equiv g(\mathbf{x}, \mathbf{e}_j(p))$, $j = 0, 1, 2, 3$.

The construction of spinors at a point p in a curved spacetime manifold $\mathcal{M}$ uses the fixed, Minkowski-space matrices $\hat{\sigma}_{(M)} \equiv \sigma^j_{AB'}$, $j = 0, 1, 2, 3$. We define the modified map $\hat{\sigma}(p) : S \otimes \bar{S} \to T_p\mathcal{M}$ by using a tetrad $\{\mathbf{e}_j(p)\}$ which maps the result of $\hat{\sigma}_{(M)}$ into $T_p\mathcal{M}$:

$$\hat{\tilde{\sigma}}(h^{AB'}) = \sum_{j=0}^3 \sigma^j_{AB'} h^{AB'} \mathbf{e}_j \in T_p\mathcal{M}.$$

This defines the curved-spacetime map $\hat{\tilde{\sigma}}(p)$ which can be thought of a world-vector-valued spinorial $(0, 1\bar{1})$-tensor,

$$\tilde{\sigma}_{AB'}(p) \equiv \sum_{j=0}^3 \sigma^j_{AB'} \mathbf{e}_j(p); \quad \tilde{\sigma}^\mu_{AB'} \equiv \sum_{j=0}^3 \sigma^j_{AB'} e^\mu_j.$$

Due to the defining property (7.16) of a tetrad, we have

$$\begin{aligned} g_{\mu\nu}(p) \tilde{\sigma}^\mu_{AA'}(p) \tilde{\sigma}^\nu_{BB'}(p) &= \sum_{j,k=0}^3 g_{\mu\nu}(p) e^\mu_j(p) \sigma^j_{AA'} e^\mu_k(p) \sigma^k_{BB'} \\ &= \sum_{j,k=0}^3 \eta_{jk} \sigma^j_{AA'} \sigma^k_{BB'}. \end{aligned}$$

Therefore, the spinorial 2-form ε_{AB} automatically satisfies the property analogous to Eq. (7.10) at every point p,

$$g_{\mu\nu}(p) \tilde{\sigma}^\mu_{AA'}(p) \tilde{\sigma}^\nu_{BB'}(p) = \frac{1}{2} \varepsilon_{AB} \varepsilon_{A'B'}.$$

Note that the definition of the spinor spaces $S(p)$ involves an arbitrary choice of an orthonormal tetrad $\{\mathbf{e}_j(p)\}$. Different choices of the tetrad are related to teach other with Lorentz transformations. (For any two orthonormal tetrads $\{\mathbf{e}_j\}$ and $\{\mathbf{e}'_j\}$, there exists a unique Lorentz transformation $\hat{L}$ such that $\hat{L}\mathbf{e}_j = \mathbf{e}'_j$.) The map $\hat{\sigma}$ and the 2-form ε_{AB} are invariant under Lorentz transformations. Thus, any Lorentz-invariant spinorial expression, such as $\phi_{AB} \bar{\varepsilon}_{A'B'} g^{AA'BB'}$, will be mapped into a world scalar, independently of the choice of the tetrad.

Remark: Since spinors are not defined directly through points of the manifold, there is no concept of arbitrary pointwise transformations (diffeomorphisms) directly applied to spinor fields! The only transformations that apply to spinors are Lorentz transformations $\hat{L}_s$ within the spinor space. Thus one cannot define the change in a spinor field due to the flow of an arbitrary vector field $\mathbf{v}$ on $\mathcal{M}$; in other words, the Lie derivative "$\mathcal{L}_\mathbf{v}\psi$" of a spinor with respect to a vector is undefined.[15] However, this does not mean that the presence of spinors in a theory violates general covariance. General point transformations *can* be applied to the tetrad vectors $\mathbf{e}_j$ and thus effectively change the map $\hat{\sigma}$, which is the only connection between the spinor space S and the tangent space $T_p\mathcal{M}$. Field theories involving spinors can be formulated in terms of Lorentz invariants in the spinor space, and invariants involving the map $\hat{\sigma}$ and the world metric g, and then these theories are generally covariant.

7.4.2 Covariant derivative on spinors

The Levi-Civita connection ∇ (but not any other connection, see below!) can be uniquely extended to a covariant derivative $\nabla_\mathbf{v}\psi$ on spinor fields. In addition to the usual properties (linearity, Leibnitz rule, torsion-freeness), the Levi-Civita connection satisfies

$$\begin{aligned} \nabla_\mathbf{v} \bar{\psi} &= \overline{\nabla_\mathbf{v} \psi}, \\ \nabla_\mathbf{v} \varepsilon_{AB} &= \nabla_\mathbf{v} \varepsilon^{AB} = 0, \\ \nabla_\mathbf{v} \hat{\sigma} &= 0. \end{aligned}$$

Naturally, the gradient operator ∇_μ is also equivalent to the $(0, 1\bar{1})$-tensor $\nabla_{AA'}$ in the spinor space,

$$\nabla_\mathbf{v} \equiv v^\mu \nabla_\mu \equiv v^\mu \sigma^{AA'}_\mu \nabla_{AA'}.$$

Perhaps, it appears strange that such "simple" operations as the coordinate derivative ∂_μ cannot be defined on spinor fields. One way to explain the fact that only the Levi-Civita connection is defined on spinors is to consider the flag field $(\mathbf{n}_\psi, F_\psi)$ corresponding to a spinor field ψ. Suppose we attempt to define a parallelly transported spinor ψ using some connection ∇. We say that ψ is "parallelly transported along a vector $\mathbf{v}$" if $\nabla_\mathbf{v}\psi^A = 0$. Then we expect that the flag $(\mathbf{n}_\psi, F_\psi)$ should also be parallelly transported along $\mathbf{v}$. However, the basic properties (7.12), (7.15) of a flag (the vector $\mathbf{n}_\psi$ should be null, F_ψ should be transverse to $\mathbf{n}_\psi$, etc.) involve the metric g. Unless the connection ∇ is compatible with the metric g, a parallelly transported flag will fail to remain a valid flag; for instance, the vector $\mathbf{n}_\psi$ may fail to remain null, F_ψ may fail to remain transverse to $\mathbf{n}_\psi$, etc. Only the metric-compatible (Levi-Civita) connection guarantees that the pair $(\mathbf{n}_\psi, F_\psi)$ will remain a valid flag after an arbitrary parallel transport.

Another argument is based on the concept of associated bundles. The spinor bundle is associated to the gauge group $SL(2, \mathbb{C})$, which is essentially equivalent to the Lorentz group $SO(3, 1)$. The Levi-Civita connection ∇ is the gauge-invariant connection with respect to the gauge group, because ∇ is compatible with the metric (see Sec. 6.3.5). No other connection is admissible in the associated bundle. Thus, no other connection can be defined on spinors.

[15] Unless the vector $\mathbf{v}$ is a conformal Killing vector for the metric. See [27], vol. 2, § 6.6.

7.4.3 Maxwell equations

The electromagnetic field corresponds to a particle of spin 1 and is described by a 2-form $F = dA$, where A is a 1-form representing the electromagnetic potential. In the index notation, F is the antisymmetric tensor $F^{\mu\nu}$. We have seen that such a tensor can be reduced to a symmetric spinorial tensor ϕ^{AB} by

$$\phi^{AB} = \frac{1}{2}\bar{\varepsilon}_{A'B'}F^{AA'BB'},$$
$$F^{AA'BB'} = \phi^{AB}\bar{\varepsilon}^{A'B'} + \varepsilon^{AB}\bar{\phi}^{A'B'}.$$

Now we shall show that the Maxwell equations can be written concisely in terms of ϕ^{AB}, namely

$$\nabla_{AA'}\phi^{AB} = \frac{1}{2}\bar{\varepsilon}_{A'B'}j^{BB'}, \tag{7.17}$$

where $j^{BB'} \equiv j^{\mu}\sigma_{\mu}^{BB'}$ is the spinorial version of the familiar charge 4-current $j^{\mu} \equiv \left(\rho, \vec{j}\right)$.

It is convenient to define the Hodge dual ${}^*F$ which is again a 2-form,

$${}^*F^{\mu\nu} \equiv \frac{1}{2}\varepsilon^{\kappa\lambda\mu\nu}F_{\kappa\lambda},$$

and to rewrite the Maxwell equations equivalently as

$$\nabla_{\mu}F^{\mu\nu} = j^{\nu},$$
$$\nabla_{\mu}\left({}^*F^{\mu\nu}\right) = 0.$$

(The last equation is equivalent to $\nabla^{[\lambda}F^{\mu\nu]} = 0$, i.e. $dF = 0$, which is a consequence of the identity $ddA = 0$, which is equivalent to $F = dA$.) Using the explicit expression (7.14) for the Levi-Civita symbol $\varepsilon_{\kappa\lambda\mu\nu}$, it is straightforward to verify that

$$F^{AA'BB'} + i\left(*F\right)^{AA'BB'} = 2\phi^{AB}\bar{\varepsilon}^{A'B'}.$$

Then, the Maxwell equations are equivalent to

$$j^{\nu} = \nabla_{\mu}(F^{\mu\nu} + i*F^{\mu\nu}) = 2\nabla_{AA'}\left(\phi^{AB}\bar{\varepsilon}^{A'B'}\right)$$
$$= 2\left(\nabla_{AA'}\phi^{AB}\right)\bar{\varepsilon}^{A'B'},$$

since ε_{AB} is "constant" under ∇. Multiplying both sides by $\bar{\varepsilon}_{B'C'}$, we find that Eq. (7.17) is equivalent to the Maxwell equations.

Statement: Consider a spinorial tensor $\phi^{AB} = \psi^{A}\psi^{B}$, where ψ^{A} is some spinor, and show that the corresponding tensor $F^{\mu\nu}$ describes an electromagnetic wave in vacuum.

Hint: In this case, $F^{\mu\nu}$ is the flag of the spinor ψ, so $F^{\mu\nu}$ must satisfy the properties (7.12), (7.15) of a flag. For an electromagnetic field described by 3-dimensional vectors $\vec{E}$ and $\vec{B}$, we have

$$F_{\mu\nu}F^{\mu\nu} = 2\left(|\vec{B}|^2 - |\vec{E}|^2\right),$$
$$\varepsilon_{\kappa\lambda\mu\nu}F^{\kappa\lambda}F^{\mu\nu} = -2\vec{E}\cdot\vec{B}.$$

Calculation: Show that the energy-momentum tensor of the electromagnetic field is

$$T_{\mu\nu} = \frac{1}{4\pi}\left(\frac{1}{4}g_{\mu\nu}F_{\alpha\beta}F^{\alpha\beta} - F_{\mu\alpha}F^{\alpha\beta}g_{\beta\nu}\right)$$
$$\equiv T_{AA'BB'} = \frac{1}{2\pi}\phi_{AB}\bar{\phi}_{A'B'}.$$

Hint: Use Eq. (7.13) to simplify the spinorial form $T_{ABA'B'}$ of the symmetric tensor $T_{\mu\nu}$.

The spinorial form of the Maxwell equation has its roots in group theory.[16] In a heuristic language, this is the simplest and the most natural relativistic equation for a symmetric spinorial tensor of rank $(2,0)$.

7.4.4 Dirac equation

The Dirac equation describes a spinor field ϕ^{A} corresponding to a massive particle of spin $\frac{1}{2}$. We shall first consider the Dirac equation in Minkowski spacetime and then generalize to a curved spacetime.

The simplest nontrivial Lorentz-invariant equation for a field ϕ^{A} would be

$$\left(\Box + m^2\right)\phi^{A} = 0, \tag{7.18}$$

where the D'Alambert operator $\Box$ is

$$\Box \equiv \nabla_{\mu}\nabla^{\mu} = \nabla_{AA'}\nabla^{AA'}.$$

Dirac's main motivation was to rewrite this second-order equation as a system of first-order equations. The standard way to achieve this is to introduce extra fields for first derivatives. So let us introduce an auxiliary (complex conjugate and dual) spinor field $\bar{\sigma}_{A'}$, which will be the first derivative of ϕ^{A},

$$\bar{\sigma}_{A'} \equiv \nabla_{AA'}\phi^{A}.$$

We expect that the equation for $\bar{\sigma}_{A'}$ will be of the form $\nabla^{BA'}\bar{\sigma}_{A'} = (...)$, and we shall now derive that equation.

We need to simplify the expression

$$\nabla^{BA'}\bar{\sigma}_{A'} = \nabla^{BA'}\nabla_{AA'}\phi^{A}. \tag{7.19}$$

Let us first consider a more general spinorial tensor of this form,

$$X_{AA'BB'C} \equiv \nabla_{BB'}\nabla_{AA'}\phi_{C};$$

when we are done simplifying $X_{AA'BB'C}$, the term (7.19) will be expressed as

$$\nabla^{BB'}\nabla_{AA'}\phi^{A} = \bar{\varepsilon}^{A'B'}\varepsilon^{AC}\varepsilon^{BD}X_{DB'AA'C}. \tag{7.20}$$

Since we are working in flat space where the curvature identically vanishes, the covariant derivatives commute (even when acting on arbitrary tensors or spinors), so

$$X_{AA'BB'C} = X_{BB'AA'C}.$$

Thus we can apply the decomposition (7.13), suitable for symmetric 4-tensors, to the first four indices of $X_{AA'BB'C}$:

$$X_{AA'BB'C} = X_{(AB)(A'B')C} + \frac{1}{4}\varepsilon_{AB}\bar{\varepsilon}_{A'B'}Y_{C}, \tag{7.21}$$
$$Y_{C} \equiv X_{AA'BB'C}\varepsilon^{AB}\bar{\varepsilon}^{A'B'}.$$

It is easy to see that

$$Y_{C} = \varepsilon^{AB}\bar{\varepsilon}^{A'B'}\nabla_{AA'}\nabla_{BB'}\phi_{C} = g^{\alpha\beta}\nabla_{\alpha}\nabla_{\beta}\phi_{C} = \Box\phi_{C}.$$

[16] It can be shown the homogeneous version of the spinorial Maxwell equation, $\nabla_{AA'}\phi^{AB} = 0$, determines the irreducible representation of the Poincaré group having mass 0 and spin 1.

When we substitute Eq. (7.21) into Eq. (7.20), the symmetric term $X_{(AB)(A'B')C}$ will be canceled after contracting with $\bar{\varepsilon}^{A'B'}$ in Eq. (7.20). Therefore, only the term containing Y_C survives. After some simplification, we find the equation for $\bar{\sigma}_{A'}$,

$$\nabla^{BA'}\bar{\sigma}_{A'} = \nabla^{BA'}\nabla_{AA'}\phi^A = \frac{1}{2}\Box\phi^B = -\frac{1}{2}m^2\phi^B.$$

The conventional definition of the auxiliary field is $\bar{\chi}_A \equiv \sqrt{2}m^{-1}\bar{\sigma}_{A'}$, which absorbs the awkward factor $\frac{1}{2}m^2$. Then the equations acquire a more symmetric form,

$$\nabla_{AA'}\phi^A = \frac{m}{\sqrt{2}}\bar{\chi}_{A'},$$
$$\nabla^{BA'}\bar{\chi}_{A'} = -\frac{m}{\sqrt{2}}\phi^B.$$

At this point, we are done rewriting the second-order equation (7.18) as a system of first-order equations. We now introduce a four-dimensional, complex-valued vector $\psi \in S \oplus \bar{S}^*$ instead of the pair $\left(\phi^B, \bar{\chi}_{A'}\right)$, and write the above system of equations as a first-order equation for ψ,

$$\gamma^\mu\nabla_\mu\psi = m\psi, \tag{7.22}$$

where γ^μ are the required 4×4 complex matrices called the **Dirac matrices**,

$$\gamma^\mu = \sqrt{2}\begin{pmatrix} 0 & -\sigma^\mu_{AA'} \\ \sigma^{\mu BB'} & 0 \end{pmatrix}.$$

The vector ψ is called a **Dirac spinor**.

The Dirac equation in curved spacetime is Eq. (7.22) rather than Eq. (7.18). In the presence of curvature, these equations are not equivalent since covariant derivatives do not commute.

A Elements of Special and General Relativity

This is a brief review of basic concepts of relativity and differential geometry. In this appendix I mostly use the traditional index notation to facilitate comparison with the prevalent physics literature. The material in this appendix is assumed known in the main part of the book; the explanations here cover only a portion of the material in standard relativity textbooks. A good introductory textbook is [29].

Special Relativity (SR) is a theory describing the motion of light and point masses in cases when gravity can be neglected. The name **General Relativity** (GR) is used to denote Einstein's theory of gravitation, which he developed as a generalization of SR. I shall now review the mathematical foundations of these theories, omitting most proofs.

A.1 Special Relativity

The theory of Special Relativity is based on two main postulates: (i) All the laws of physics are the same in every inertial reference frame.[1] (ii) The speed of light, denoted c, is independent of the speed of the light source. One can derive all the statements of SR from these two postulates. For instance, it is shown that no massive body can be accelerated from rest to a velocity greater than or equal to c. Also, the distance x' and time t' measured in a moving reference frame must be different from the distance x and time t measured in a rest frame. Namely, the results of these measurements are related by the **Lorentz transformation**

$$ct' = \gamma\,(ct - \beta x)\,, \quad x' = \gamma(x - \beta ct),$$
$$\gamma \equiv \frac{1}{\sqrt{1-\beta^2}}, \quad \beta \equiv \frac{v}{c},$$

where we assume that the motion is in the positive x direction, and $\beta < 1$ is the relative velocity of the two reference frames, measured in the units of c. Since c is an absolute constant, one can measure velocities in the units of c and distances in the units of time. This choice of measurement units is mathematically equivalent to setting $c \equiv 1$ in all equations, and we adopt this convention everywhere in this text.

The fact that coordinates and times are generally not measured to be the same in different reference frames is at the core of the theory of relativity; for instance, events seen as simultaneous for one observer may precede one another for a different observer. Thus, simultaneity is not absolute but is only defined *relative* to an observer's reference frame (this is one justification for the name "relativity").

A.1.1 Spacetime

Rather than use distances and times measured in different reference frames, one adopts a more convenient way to describe events. Namely, one introduces the **spacetime**, which is an auxiliary four-dimensional vector space $\mathcal{M}$ with coordinates $\{t, x, y, z\}$. This space $\mathcal{M} \equiv \mathbb{R}^4$ is called the **Minkowski spacetime**. In the spacetime picture, events $\mathbf{x} \in \mathcal{M}$, $\mathbf{x} \equiv \{t, x, y, z\}$ are *points* of the spacetime (vectors such as $\mathbf{x}$ are denoted by boldface letters). Bodies follow **worldlines** $x(t), y(t), z(t)$ which may be drawn as lines in $\mathcal{M}$ and parametrized by functions $x(t), y(t), z(t)$. It is usually more convenient to parametrize worldlines by four functions $t(\tau), x(\tau), y(\tau), z(\tau)$, where τ is a real-valued parameter. Then a worldline can be represented by a point-valued function $\mathbf{x}(\tau)$.

A Lorentz transformation between inertial frames will then be seen as a linear transformation of the coordinates in the spacetime, $\{t, x, y, z\} \to \{t', x', y', z'\}$. However, it is natural to think of the spacetime as an independent arena where events happen, regardless of coordinates introduced in it. In other words, the coordinate values (t, x, y, z) ascribed to an event $\mathbf{x}$ by a particular inertial observer may vary, but the event $\mathbf{x}$ happens by itself at a particular "place-and-time," whether observed or not. (We may imagine the spacetime $\mathcal{M}$ as a chart containing complete histories of all the bodies and a complete record of all the events that happened or will ever happen.) To differentiate between ordinary three-dimensional vectors and four-dimensional vectors from $\mathcal{M}$, the latter are frequently called **4-vectors**.

It is easy to see that the Lorentz transformation preserves the quadratic form,

$$g(\mathbf{x}, \mathbf{x}) \equiv t^2 - x^2 - y^2 - z^2,$$

called the **relativistic interval** or the **Minkowski metric**. In fact, the metric g has a far greater significance than merely an invariant of Lorentz transformations. The most important use of g is in computing distances and time intervals between events. Given two events, $\mathbf{x}$ and $\mathbf{y}$, we may ask whether an inertial observer might pass through both $\mathbf{x}$ and $\mathbf{y}$. If this is the case then these events will have coordinates $\{t_1, 0, 0, 0\}$ and $\{t_2, 0, 0, 0\}$ in this observer's rest frame. The time interval, $t_2 - t_1$, measured according to this observer's clock, is called the observer's **proper time** interval between $\mathbf{x}$ and $\mathbf{y}$. This interval can be computed as

$$\Delta\tau(\mathbf{x}, \mathbf{y}) \equiv \sqrt{(t_2 - t_1)^2} = \sqrt{g(\mathbf{y} - \mathbf{x}, \mathbf{y} - \mathbf{x})}.$$

The 4-vector $\mathbf{y} - \mathbf{x}$ connecting the events in the Minkowski space is then called a **timelike vector**. The last term in the above formula is expressed through g and is thus Lorentz-invariant. Hence, any observer can compute the proper time $\Delta\tau\,(\mathbf{x}, \mathbf{y})$ using this formula.

Another possibility is the existence of a lightray connecting the events $\mathbf{x}$ and $\mathbf{y}$. **Lightrays** are worldlines of bodies that move with the speed of light; for example, the line $x = t, y = z = 0$ is a lightray. It is easy to see that the squared interval in this case is $g(\mathbf{y} - \mathbf{x}, \mathbf{y} - \mathbf{x}) = 0$. The 4-vector $\mathbf{y} - \mathbf{x}$ connecting the events is then called a **null vector**.

Finally, there may be an inertial observer for whom the events $\mathbf{x}, \mathbf{y}$ are *simultaneous* and thus have coordinates

[1] Of course, it is also postulated that such inertial reference frames *exist* and, in particular, are approximately realized in a laboratory freely floating in empty space.

$\{t_0, x_1, x_2, x_3\}$ and $\{t_0, y_1, y_2, y_3\}$. In this case, $g(\mathbf{y}-\mathbf{x}, \mathbf{y}-\mathbf{x}) < 0$, the 4-vector $\mathbf{y}-\mathbf{x}$ is called a **spacelike vector**, and the real number

$$\Delta L(\mathbf{x}, \mathbf{y}) \equiv \sqrt{\sum_{j=1}^{3} (y_j - x_j)^2} = \sqrt{-g(\mathbf{y}-\mathbf{x}, \mathbf{y}-\mathbf{x})}$$

is equal to the distance between the points $\mathbf{x}, \mathbf{y}$ in the reference frame where both events occur simultaneously. The quantity $\Delta L(\mathbf{x}, \mathbf{y})$ is called the **proper distance** between the events. Again, this distance can be calculated in any other reference frame using the above formula involving g.

Conversely, it can be shown that the sign of $g(\mathbf{y}-\mathbf{x}, \mathbf{y}-\mathbf{x})$ unambiguously specifies which of the three above cases (timelike, null, or spacelike) takes place. Thus, the metric g provides not only a means to compute the proper time and the proper distance between events in an arbitrary reference frame, but also *classifies* pairs of events according to whether their separation is timelike, null, or spacelike. Since no material bodies or signals can propagate faster than light, only events separated by a timelike or null interval can influence or cause each other. Hence, the metric g determines which points of $\mathcal{M}$ can causally influence each other, i.e. describes the **causal structure** of the spacetime.

We have arrived at a picture of the Minkowski spacetime $\mathcal{M}$ whose points are events and where a metric g is defined. Reference frames corresponding to different inertial observers are merely coordinate systems that we may introduce on $\mathcal{M}$. Using this picture, known physical laws can be reformulated only using 4-vectors $\mathbf{x}$ and the metric g, without reference to a particular coordinate system on $\mathcal{M}$. In other words, the laws of physics have a frame-invariant character compatible with the Lorentz transformations, which means that these laws are **relativistic** (agree with Special Relativity).

Note that, according to a standard result of linear algebra, the quadratic form $g(\mathbf{x}, \mathbf{x})$ gives rise to a symmetric bilinear form $g(\mathbf{x}, \mathbf{y})$ which we may define by

$$g(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\left(g(\mathbf{x}+\mathbf{y}, \mathbf{x}+\mathbf{y}) - g(\mathbf{x}, \mathbf{x}) - g(\mathbf{y}, \mathbf{y})\right).$$

Clearly, g plays the role of the scalar product in the 4-dimensional vector space $\mathcal{M}$. In coordinates: if $\mathbf{x} \equiv \{x_0, x_1, x_2, x_3\}$ and $\mathbf{y} \equiv \{y_0, y_1, y_2, y_3\}$ then

$$g(\mathbf{x}, \mathbf{y}) = x_0 y_0 - x_1 y_1 - x_2 y_2 - x_3 y_3.$$

This bilinear form is also called the **Minkowski metric**. Two 4-vectors $\mathbf{x}, \mathbf{y}$ are called **orthogonal** to each other if $g(\mathbf{x}, \mathbf{y}) = 0$. Note that the unusual signs in the metric g make the geometric interpretation of the scalar product $g(\mathbf{x}, \mathbf{y})$ and of the orthogonality somewhat difficult. For instance, a null vector $\mathbf{x}$, such as $\{1, 1, 0, 0\}$, is "orthogonal to itself" since $g(\mathbf{x}, \mathbf{x}) = 0$. However, in any given reference frame, the time direction is always orthogonal to every spatial direction, and the three spatial directions are also mutually orthogonal, just like in normal Euclidean geometry. From the point of view of linear algebra, the assumption of an "inertial frame" is equivalent to choosing a basis $\{\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ in the vector space $\mathcal{M}$ such that $\mathbf{e}_0$ is timelike, $\mathbf{e}_{1,2,3}$ are spacelike, and $g(\mathbf{e}_0, \mathbf{e}_0) = 1$, $g(\mathbf{e}_0, \mathbf{e}_j) = 0$, $g(\mathbf{e}_j, \mathbf{e}_k) = -\delta_{jk}$ for $j, k = 1, 2, 3$ (the basis vectors are orthonormal with respect to the bilinear form g).

A.1.2 Motion of bodies in SR

The motion of a massive body is described by a worldline $\mathbf{x}(\tau)$, where τ is an arbitrary parameter. It is convenient to choose τ to be the proper time in the body's rest frame. With that choice, many equations are simplified; in particular, the 4-vector called **4-velocity**, defined by

$$\mathbf{v} \equiv \dot{\mathbf{x}} \equiv \frac{d}{d\tau}\mathbf{x},$$

satisfies $g(\mathbf{v}, \mathbf{v}) = 1$ because $g(\mathbf{v}, \mathbf{v})$ is frame-invariant and in the body's rest frame we have $\mathbf{x} = \{\tau, 0, 0, 0\}$ and thus $\mathbf{v} = \{1, 0, 0, 0\}$.

The **4-acceleration,**

$$\mathbf{a} \equiv \frac{d}{d\tau}\mathbf{v} = \frac{d^2}{d\tau^2}\mathbf{x},$$

is always orthogonal to $\mathbf{v}$ in the sense of the metric g, namely $g(\mathbf{v}, \mathbf{a}) = 0$. When τ is the proper time, a noninteracting body moves according to the equation

$$\frac{d^2}{d\tau^2}\mathbf{x} = 0.$$

Solutions of this equation are straight worldlines, which describe motion with constant velocity.

The above equation of motion can be derived from a variational principle,

$$\delta \int \sqrt{g(\dot{\mathbf{x}}, \dot{\mathbf{x}})}\, \mathrm{d}\tau = 0,$$

which means that the inertial trajectory is an extremum of the proper time among all the possible worldlines $\mathbf{x}(\tau)$. (It can be shown that the extremum in fact a global maximum.) With the normalization $g(\dot{\mathbf{x}}, \dot{\mathbf{x}}) = 1$, the variational principle becomes simply $\delta \int \mathrm{d}\tau = 0$.

To describe a body of rest mass m that interacts with other bodies, one writes the action

$$S = -m \int \mathrm{d}\tau + \int L_{\text{int}}[\mathbf{x}, \dot{\mathbf{x}}]\mathrm{d}\tau,$$

where the Lagrangian $L_{\text{int}}[\mathbf{x}, \dot{\mathbf{x}}]$ is a function of the worldline $\mathbf{x}(\tau)$ that represents interactions with other bodies. The trajectory $\mathbf{x}(\tau)$ of the body should extremize L with respect to all possible worldlines. The resulting equation of motion is the "relativistic Newton's law,"

$$m\frac{d^2}{d\tau^2}\mathbf{x} = \mathbf{f}, \tag{A.1}$$

where the 4-vector $\mathbf{f}$ is the **4-force** which may in general depend on $\mathbf{x}$ and $\dot{\mathbf{x}}$. The 4-vector $\mathbf{p} \equiv m\dot{\mathbf{x}}$ is called the **4-momentum** and carries the information about the energy and the momentum of the body, $\mathbf{p} = \{E, p_1, p_2, p_3\}$, where $\{p_1, p_2, p_3\} \equiv \vec{p}$ are the components of the 3-momentum. Note the normalization, $g(\mathbf{p}, \mathbf{p}) = m^2$, which implies the energy law

$$E = \sqrt{m^2 + \vec{p}^2}.$$

A.2 Index notation

Let us now introduce the index notation. In a chosen reference frame, a 4-vector $\mathbf{x}$ has four coordinates (also called **components**) that are usually denoted x^0, x^1, x^2, x^3 with an upper

(superscript) index; the index value 0 means "time" and the values 1, 2, 3 mean spatial coordinates. The scalar product of two 4-vectors $\mathbf{x}, \mathbf{y}$ can then be written as

$$g(\mathbf{x}, \mathbf{y}) = \sum_{\mu=0}^{3} \sum_{\nu=0}^{3} g_{\mu\nu} x^\mu x^\nu,$$

where $g_{\mu\nu}$ is a 4×4 matrix,

$$g_{\mu\nu} \equiv \operatorname{diag}(1, -1, -1, -1) \equiv \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}.$$

The low (subscript) position of indices in the symbol $g_{\mu\nu}$ is not accidental. For brevity, it is customary to use the **Einstein summation convention**, which consists of dropping the summation signs and implying summation every time an identical index letter appears once as a subscript and once as a superscript. Then, $g(\mathbf{x}, \mathbf{y})$ is written as $g_{\mu\nu} x^\mu x^\nu$. Also, one does not distinguish between a vector $\mathbf{x}$ and the array x^μ of its components in some basis; one simply says "a vector x^μ." This is the essence of the index notation: It is always assumed that a basis is chosen and that one is performing calculations with components of vectors and tensors in the chosen basis. All 4-vectors and 4-tensors then appear as tables of numbers, which are indexed by (Greek) letters, for example: x^μ, $g^{\alpha\beta}$, $\varepsilon_{\kappa\lambda\mu\nu}$.

A Lorentz transformation is described by a matrix Λ^α_β which acts on 4-vectors x^μ as

$$\Lambda : \mathbf{x} \to \mathbf{x}'; \quad x'^\mu = \Lambda^\mu_\nu x^\nu.$$

It is easy to see that the condition of Lorentz invariance of the metric is written as

$$\Lambda^\mu_\alpha \Lambda^\nu_\beta g_{\mu\nu} = g_{\alpha\beta}.$$

Furthermore, one introduces the Kronecker delta symbol δ^ν_μ, which corresponds to an identity matrix, and the inverse metric $g^{\mu\nu}$ such that

$$g_{\alpha\beta} g^{\alpha\gamma} = \delta^\gamma_\beta.$$

Numerically, $g^{\mu\nu} = \operatorname{diag}(1, -1, -1, -1)$ is the same matrix as $g_{\mu\nu}$.

One also introduces the Levi-Civita symbol $\varepsilon_{\kappa\lambda\mu\nu}$ which is a rank 4 totally antisymmetric tensor, defined by the relation $\varepsilon_{0123} = 1$ in an inertial frame.

Finally, it is often convenient to work with **dual vectors**, also called **covectors** or **covariant vectors**. A covectors can be pictured geometrically as the *slope* of the (multidimensional) graph of a function of x^μ. The slope at a chosen point describes the derivative of a function in every direction at once, in the following sense. The derivative of a function $f(\mathbf{x})$ in the direction given by a vector $\mathbf{v}$ is

$$D_{\mathbf{v}} f \equiv v^\lambda \frac{\partial f}{\partial x^\lambda}.$$

Thus the slope of a function $f(\mathbf{x})$ is adequately described by the collection of components $s_\lambda \equiv \partial f / \partial x^\lambda$. If s_μ is the "slope of a function," then the derivative of a function in a direction v^μ is the number $s_\mu v^\mu$. Thus, covectors can be thought of as linear functions of vectors.

A linear function $\mathbf{s}$ of a vector $\mathbf{v}$ must be a linear combination of the components v^μ, $\mu = 0, 1, 2, 3$, i.e. the function $\mathbf{s}$ must have the form

$$\mathbf{s}(\mathbf{v}) = s_0 v^0 + s_1 v^1 + s_2 v^2 + s_3 v^3 \equiv s_\mu v^\mu.$$

The numbers s_μ are the **components** of the covector $\mathbf{s}$. Thus, in view of the Einstein summation convention, it is natural to denote covectors by letters with lower indices.

A standard example of a covector is the function which maps $\mathbf{x}$ to the scalar product of $\mathbf{x}$ with a fixed vector $\mathbf{b}$:

$$s : \mathbf{x} \to g(\mathbf{b}, \mathbf{x}).$$

(This is the slope of the hyperplane $z = g(\mathbf{b}, \mathbf{x})$, where the coordinate z is the "ordinate axis" of the plot.) It is easy to see that the components of the covector s are $s_\mu = g_{\mu\nu} b^\nu$. It is customary to refer to the relation between the vector $\mathbf{b}$ and the covector s as the "lowering of an index" of b^μ, and to denote the two objects $\mathbf{b}, \mathbf{s}$ by the same letter (rather than by different letters as I have done here). Thus, one writes b_μ instead of s_μ and calls it the "covariant version of the vector b^μ" or the "vector b^μ with the index lowered." In this way, the metric $g_{\mu\nu}$ can be seen as the "operator that lowers indices," while the inverse metric $g^{\mu\nu}$ "raises indices." (In a more geometric language, $g_{\mu\nu}$ is a map from vectors to covectors and $g^{\mu\nu}$ is the inverse map.)

A.3 Transition to General Relativity

Einstein's main motivation for introducing General Relativity was to remove the restriction to inertial frames and to admit arbitrary non-inertial, that is, *accelerated* reference frames. Since gravitation is locally equivalent to an accelerated reference frame (the **equivalence principle**), he hoped to achieve a description of gravitation in this way.

Once we admit arbitrary non-inertial reference frames, we must assume that the coordinates x^μ may be curvilinear, while the metric $g_{\mu\nu}$ may be non-diagonal and generally coordinate-dependent. Thus, the theory of General Relativity is obtained from Special Relativity by the following steps:

(1) We replace the Minkowski spacetime $\mathcal{M} = \mathbb{R}^4$, which is itself a vector space, by an arbitrary four-dimensional space $\mathcal{M}$, called the **spacetime manifold**. Coordinates x^μ in $\mathcal{M}$ are chosen arbitrarily.

(2) The Minkowski metric $\eta_{\mu\nu} = \operatorname{diag}(1, -1, -1, -1)$ is replaced by a symmetric tensor $g_{\mu\nu}(x) = g_{\nu\mu}(x)$ which can depend on the coordinates x^μ. Therefore, we imagine that $\mathcal{M}$ is an arbitrarily curved spacetime rather than a "flat" Minkowski spacetime. (A spacetime manifold is **flat** if there exists a coordinate system in which the metric tensor is $g_{\mu\nu}(p) = \eta_{\mu\nu}$ at every point p. A manifold is **curved** if no such coordinate system can be found.)

(3) The signature of the matrix $g_{\mu\nu}$ must remain $(+---)$ everywhere. In particular, for each event $p \in \mathcal{M}$ there must exist a local coordinate system where $g_{\mu\nu}(p) = \operatorname{diag}(1, -1, -1, -1)$. Moreover, there exists a local coordinate system (called a **locally inertial frame**) where the physical laws at the event p are exactly the same as in Special Relativity in the absence of gravitation. Thus, gravitation is locally equivalent to a non-inertial reference frame (the **equivalence principle**).

(4) Physical laws, which were previously made relativistic (compatible with arbitrary inertial frames and rewritten

through the Minkowski metric), are now reformulated in a way that is independent of the choice of coordinate systems and admits arbitrary, non-inertial coordinates. In particular, in every equation the Minkowski metric is replaced by the tensor $g_{\mu\nu}(x)$. Scalar products and the raising/lowering of tensor indices is performed using $g_{\mu\nu}(x)$.

(5) The theory prescribes equations (the Einstein equations) that can be used to determine the metric $g_{\mu\nu}(x)$ from a given distribution of matter in the entire spacetime. The Einstein equations involve the *curvature* of the spacetime and the energy-momentum tensor of matter (see below).

It follows from (1) that the coordinates x^μ are not themselves 4-vectors any more, and for instance the coordinates x^μ and y^μ of two events cannot be simply subtracted to obtain a 4-vector $y^\mu - x^\mu$. Moreover, the theory must be invariant under arbitrary coordinate transformations,

$$x^\mu \to \tilde{x}^\mu(x^0, x^1, x^2, x^3). \tag{A.2}$$

However, the derivative $\dot{x}^\mu \equiv dx^\mu/d\tau$ actually *is* a 4-vector because it is the difference of coordinates at infinitesimally close points, and such points belong to an almost-Minkowski local environment which exists according to (3). The fact that $u^\mu \equiv dx^\mu/d\tau$ is a 4-vector can be also formally established by considering the change of the components u^μ under the transformation (A.2),

$$u^\mu \to \tilde{u}^\mu \equiv \frac{d}{d\tau}\tilde{x}^\mu = \frac{\partial \tilde{x}^\mu}{\partial x^\nu}\frac{dx^\nu}{d\tau} = \frac{\partial \tilde{x}^\mu}{\partial x^\nu}u^\nu, \tag{A.3}$$

which is the usual linear transformation law for components of a vector under a general change of basis.

Hence, the metric $g_{\mu\nu}(x)$ still determines proper times and proper distances, but only between infinitesimally close events. To make this statement, one frequently writes a somewhat strange-looking equation

$$ds^2 = g_{\mu\nu}dx^\mu dx^\nu,$$

where "dx^μ" means an infinitesimal displacement between two nearby points, and ds^2 is not really a square of any "ds" since ds^2 may also be negative. (One can regard the above equation simply as a **jargon** notation, i.e. a meaningless but convenient and widely understood shorthand for the words "the infinitesimal interval is equal to $g_{\mu\nu}dx^\mu dx^\nu$.") To determine the proper time interval between two widely separated events $x^\mu_{(0)}$ and $x^\mu_{(1)}$, we need to select a line in $\mathcal{M}$ leading from $x^\mu_{(0)}$ and $x^\mu_{(1)}$, say a curve $x^\mu(s)$, where s is some parameter such that $x^\mu(0) = x^\mu_{(0)}$ and $x^\mu(1) = x^\mu_{(1)}$, and integrate along the curve,

$$\Delta\tau \equiv \int_0^1 \sqrt{g_{\mu\nu}\left(x^\mu(s)\right)\frac{dx^\mu}{ds}\frac{dx^\nu}{ds}}\,\mathrm{d}s.$$

Here we assume that curve $x^\mu(s)$ is such that dx^μ/ds is everywhere timelike, so the expression under the square root is positive. Such curves are called **timelike worldlines**. The resulting time interval $\Delta\tau$ will of course depend on the chosen curve $x^\mu(s)$, but not on the choice of the parameter s along the curve. In general, it is not easy to determine whether there exists a timelike curve connecting two given points in the spacetime. It might happen that only a null or a spacelike curve connects two given events. (Two events are **timelike separated** if they can be connected by a timelike worldline.) Without a detailed analysis of the behavior of the metric $g_{\mu\nu}$ everywhere in $\mathcal{M}$, it is not immediately clear whether two given events can causally influence each other.

A.4 Covariant derivative

The requirement (4) above means, in particular, that the relativistic Newton's law (A.1) should be rewritten so that it holds in all coordinate systems. An immediate difficulty with this requirement is that the laws of physics involve derivatives of vectors, and taking a derivative of a vector or a tensor in an arbitrary coordinate system in curved space is a somewhat complicated operation. For instance, $d^2x^\mu/d\tau^2$ is *not* a correct expression for the μ-th component of the 4-acceleration, unless we are using Cartesian coordinates in flat Minkowski spacetime. Let us examine this problem in more detail; we shall begin by considering derivatives in curved coordinates in *flat* space, and then generalize to curved space.

A.4.1 Curved coordinates

Suppose that $x^\mu(\tau)$ is the worldline of a particle in the flat Minkowski spacetime specified in a rectangular coordinate system $\{x^\mu\}$, and suppose that $\{\tilde{x}^\mu\}$ is a curvilinear coordinate system related to $\{x^\mu\}$ by four given functions $f^\mu(x)$, $\mu = 0, 1, 2, 3$, as $\tilde{x}^\mu = f^\mu(x^0, x^1, x^2, x^3)$. The components of 4-velocity in the coordinate system $\{x^\mu\}$ are $u^\mu \equiv dx^\mu/d\tau$, and the components of the 4-acceleration are $a^\mu \equiv d^2x^\mu/d\tau^2$. In the coordinate system $\{\tilde{x}^\mu\}$, the new components of the 4-velocity, $\tilde{u}^\mu$, and the 4-acceleration, $\tilde{a}^\mu$, must be calculated according to Eq. (A.3):

$$\tilde{u}^\mu = \frac{\partial f^\mu}{\partial x^\nu}u^\nu, \quad \tilde{a}^\mu = \frac{\partial f^\mu}{\partial x^\nu}a^\nu = \frac{\partial f^\mu}{\partial x^\nu}\frac{du^\nu}{d\tau}. \tag{A.4}$$

The worldline of the particle in the new coordinate system is $\tilde{x}^\mu(\tau) \equiv f^\mu(x(\tau))$, and so one might try to compute the 4-velocity and the 4-acceleration directly in the coordinate system $\{\tilde{x}^\mu\}$ by computing derivatives $d\tilde{x}^\mu/d\tau$ and $d^2\tilde{x}^\mu/d\tau^2$. Now, the 4-velocity will be computed correctly because, by virtue of the chain rule, $\partial_\alpha f(x) = (\partial f/\partial x)\,\partial_\alpha x$, and hence

$$\frac{d\tilde{x}^\mu}{d\tau} = \frac{\partial f^\mu}{\partial x^\nu}\frac{dx^\nu}{d\tau} = \frac{\partial f^\mu}{\partial x^\nu}u^\nu = \tilde{u}^\mu.$$

However, the 4-acceleration would be found incorrectly if one uses the formula $d\tilde{u}^\mu/d\tau$:

$$\begin{aligned}\frac{d}{d\tau}\tilde{u}^\mu &= \frac{d}{d\tau}\left(\frac{\partial f^\mu}{\partial x^\nu}u^\nu\right) = u^\nu\frac{d}{d\tau}\left(\frac{\partial f^\mu}{\partial x^\nu}\right) + \frac{\partial f^\mu}{\partial x^\nu}\frac{du^\nu}{d\tau} \\ &= u^\nu\frac{d}{d\tau}\left(\frac{\partial f^\mu}{\partial x^\nu}\right) + \tilde{a}^\mu.\end{aligned} \tag{A.5}$$

For a general coordinate transformation (not merely a linear change of coordinates where $\partial f^\mu/\partial x^\nu = \text{const}$), we have

$$\frac{d}{d\tau}\left(\frac{\partial f^\mu}{\partial x^\nu}\right) \neq 0,$$

therefore $\tilde{a}^\mu \neq d\tilde{u}^\mu/d\tau$. So the formula $d\tilde{u}^\mu/d\tau$ cannot be used in a general coordinate system $\{\tilde{x}^\mu\}$ to compute the components of the 4-acceleration $\tilde{a}^\mu$. The reason for the problem is that the coordinate system $\{\tilde{x}^\mu\}$ is curved and so the derivative $d\tilde{u}^\mu/d\tau$ reflects not only the change in the 4-velocity, but also the change in the directions of coordinate axes.

A correct expression for $\tilde{a}^\mu$ is given in Eq. (A.4). Nevertheless, we would like to have a formula for $\tilde{a}^\mu$ that can be used directly in the coordinate system $\{\tilde{x}^\mu\}$. The solution of the problem is evident from Eq. (A.5),

$$\tilde{a}^\mu = \frac{d\tilde{u}^\mu}{d\tau} - u^\nu\frac{d}{d\tau}\left(\frac{\partial f^\mu}{\partial x^\nu}\right).$$

Here, u^ν can be expressed through $\tilde{u}^\mu$ using Eq. (A.4). Therefore, the 4-acceleration $\tilde{a}^\mu$ in a curved coordinate system can be computed from $d\tilde{u}^\mu/d\tau$ and $\tilde{u}^\mu$, if additionally we know the functions $f^\mu(x^\nu)$, $\mu = 0, 1, 2, 3$, that relate the curved coordinates to the flat ones. Note that $d/d\tau$ is effectively a derivative in the direction of the 4-velocity u^μ and can be applied to arbitrary functions of coordinates as

$$\frac{d}{d\tau}(...) = u^\lambda \frac{\partial}{\partial x^\lambda}(...).$$

Since f^μ are functions of *coordinates* (not of τ), we may write

$$\frac{d}{d\tau}\left(\frac{\partial f^\mu}{\partial x^\nu}\right) = u^\lambda \frac{\partial}{\partial x^\lambda}\left(\frac{\partial f^\mu}{\partial x^\nu}\right).$$

Therefore, we can compute the derivative of a vector field $A^\mu(x)$ in the direction u^λ in any coordinate system: in flat coordinates $\{x^\mu\}$ as

$$\frac{d}{d\tau}A^\mu \equiv u^\lambda \frac{\partial A^\mu}{\partial x^\lambda}$$

and in curved coordinates $\{\tilde{x}^\mu\}$ as

$$\tilde{u}^\lambda\left[\frac{\partial \tilde{A}^\mu}{\partial \tilde{x}^\lambda} - A^\nu \frac{\partial^2 f^\mu}{\partial x^\lambda \partial x^\nu}\right] = \tilde{u}^\lambda\left[\frac{\partial \tilde{A}^\mu}{\partial \tilde{x}^\lambda} - \tilde{A}^\rho\left(\frac{\partial x^\nu}{\partial f^\rho}\right)\frac{\partial^2 f^\mu}{\partial x^\lambda \partial x^\nu}\right].$$

The expression in brackets above is called the **covariant derivative** and denoted either by a semicolon or by the symbol ∇ (pronounced "nabla" or "del"),

$$\nabla_\lambda \tilde{A}^\mu \equiv \tilde{A}^\mu{}_{;\lambda} \equiv \frac{\partial \tilde{A}^\mu}{\partial \tilde{x}^\lambda} - \tilde{A}^\rho\left(\frac{\partial x^\nu}{\partial f^\rho}\right)\frac{\partial^2 f^\mu}{\partial x^\lambda \partial x^\nu}. \tag{A.6}$$

In this notation, the derivative of a vector $\tilde{A}^\mu$ in the direction $\tilde{u}^\mu$ is written as $\tilde{u}^\lambda \tilde{A}^\mu{}_{;\lambda}$ or $\tilde{u}^\lambda \nabla_\lambda \tilde{A}^\mu$.

By construction, the covariant derivative $\tilde{A}^\mu{}_{;\lambda}$ is found by first computing the derivative $\partial A^\mu/\partial x^\lambda$ in flat coordinates (where we know the correct way to differentiate vectors) and then recalculating the components to the curved coordinates $\{\tilde{x}^\mu\}$ by using the transformation law for tensors,

$$\tilde{A}^\mu{}_{;\lambda} = \left(\frac{\partial A^\nu}{\partial x^\rho}\right)\frac{\partial f^\mu}{\partial x^\nu}\frac{\partial x^\rho}{\partial f^\lambda}.$$

The formula (A.6) gives these components $\tilde{A}^\mu{}_{;\lambda}$ directly through the components $\tilde{A}^\mu(\tilde{x})$, without need to know the components A^μ in flat coordinates. Of course, the components $\tilde{A}^\mu{}_{;\lambda}$ transform correctly ("covariantly") under a change of coordinates. The origin of the name "covariant" derivative.

Remark: I would like to emphasize that the covariant derivative $\nabla_\lambda A^\mu$ is not a different *kind* of derivative. If the notion of the covariant derivative causes you difficulties, it might be instructive to focus attention on the case of flat space. In flat space, there is already a well-defined, familiar directional derivative of vectors, namely $\partial A^\mu/\partial x^\lambda$, where $\{x^\lambda\}$ is a flat (Cartesian) coordinate system. However, we have seen that the formula $\partial A^\mu/\partial x^\lambda$ only holds in flat coordinates, while in curved coordinates the correct expression for the directional derivative is the formula (A.6). Thus, the "covariant derivative" (A.6) can be thought of as a *covariant formula for* the (already well-defined and familiar) directional derivative of a vector field in flat space. The covariant formula is preferable because it holds in curved coordinates as well as in Cartesian coordinates. ■

Suppose that $\{\tilde{x}^\mu\}$ were actually the same coordinate system as $\{x^\mu\}$; then we would have $\partial^2 f^\mu/(\partial x^\lambda \partial x^\nu) = 0$ and Eq. (A.6) would still define the correct derivative $\partial A^\mu/\partial x^\lambda$. Therefore, the covariant derivative reduces to the coordinate derivative $\partial/\partial x^\lambda$ in flat coordinates, and so it makes sense to use the covariant derivative *always*. A tensor expression involving covariant derivatives, such as

$$a^\mu = u^\nu u^\mu{}_{;\nu},$$

is valid in *every* coordinate system.

A more concise way to represent the covariant derivative (A.6) is to write

$$\tilde{A}^\mu{}_{;\lambda} \equiv \frac{\partial \tilde{A}^\mu}{\partial \tilde{x}^\lambda} + \tilde{A}^\rho \Gamma^\mu_{\rho\lambda}.$$
$$\Gamma^\mu_{\rho\lambda} \equiv -\left(\frac{\partial x^\nu}{\partial f^\rho}\right)\frac{\partial^2 f^\mu}{\partial x^\lambda \partial x^\nu}.$$

Once the quantities $\Gamma^\mu_{\rho\lambda}$ are computed for a given coordinate system $\{\tilde{x}^\mu\}$, it is easy to differentiate vectors directly in that coordinate system, without needing to convert tensors back to flat coordinates every time.

Note that $\Gamma^\mu_{\rho\lambda}$ is an array of numbers that depends on the coordinate system in a nontrivial way and, despite its appearance as an indexed quantity, is *not* a tensor of rank (1,2). The transformation law for $\Gamma^\mu_{\rho\lambda}$ is given in Eq. (A.9) below.

A.4.2 Curved space and induced metric

In the previous section, we found how a vector field $A^\mu(x)$ should be differentiated in an arbitrary, curved coordinate system in flat space. The method was to differentiate A^μ in a flat coordinate system and then recalculate the components to the curved coordinates. The result was a covariant formula for the derivative, $\nabla_\lambda A^\mu$. Let us now consider a more general case: a *curved space*.

A general **manifold** $\mathcal{M}$ can be visualized as a "non-straight" (curved) surface in flat Euclidean space $\mathbb{R}^n$. A direction *within* the manifold is represented by a vector *tangent* to the surface. Since tangent vectors are at the same time vectors in $\mathbb{R}^n$, we can compute scalar products of such tangent vectors by using the standard scalar product in $\mathbb{R}^n$. Thus we have automatically defined a metric on the surface. (A **metric** on a manifold is a quadratic form on tangent vectors.) This naturally defined metric on the surface is called the **induced metric** on the surface with respect to the given embedding in $\mathbb{R}^n$. The idea is that the known scalar product in $\mathbb{R}^n$ "induces" the scalar product of tangent vectors.

Example: A unit 2-sphere in three-dimensional Euclidean space with coordinates $\{x, y, z\}$ is defined as the locus of points satisfying

$$x^2 + y^2 + z^2 = 1.$$

The 2-sphere a curved two-dimensional manifold denoted S^2. A tangent vector to the sphere at a point $\{x, y, z\}$ can be visualized as a three-dimensional vector with components $\{v_1, v_2, v_3\}$ such that

$$xv_1 + yv_2 + zv_3 = 0.$$

(This equation selects the tangent plane at point $\{x, y, z\}$ on the sphere.) The scalar product of two tangent vectors

$\{u_1, u_2, u_3\}$ and $\{v_1, v_2, v_3\}$ is the usual Euclidean product,

$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 + u_3 v_3.$$

Points on the sphere may be labeled by intrinsic coordinates, e.g. by the spherical coordinates $\{\theta, \phi\}$. The point $\{\theta, \phi\}$ has Euclidean coordinates $\{X, Y, Z\}$, where X, Y, Z are functions of θ and ϕ,

$$X = \cos\theta \cos\phi,\ Y = \cos\theta \sin\phi,\ Z = \sin\theta.$$

A tangent vector with three-dimensional components $\{v_1, v_2, v_3\}$ is visualized as a "velocity" of a point moving within the sphere. If this point moves along a trajectory $\gamma(\tau) \equiv \{\Theta(\tau), \Phi(\tau)\}$, where Θ and Φ are some functions, then the velocity vector is described by *two* intrinsic components $\{t_1, t_2\}$,

$$\mathbf{v} = \{t_1, t_2\} = \left\{ \frac{d\Theta}{d\tau}, \frac{d\Phi}{d\tau} \right\}.$$

It is straightforward to see that the three-dimensional components $\{v_1, v_2, v_3\}$ of the same velocity vector $\mathbf{v}$ are found as

$$v_1 = \frac{\partial X}{\partial \theta}\frac{d\Theta}{d\tau} + \frac{\partial X}{\partial \phi}\frac{d\Phi}{d\tau}, \quad \text{etc.}$$

Thus the correspondence between the intrinsic components $\{t_1, t_2\}$ and the three-dimensional components $\{v_1, v_2, v_3\}$ is a linear transformation,

$$v_1 = \frac{\partial X}{\partial \theta} t_1 + \frac{\partial X}{\partial \phi} t_2,\ v_2 = \frac{\partial Y}{\partial \theta} t_1 + \frac{\partial Y}{\partial \phi} t_2,\ v_3 = \frac{\partial Z}{\partial \theta} t_1 + \frac{\partial Z}{\partial \phi} t_2.$$

The scalar product of two tangent vectors $\mathbf{u}$ and $\mathbf{v}$ can be written explicitly through the intrinsic components $\mathbf{u} \equiv \{s_1, s_2\}$ and $\mathbf{v} \equiv \{t_1, t_2\}$ as

$$\begin{aligned} \mathbf{u} \cdot \mathbf{v} &= u_1 v_1 + u_2 v_2 + u_3 v_3 \\ &= \left(\frac{\partial X}{\partial \theta} s_1 + \frac{\partial X}{\partial \phi} s_2 \right) \left(\frac{\partial X}{\partial \theta} s_1 + \frac{\partial X}{\partial \phi} t_2 \right) + \ldots \end{aligned}$$

After a short calculation with the functions X, Y, Z given above, we find

$$\mathbf{u} \cdot \mathbf{v} = s_1 t_1 + \left(\sin^2\theta \right) s_2 t_2.$$

Thus the induced metric on the sphere (in coordinates $\{\theta, \phi\}$) is written as

$$ds^2 = d\theta^2 + d\phi^2 \sin^2\theta.$$

Note that there is no set of coordinates $\{x, y\}$ that would make a sphere a flat manifold, i.e. a space with the Euclidean metric $dx^2 + dy^2$. (If that were possible, there would be a *unique* shortest path connecting opposite poles of the sphere, which is clearly not the case.) ■

A.4.3 Covariant derivative

We now need to derive a formula for the directional derivative for vector fields in a curved space, when a flat coordinate system is *not available*. The formula must involve only intrinsic coordinates on the manifold and require no information about an embedding of $\mathcal{M}$ into a larger space.

It is a known theorem in differential geometry that *any* manifold $\mathcal{M}$ with a given metric g can be represented as a hypersurface embedded into a higher-dimensional flat space (the higher-dimensional space is sometimes called the **bulk** in modern physics). Then the flat coordinates $\{X^\mu\}$ and the flat metric $G_{\mu\nu}$ are available in the bulk, and the metric $g_{\mu\nu}$ defined within the manifold $\mathcal{M}$ is equal to the induced metric with respect to the embedding into the bulk. A vector field defined within the manifold $\mathcal{M}$ is visualized as a field of vectors $A^\mu(x)$ everywhere tangent to the surface. We can compute the derivative of a tangent vector A^μ with respect to a tangent direction u^μ as $u^\lambda \partial A^\mu / \partial X^\lambda$, but this vector may have a nonzero component normal to the surface. If we simply discard this component, i.e. if we project the vector $u^\lambda \partial A^\mu / \partial X^\lambda$ orthogonally onto the surface, we obtain a vector tangent to the surface. Let us denote this vector by $u^\lambda \nabla_\lambda A^\mu$ (this notation is justified since the projection of $u^\lambda \partial A^\mu / \partial X^\lambda$ is linear in u^λ). The tensor $\nabla_\lambda A^\mu$ is called the **covariant derivative** of the vector field A^μ. Thus we have obtained a derivative operation ∇_λ defined on tangent vectors within the manifold $\mathcal{M}$. Similarly, the covariant derivative is defined on arbitrary tensors.

The procedure we employed to define the covariant derivative involves the orthogonal projection of $\partial A^\mu / \partial X^\lambda$ onto the tangent space of the hypersurface $\mathcal{M}$. It may be unclear why it is useful to apply such a projection rather than some other operation. One can motivate this procedure by the following considerations. Imagine that a massive body is constrained to move without friction along a hypersurface $\mathcal{M}$, while no other forces influence its motion. Further, imagine that the motion of the body is observed by a "flat" observer who lives entirely within the hypersurface and thus cannot see the extra dimensions. To the "flat" observer, such a body appears to be unforced because the constraining force normal to the surface is invisible. Therefore, the "flat" observer expects that the acceleration of the body, as seen from within the hypersurface, is equal to zero. Since the force acting on the body is always normal to the hypersurface, indeed the *tangential* components of the acceleration are always zero (but not the normal component). Therefore, the orthogonal projection of the acceleration onto the tangent space is always zero. This condition is equivalent to saying that the covariant derivative of the velocity vector field u^μ in the direction of motion is equal to zero, $u^\lambda \nabla_\lambda u^\mu = 0$. Therefore, the construction of the covariant derivative through the orthogonal projection indeed expresses the idea that the velocity of a freely moving body remains constant along the trajectory.

Note that the normal projection is constructed through the metric $G_{\mu\nu}$ and depends on the embedding of the hypersurface $\mathcal{M}$ into the flat space. However, the derivative operation can be written solely in terms of intrinsic coordinates $\{x^\mu\}$ within the manifold $\mathcal{M}$ and the metric $g_{\mu\nu}$, without referring to any embedding into a larger space. The derivation is standard and somewhat lengthy, so we omit it.[2] We present only the resulting final formula for the covariant derivative,

$$\nabla_\nu u^\mu = \frac{\partial u^\mu}{\partial x^\nu} + \Gamma^\mu_{\alpha\nu} u^\alpha, \tag{A.7}$$

$$\Gamma^\lambda_{\alpha\beta} \equiv \frac{1}{2} g^{\lambda\mu} \left(g_{\mu\alpha,\beta} + g_{\mu\beta,\alpha} - g_{\alpha\beta,\mu} \right). \tag{A.8}$$

The array of numbers $\Gamma^\mu_{\alpha\beta}$ is called the **Christoffel symbol** corresponding to the given coordinate system and the metric $g_{\mu\nu}$. As before, the numbers $\Gamma^\mu_{\alpha\nu}$ depend on the coordinate system

[2] A derivation of an explicit formula the covariant derivative is performed in Sec. 1.6.6 using coordinate-free calculations.

in such a way that the sum $\partial u^\mu/\partial x^\nu + \Gamma^\mu_{\alpha\nu} u^\nu$ transforms correctly as a rank (1,1) tensor under an arbitrary change of coordinates, even though the two terms $\partial_\nu u^\mu$ and $\Gamma^\mu_{\alpha\nu} u^\nu$, taken separately, do not transform correctly. The required transformation law for $\Gamma^\mu_{\alpha\nu}$ is easy to derive. Let us temporarily write $\tilde\nabla$ with a tilde when the covariant derivative is computed in the coordinate system $\{\tilde x^\mu\}$ (this is only for clarity; normally we do not use such a notation). We assume that $\nabla_\nu u^\mu$ obeys the transformation law for rank (1,1) tensors,

$$\nabla_\nu u^\mu = \left(\tilde\nabla_\alpha \tilde u^\beta\right) \frac{\partial x^\mu}{\partial \tilde x^\beta} \frac{\partial \tilde x^\alpha}{\partial x^\nu},$$

and substitute

$$\nabla_\nu u^\mu \equiv \frac{\partial}{\partial x^\nu} u^\mu + \Gamma^\mu_{\alpha\nu} u^\alpha,$$
$$\tilde\nabla_\nu \tilde u^\mu \equiv \frac{\partial}{\partial \tilde x^\nu} \tilde u^\mu + \tilde\Gamma^\mu_{\alpha\nu} \tilde u^\alpha.$$

The result is

$$\Gamma^\lambda_{\alpha\beta} = \tilde\Gamma^\mu_{\nu\rho} \frac{\partial \tilde x^\nu}{\partial x^\alpha} \frac{\partial \tilde x^\rho}{\partial x^\beta} \frac{\partial x^\lambda}{\partial \tilde x^\mu} + \frac{\partial^2 x^\lambda}{\partial \tilde x^\nu \partial \tilde x^\rho} \frac{\partial \tilde x^\nu}{\partial x^\alpha} \frac{\partial \tilde x^\rho}{\partial x^\beta}. \tag{A.9}$$

The above transformation law contains a term that is not proportional to $\Gamma^\lambda_{\alpha\beta}$, so the numbers $\Gamma^\lambda_{\alpha\beta}$ may all vanish in one coordinate system but become nonzero in another. Clearly, the numbers $\Gamma^\lambda_{\alpha\beta}$ cannot represent the components of any tensor.

A quantity with the transformation law (A.9) is called a **connection**.

A.4.4 Properties of covariant derivative

As defined above by Eq. (A.7), the covariant derivative operator ∇_ν produces tensors of rank (1,1) out of vectors. Despite an extra term in the formula (A.7), the derivative satisfies the usual properties, such as

$$\nabla_\lambda (f A^\mu) = f \nabla_\lambda A^\mu + A^\mu \nabla_\lambda f,$$

where f is a scalar function and A^μ is a vector field. This is to be expected since ∇_ν can be thought of as merely the standard derivative operator recalculated in a curved coordinate system. Note that $\nabla_\lambda f \equiv \partial f/\partial x^\lambda$ is an ordinary derivative, but no harm is done by writing $\nabla_\lambda f$ instead of $\partial_\lambda f$.

Since the derivative operation ∇_λ is a projection of the ordinary directional derivative $\partial/\partial X^\lambda$ in the embedding space, it follows that ∇_λ satisfies the standard properties of a derivative, such as linearity,

$$\nabla_\lambda (A^{\mu\nu\ldots}_{\alpha\beta\ldots} + B^{\mu\nu\ldots}_{\alpha\beta\ldots}) = \nabla_\lambda A^{\mu\nu\ldots}_{\alpha\beta\ldots} + \nabla_\lambda B^{\mu\nu\ldots}_{\alpha\beta\ldots},$$

and the Leibnitz rule, for instance

$$\nabla_\lambda (a^\alpha_\mu b^{\beta\gamma}) = \left(\nabla_\lambda a^\alpha_\mu\right) b^{\beta\gamma} + a^\alpha_\mu \left(\nabla_\lambda b^{\beta\gamma}\right).$$

Using this property, a generalization of the formula (A.7) can be easily found for covectors or arbitrary tensors. In particular, we demand that

$$\nabla_\nu \left(a^\alpha b^\beta\right) = \left(\nabla_\nu a^\alpha\right) b^\beta + a^\alpha \left(\nabla_\nu b^\beta\right),$$
$$\nabla_\nu \left(x^\alpha y_\alpha\right) = \partial_\nu \left(x^\alpha y_\alpha\right).$$

From this we can derive an explicit formula for the covariant derivative for an arbitrary tensor, in terms of coordinate derivatives ∂_λ and the Christoffel symbol. For example,

$$\nabla_\nu a_\mu = \partial_\nu a_\mu - \Gamma^\lambda_{\nu\mu} a_\lambda,$$
$$\nabla_\nu T^\alpha_\beta = \partial_\nu T^\alpha_\beta + \Gamma^\alpha_{\mu\nu} T^\mu_\beta - \Gamma^\lambda_{\nu\beta} T^\alpha_\lambda.$$

For brevity, partial derivatives with respect to a coordinate are often denoted by an index with a preceding comma,

$$\frac{\partial u^\mu}{\partial x^\nu} \equiv u^\mu_{,\nu},$$

while covariant derivatives are denoted by an index with a preceding semicolon, $\nabla_\nu u^\mu \equiv u^\mu_{;\nu}$.

All the physical laws can be formulated in a **generally covariant** way (i.e., valid in any coordinate system) if we replace all coordinate derivatives ∂_μ by covariant derivatives ∇_μ. Thus, the generally covariant analog of Eq. (A.1) is

$$m u^\nu \nabla_\nu u^\mu = f^\mu. \tag{A.10}$$

Here $u^\mu \equiv dx^\mu/d\tau$ is the 4-velocity of the particle, and we replaced $d/d\tau$ by the covariant derivative $u^\nu \nabla_\nu$. So far, all we have achieved is a rewriting of the known physical laws in an arbitrary coordinate system; no new physical information is introduced or derived. The second step towards General Relativity is to postulate that the correct form of the physical equations (e.g., Newton's law or the Maxwell equations) is obtained from the known laws in Special Relativity by substituting an arbitrary (but nondegenerate and Lorentzian-signature) spacetime-dependent metric $g_{\mu\nu}(x)$ instead of the flat Minkowski metric $\eta_{\mu\nu}$ and by replacing derivatives ∂_μ by covariant derivatives ∇_μ, where the Christoffel symbol $\Gamma^\lambda_{\alpha\beta}$ is defined by Eq. (A.8). This step of course leads to certain changes in the physical laws due to a different $g_{\mu\nu}$ and a nontrivial $\Gamma^\lambda_{\alpha\beta}$; these changes are interpreted as the influence of gravitation. The result is a reformulation of all the laws of physics in an arbitrary curved spacetime. This is, of course, a major assumption about the way gravitation affects the behavior of particles and fields: "gravitation" is not a special force but the natural and inavoidable influence of the nontrivial geometry on matter in a curved spacetime. Ultimately, this statement must be tested by experiments. The experimental status of General Relativity is quite satisfactory at present. The widespread acceptance of GR is due to the experimental confirmation as well as to the simplicity of this theory compared with other theories of gravitation.

A.4.5 *Choice of connection

We have shown that the transformation law (A.9) can be derived merely from the requirement that $\nabla_\nu A^\mu$ should transform as a tensor of rank (1,1). It follows that *any* connection $\Gamma^\lambda_{\alpha\beta}$ that transforms according to Eq. (A.9) will give rise to a tensor

$$\frac{\partial A^\mu}{\partial x^\beta} + \Gamma^\mu_{\alpha\beta} A^\alpha.$$

So one may consider "alternative" covariant derivatives of the form (A.7), where $\Gamma^\mu_{\alpha\beta}$ is not necessarily the Christoffel connection. Since the physical influence of gravity is described by terms containing $\Gamma^\mu_{\alpha\beta}$, a natural question is to decide which connection $\Gamma^\mu_{\alpha\beta}$ is "correct."

In GR, one chooses the Christoffel symbol (A.8) as the connection. This can be justified on the basis of experimental data confirming GR, and also by considerations of mathematical simplicity. In fact, the choice (A.8) is equivalent to the following two conditions: (i) $\Gamma^\lambda_{\alpha\beta} = \Gamma^\lambda_{\beta\alpha}$, and (ii) $g_{\alpha\beta;\nu} = 0$. Let us review various arguments supporting these assumptions.[3]

The condition (i) can be seen as a consequence of the equivalence principle, which says that each event p admits a locally inertial frame where Special Relativity holds in an infinitesimal neighborhood of p. In other words, there exists a coordinate system x^μ in which the physical laws hold with ordinary derivatives $\partial/\partial x^\mu$ instead of ∇_μ. Hence, at the event p we have $\Gamma^\lambda_{\alpha\beta}(p) = 0$. Recalculating $\Gamma^\lambda_{\alpha\beta}$ for other reference frames using Eq. (A.9), we find that $\Gamma^\lambda_{\alpha\beta}(p) = \Gamma^\lambda_{\beta\alpha}(p)$ in *every* reference frame. Since this argument holds for any event p, we find that the equivalence principle entails the condition $\Gamma^\lambda_{\alpha\beta} = \Gamma^\lambda_{\beta\alpha}$. Alternatively, we may demand that the covariant derivatives commute on functions, i.e. $f_{;\mu\nu} = f_{;\nu\mu}$, where f is any scalar function. It can be easily seen that this condition also forces $\Gamma^\lambda_{\alpha\beta} = \Gamma^\lambda_{\beta\alpha}$.

The condition (ii) is another fundamental assumption that can be motivated in many ways. For instance, $g_{\mu\nu}$ in the Minkowski spacetime is given by a constant matrix, $g_{\mu\nu} = \eta_{\mu\nu}$, thus $\partial_\alpha g_{\mu\nu} = 0$, while we expect that the same property (but involving the covariant derivative) should hold in curved spacetimes. Alternatively, we may require that the operation of lowering an index of a vector or a tensor should commute with the covariant derivative:

$$g_{\mu\alpha}A^\mu{}_{;\nu} = (A_\alpha)_{;\nu}, \quad \text{where } A_\alpha \equiv A^\mu g_{\mu\alpha}.$$

This property is clearly equivalent to $g^{\mu\alpha}_{;\nu} = 0$. Alternatively, let us consider the properties of a "locally constant" vector field. A vector field u^μ is "locally constant" if its covariant derivative vanishes at a point p, i.e. $u^\mu_{;\nu}(p) = 0$. (In a locally inertial frame at p, this condition is identical to $\partial_\nu u^\mu(p) = 0$.) If the vector u^μ is "constant" in this sense, it is natural to expect that the length of the vector u^μ is also locally constant: $\nabla_\alpha(g_{\mu\nu}u^\mu u^\nu)|_p = 0$. However, this condition is equivalent to $g_{\mu\nu;\alpha}u^\mu u^\nu = 0$. Since the direction u^μ and the point p are arbitrary, we must require that $g_{\mu\nu;\alpha} = 0$ everywhere.

It is a straightforward exercise to derive an explicit form of $\Gamma^\lambda_{\alpha\beta}$ from the conditions (i) and (ii), and we omit the derivation (which can be found in standard textbooks). The result is the formula (A.8). Thus the intrinsic metric $g_{\mu\nu}$ uniquely defines the Christoffel symbol and the covariant derivative. In General Relativity, one never uses any other connection than the Christoffel symbol given above.

Remark: The constancy of the metric under the covariant derivative, $\nabla_\lambda g_{\mu\nu} = 0$, is a direct consequence of the construction of ∇_λ through a projection from an embedding of the manifold $\mathcal{M}$ as a hypersurface in a "bulk" space with coordinates X^λ. Since the bulk has a flat metric $G_{\mu\nu}$, we have $\partial_\lambda G_{\mu\nu} = 0$, where $\partial_\lambda \equiv \partial/\partial X^\lambda$ is the derivative with respect to the flat coodinates in the bulk. Suppose u^μ and v^μ are two tangent vector fields; then by definition of the induced metric, $g_{\mu\nu}u^\mu v^\nu = G_{\mu\nu}u^\mu v^\nu$. Hence,

$$\partial_\lambda\left(G_{\mu\nu}u^\mu v^\nu\right) = G_{\mu\nu}\left(\partial_\lambda u^\mu\right)v^\nu + G_{\mu\nu}u^\mu\left(\partial_\lambda v^\nu\right).$$

Since u^μ and v^μ are tangent, the normal component of $\partial_\lambda u^\mu$ will disappear from the scalar product $G_{\mu\nu}\left(\partial_\lambda u^\mu\right)v^\nu$. Thus

$$\partial_\lambda\left(G_{\mu\nu}u^\mu v^\nu\right) = g_{\mu\nu}\left(\nabla_\lambda u^\mu\right)v^\nu + g_{\mu\nu}u^\mu\left(\nabla_\lambda v^\nu\right).$$

On the other hand,

$$\partial_\lambda\left(G_{\mu\nu}u^\mu v^\nu\right) = \nabla_\lambda\left(g_{\mu\nu}u^\mu v^\nu\right),$$

and it follows that $\nabla_\lambda g_{\mu\nu} = 0$.

A.5 Curvature

A.5.1 Parallel transport

The concept of parallel transport of vectors on a curved manifold $\mathcal{M}$ can be easily visualized if we think of the manifold $\mathcal{M}$ as a surface embedded in flat space. Suppose a path is given within the surface, with a tangent vector v^μ. An arbitrary tangent vector u^μ can be transported along the path in such a way that the derivative along the path in the bulk space, $v^\lambda\partial u^\mu/\partial X^\lambda$, is everywhere normal to the surface. Since (by definition) ∇_λ is $\partial/\partial X^\lambda$ projected onto the tangent plane, the condition $v^\lambda\partial u^\mu/\partial X^\lambda = 0$ is equivalent to the condition $v^\lambda\nabla_\lambda u^\mu = 0$.

This motivates the following definition: A vector u^μ is **parallelly transported** along a path if $v^\lambda\nabla_\lambda u^\mu = 0$, where v^λ is a tangent vector to the path.

The covariant derivative can then be given a different interpretation using the parallel transport operation. Let us denote by $\hat{T}_{\varepsilon\mathbf{a}}u^\mu$ the parallel transport of a vector u^μ along a straight line segment εa^μ. Then we can say that the covariant derivative measures the rate of change in the vector u^μ compared with the change due to the parallel transport:

$$a^\lambda\nabla_\lambda u^\mu = \lim_{\varepsilon\to 0}\frac{u(x^\mu + \varepsilon a^\mu) - \hat{T}_{\varepsilon\mathbf{a}}u(x^\mu)}{\varepsilon}.$$

It is easy to see that the scalar product of vectors is constant under a parallel transport:

$$v^\lambda\nabla_\lambda\left(g_{\mu\nu}u^\mu w^\nu\right) = 0 \text{ if } v^\lambda\nabla_\lambda u^\mu = v^\lambda\nabla_\lambda w^\mu = 0.$$

Note that in flat space we may introduce flat coordinates where we have $\nabla_\lambda = \partial_\lambda$, and thus a vector u^μ is parallelly transported along any curve iff all the components u^μ remain constant. If we use *curved* coordinates in flat space, the components of a vector may change during a parallel transport because the directions of the basis vectors change in space. However, if we execute a parallel transport of a vector along a *closed* path, the components of the vector will return to their initial values.

Statement: Show that the operator $a^\mu\nabla_\mu$ (acting on an arbitrary vector field u^λ) commutes with parallel transport along a^μ. In other words,

$$a^\mu\nabla_\mu\left(T_{\varepsilon\mathbf{a}}u^\lambda\right) = T_{\varepsilon\mathbf{a}}\left(a^\mu\nabla_\mu u^\lambda\right).$$

[3] I wish to emphasize that these are **physical assumptions**, i.e. something to be ultimately tested by experiments. Theories similar to GR but with $\Gamma^\lambda_{\alpha\beta} \neq \Gamma^\lambda_{\beta\alpha}$ or $g_{\alpha\beta;\nu} \neq 0$ can be constructed as well. Although such theories are of course more complicated than GR, physicists do consider their physical consequences and test them experimentally. So far, no alternative theory of gravitation has been shown to surpass GR in correct experimental predictions.

A.5.2 Riemann tensor

In flat space, the operation of parallel transport along a closed path does not change any vectors. However, in a curved manifold, this is no longer true; e.g., on a spherical surface, a parallel transport of a vector along a spherical triangle will generally rotate the vector. The **curvature** of a manifold measures the failure of parallel transport around a closed curve to preserve vectors.

In general, a vector u^μ parallelly transported along a closed path γ will undergo a linear transformation (a rotation and a dilation). We may describe this transformation by a matrix $\hat{T}^\mu_\nu(\gamma)$, so that the new vector is expressed as $\tilde{u}^\mu = \hat{T}^\mu_\nu(\gamma)u^\nu$. In flat space, parallel transport along a closed path will not change any vectors (even in curved coordinates!) and thus $\hat{T}^\mu_\nu(\gamma) \equiv \delta^\mu_\nu$. In the general case, the dependence of $\hat{T}^\mu_\nu(\gamma)$ on the path γ is complicated but can be simplified if we consider *infinitesimally small* closed paths. A simple example of such a path is a parallelogram spanned by two vectors a^μ and b^μ with sides εa^μ and εb^μ. One can compute the change in the vector u^μ after a parallel transport along the parallelogram (see the calculation below). The result is

$$\tilde{u}^\mu = u^\mu + \varepsilon^2 R^\mu_{\lambda\alpha\beta} u^\lambda a^\alpha b^\beta,$$

where $R^\mu_{\alpha\beta\lambda}$ is called the **Riemann tensor** (also called the **curvature tensor**), which is given by the formula

$$R^\mu_{\lambda\alpha\beta} = \partial_\beta \Gamma^\mu_{\nu\alpha} - \partial_\alpha \Gamma^\mu_{\nu\beta} + \Gamma^\mu_{\beta\lambda}\Gamma^\lambda_{\alpha\nu} - \Gamma^\mu_{\alpha\lambda}\Gamma^\lambda_{\nu\beta} \tag{A.11}$$

Note that the Christoffel symbol involves first derivatives of the metric $g_{\mu\nu}$, and thus the Riemann tensor involves second derivatives of the metric.

One also defines the **Ricci tensor**, $R_{\mu\nu} \equiv R^\alpha_{\mu\alpha\nu}$, and the **curvature scalar** $R \equiv g^{\mu\nu}R_{\mu\nu}$. These quantities are convenient for writing the equation for the gravitational field (the Einstein equation).

A.5.3 *Expressing Riemann tensor through $\Gamma^\lambda_{\alpha\beta}$

In this section we give a detailed derivation of Eq. (A.11).

We will compute the matrix $\hat{T}^\mu_\nu$ that performs the parallel transport of a vector u^μ along a closed parallelogram with vertices $x^\mu_{(0)}$, $x^\mu_{(1)} \equiv x^\mu_{(0)} + \varepsilon a^\mu$, $x^\mu_{(2)} \equiv x^\mu_{(0)} + \varepsilon(a^\mu + b^\mu)$, and $x^\mu_{(3)} \equiv x^\mu_{(0)} + \varepsilon b^\mu$, where $x^\mu_{(0)}$ is an arbitrary point in spacetime and a^μ, b^μ are arbitrary vectors. The matrix $\hat{T}^\mu_\nu$ describes the change in vectors as a result of the parallel transport as $\tilde{u}^\mu = \hat{T}^\mu_\nu u^\nu$.

The calculations are performed in an arbitrary coordinate system covering the entire parallelogram. For brevity, we omit the indices on coordinates of points and write simply $x_{(0)}$, $x_{(0)} + \varepsilon a$, etc., instead of $x^\mu_{(0)}$, $x^\mu_{(0)} + \varepsilon a^\mu$, etc. Let us denote by $T^\mu_\nu(x; \varepsilon a)$ the parallel transport matrix between points x and $x + \varepsilon a$. Then the total transport matrix will be expressed as the (matrix) product of the transport matrices computed for the four individual segments:

$$\hat{T}^\mu_\nu = T^\mu_\alpha(x_{(3)}; -\varepsilon b) T^\alpha_\beta(x_{(2)}; -\varepsilon a) T^\beta_\gamma(x_{(1)}; \varepsilon b) T^\gamma_\nu(x_{(0)}; \varepsilon a).$$

Since we are only interested in considering an infinitesimally small parallelogram, the required value of ε will be small, so we need to find T^μ_ν only up to a certain order in ε. In the course of the calculation it will become clear that only terms up to and including ε^2 are relevant.

We begin by considering the parallel transport along the segment $x_{(0)} \to x_{(0)} + \varepsilon a$. The parallel-transported vector $T^\mu_\nu(x_{(0)}; \varepsilon a)u^\nu$ can be thought of as a function of ε, which we temporarily denote by $u^\mu(\varepsilon)$. By construction, the derivative $a^\mu \partial_\mu$ is equal to $d/d\varepsilon$, and therefore the vector $u^\mu(\varepsilon)$ satisfies the differential equation

$$0 = a^\mu \nabla_\mu u^\nu = \frac{du^\nu(\varepsilon)}{d\varepsilon} + a^\mu u^\alpha(\varepsilon)\Gamma^\nu_{\alpha\mu}(x_{(0)} + \varepsilon a).$$

Here we explicitly indicated the point, $x_{(0)} + \varepsilon a$, where the Christoffel symbol Γ is evaluated. Hence, the parallel transport operator $T^\mu_\nu(x_{(0)}; \varepsilon a)$ satisfies the equation

$$\frac{d}{d\varepsilon} T^\mu_\nu(x_{(0)}; \varepsilon a) + a^\lambda \Gamma^\mu_{\alpha\lambda}(x_{(0)} + \varepsilon a) T^\alpha_\nu(x_{(0)}; \varepsilon a) = 0, \tag{A.12}$$

with the initial condition $T^\mu_\nu\Big|_{\varepsilon=0} = \delta^\mu_\nu$.

It is easy to find an approximate solution of Eq. (A.12) to first order in ε,

$$T^\mu_\nu(x_{(0)}; \varepsilon a) = \delta^\mu_\nu - \varepsilon a^\lambda \Gamma^\mu_{\nu\lambda}(x_{(0)}) + O(\varepsilon^2), \tag{A.13}$$

where $\Gamma^\mu_{\nu\lambda}$ may be computed at $x_{(0)}$ since the variation of Γ across the parallelogram is of order ε and we are disregarding ε^2. The other transport operators ($T^\mu_\nu(x_{(1)}; \varepsilon b)$, etc.) are found by the same method. Combining the four operators, we obtain

$$\begin{aligned} T^\mu_\nu &= \delta^\mu_\nu - \varepsilon \left(a^\lambda + b^\lambda - a^\lambda - b^\lambda \right) \Gamma^\mu_{\nu\lambda}(x_{(0)}) + O(\varepsilon^2) \\ &= \delta^\mu_\nu + O(\varepsilon^2). \end{aligned}$$

It is clear that the current precision is insufficient to describe the change in vectors due to parallel transport. Thus, we need to compute second-order terms.

Now it is necessary to take into account the variation of Γ across the parallelogram. To this end, we expand Γ in Eq. (A.12) as

$$\Gamma^\mu_{\alpha\lambda}(x_{(0)} + \varepsilon a) = \Gamma^\mu_{\alpha\lambda}(x_{(0)}) + \varepsilon a^\beta \partial_\beta \Gamma^\mu_{\alpha\lambda}(x_{(0)}). \tag{A.14}$$

Since Γ is always multiplied by ε, it is sufficient to retain terms of order ε in Eq. (A.14). Let us also compress the notation by introducing the matrix $\Gamma(a) \equiv \Gamma^\mu_{\nu\lambda} a^\lambda$ and omitting the indices in matrix multiplications. Then Eq. (A.12) with the substitution (A.14) becomes

$$\frac{d}{d\varepsilon} T + \Big(\Gamma(a) + \varepsilon a^\beta \partial_\beta \Gamma(a) \Big)\Big|_{x_{(0)}} T + O(\varepsilon^2) = 0, \tag{A.15}$$

while the first-order solution (A.13) is written as

$$T(x_{(0)}; \varepsilon a) = \hat{1} - \varepsilon \Gamma(a) + O(\varepsilon^2),$$

where $\hat{1} \equiv \delta^\mu_\nu$ is the identity matrix. For the present purposes, it suffices to obtain a solution of Eq. (A.15) to second order in ε. Therefore we consider the ansatz

$$T(x_{(0)}; \varepsilon a) = \hat{1} - \varepsilon \Gamma(a)|_{x_{(0)}} + \varepsilon^2 U, \tag{A.16}$$

where the unknown matrix $U \equiv U^\mu_\nu$ is to be found, while the first-order term is copied from Eq. (A.13). Substituting Eq. (A.16) into Eq. (A.15), we find

$$U^\mu_\nu = \frac{1}{2} \Big(\Gamma(a)\Gamma(a) - a^\lambda \partial_\lambda \Gamma(a) \Big)\Big|_{x_{(0)}}.$$

The second-order solution is then

$$T(x_{(0)};\varepsilon a) = \hat{1} + \left[-\varepsilon\Gamma(a) - \frac{\varepsilon^2}{2}\left(\Gamma(a)\Gamma(a) - a^\lambda\partial_\lambda\Gamma(a)\right)\right]_{x_{(0)}}.$$

Note that the second-order terms,

$$\Gamma(b)\Gamma(b) - b^\lambda\partial_\lambda\Gamma(b),$$

may be evaluated at any point, say at $x_{(0)}$, because the variation of Γ across the parallelogram is of order ε. Therefore, we may simplify the above expression to

$$T(x_{(0)};\varepsilon a) = \hat{1} - \varepsilon\Gamma(a)|_{x_{(0)}} + \frac{\varepsilon^2}{2}\left(\Gamma(a)\Gamma(a) - a^\lambda\partial_\lambda\Gamma(a)\right).$$

By the same method, we compute the matrix describing parallel transport from $x_{(1)}$ to $x_{(2)}$,

$$T(x_{(1)};\varepsilon b) = \hat{1} - \varepsilon\Gamma(b)|_{x_{(1)}} + \frac{\varepsilon^2}{2}\left(\Gamma(b)\Gamma(b) - b^\lambda\partial_\lambda\Gamma(b)\right),$$

and all the other parallel transport matrices.

Finally, we have to multiply the four transport matrices, preserving the order of matrix multiplications and discarding any terms of order ε^3 or higher. Within first-order terms, we need to expand Γ to first order in ε, e.g.

$$\Gamma(b)|_{x_{(1)}} = \Gamma(b)|_{x_{(0)}} + \varepsilon a^\lambda\,\partial_\lambda\Gamma(b)|_{x_{(0)}},$$

so that

$$\Gamma(a)|_{x_{(2)}} - \Gamma(a)|_{x_{(0)}} = \varepsilon\left(a^\lambda + b^\lambda\right)\partial_\lambda\Gamma(a)|_{x_{(0)}} + O(\varepsilon^2),$$
$$\Gamma(b)|_{x_{(3)}} - \Gamma(b)|_{x_{(1)}} = \varepsilon\left(-a^\lambda + b^\lambda\right)\partial_\lambda\Gamma(b)|_{x_{(0)}} + O(\varepsilon^2).$$

Within second-order terms, we may evaluate all Γ's at $x_{(0)}$. After a straightforward calculation, we obtain

$$\begin{aligned}\hat{T} &= T(x_{(3)};-\varepsilon b)T(x_{(2)};-\varepsilon a)T(x_{(1)};\varepsilon b)T(x_{(0)};\varepsilon a)\\ &= \hat{1} - \varepsilon\Gamma(a)|_{x_{(0)}} - \varepsilon\Gamma(b)|_{x_{(1)}} + \varepsilon\Gamma(a)|_{x_{(2)}} + \varepsilon\Gamma(b)|_{x_{(3)}}\\ &\quad +\varepsilon^2\big(\Gamma(b)\Gamma(a) - \Gamma(a)\Gamma(b) - a^\lambda\partial_\lambda\Gamma(a) - b^\lambda\partial_\lambda\Gamma(b)\big)\\ &= \hat{1} + \varepsilon^2\big(\Gamma(b)\Gamma(a) - \Gamma(a)\Gamma(b) + b^\lambda\partial_\lambda\Gamma(a) - a^\lambda\partial_\lambda\Gamma(b)\big).\end{aligned}$$

Restoring the full index notation, we can write the result as

$$\hat{T}^\mu_\nu = \delta^\mu_\nu + \varepsilon^2 R^\mu_{\ \nu\alpha\beta}a^\alpha b^\beta,$$
$$R^\mu_{\ \nu\alpha\beta} \equiv \partial_\beta\Gamma^\mu_{\nu\alpha} - \partial_\alpha\Gamma^\mu_{\nu\beta} + \Gamma^\mu_{\beta\lambda}\Gamma^\lambda_{\alpha\nu} - \Gamma^\mu_{\alpha\lambda}\Gamma^\lambda_{\nu\beta}.$$

The last line coincides with Eq. (A.11). This calculation shows that the Riemann tensor $R^\mu_{\ \nu\alpha\beta}$ describes the effect of an infinitesimal parallel transport along a closed curve on vectors.

A.6 Covariant integration

A.6.1 Determinant of the metric

An important object in GR is the determinant of the metric, usually denoted by g,

$$g \equiv \det\left(g_{\mu\nu}\right).$$

It follows that the determinant of the contravariant metric is $g^{-1} = \det\left(g^{\mu\nu}\right)$. Note that the function g is a determinant of a (0,2)-tensor $g_{\mu\nu}$ which not a scalar function but depends nontrivially on the coordinate system. For instance, given two coordinate systems $\{x^\mu\}$ and $\{\tilde{x}^\mu\}$, the corresponding components $g_{\mu\nu}$ and $\tilde{g}_{\mu\nu}$ of the metric are related by

$$g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu = \tilde{g}_{\mu\nu}\mathrm{d}\tilde{x}^\mu\mathrm{d}\tilde{x}^\nu,$$

which can be rewritten in the matrix notation (temporarily using capital letters to denote the matrices $G \equiv g_{\mu\nu}$ and $S \equiv \partial\tilde{x}^\mu/\partial x^\nu$) as

$$G = S^T\tilde{G}S.$$

Since $\det A^T = \det A$ for any matrix A, we find

$$\frac{\det g_{\mu\nu}}{\det\tilde{g}_{\mu\nu}} = \left(\det\frac{\partial\tilde{x}^\mu}{\partial x^\nu}\right)^2. \tag{A.17}$$

It is clear that the metric determinant is a coordinate system-dependent quantity.

Note that a physical metric has the signature $(+---)$, so its determinant is always *negative*.

A.6.2 Covariant volume element

The determinant of the metric is used most frequently to express integration over the manifold in a covariant way, such that functions can be integrated in arbitrary coordinate systems.

Suppose we would like to integrate a scalar function $f(x)$ over a region of the spacetime manifold. The ordinary formula $\int d^4x\, f(x)$ is good in flat (Cartesian) coordinates but unsatisfactory in curved coordinates. Namely, a different choice of coordinates, $\{\tilde{x}^\mu\}$, will generally yield a different result,

$$\int \mathrm{d}^4\tilde{x}\, f(\tilde{x}) = \int \mathrm{d}^4x\, f(x)\det\left[\frac{\partial\tilde{x}^\mu}{\partial x^\nu}\right] \neq \int \mathrm{d}^4x\, f(x),$$

unless the Jacobian happens to be equal to 1, $\det\left[\partial\tilde{x}^\mu/\partial x^\nu\right] = 1$, which is certainly a very special case. The problem is that we have a formula, $\int \mathrm{d}^4x\, f(x)$, that works only in flat coordinates. The solution is to write all the integrals using the **covariant volume element**,

$$\mathrm{d}^4V \equiv \mathrm{d}^4x\sqrt{-g(x)},$$

instead of d^4x. Because of the transformation law (A.17), it is clear that the covariant formula for the integral of f,

$$\int \mathrm{d}^4x\sqrt{-g(x)}f(x),$$

yields the same answer in every coordinate system, including flat coordinate systems where $\sqrt{-g(x)} \equiv 1$ (if such coordinate systems exist).

A.6.3 Derivative of the determinant

It is often necessary to compute derivatives of the metric determinant, for instance $\partial_\mu\sqrt{-g}$. Such derivatives can be easily expressed through $g_{\mu\nu}$ and g.

In order to derive this expression, we use a general formula for the derivative of the determinant of a (nondegenerate) matrix:

$$\frac{d}{dt}\left[\det A(t)\right] = \mathrm{Tr}\left(A^{-1}\frac{d}{dt}A(t)\right)\det A(t). \tag{A.18}$$

This formula can be found quickly from the matrix identity

$$\det e^X = e^{\operatorname{Tr} X},$$

where X is an arbitrary matrix, after a substitution $A \equiv e^X$.

Using Eq. (A.18), we find, in particular,

$$\partial_\mu g = \left[g^{\alpha\beta}\partial_\mu g_{\alpha\beta}\right] g,$$
$$\partial_\mu \sqrt{-g} = \left[g^{\alpha\beta}\partial_\mu g_{\alpha\beta}\right] \frac{\sqrt{-g}}{2}. \tag{A.19}$$

A.6.4 Covariant divergence

Divergence of a vector field,

$$\operatorname{div} \mathbf{a} \equiv \nabla_\mu a^\mu \equiv a^\mu{}_{;\mu},$$

is an important special case where the covariant formula for the derivative is simpler than in the general form.

Using Eqs. (A.7)-(A.8), we find

$$\nabla_\mu a^\mu = \partial_\mu a^\mu + \Gamma^\mu_{\mu\nu} a^\nu;$$
$$\Gamma^\mu_{\mu\nu} = \frac{1}{2} g^{\mu\alpha}\left(g_{\alpha\mu,\nu} + g_{\alpha\nu,\mu} - g_{\mu\nu,\alpha}\right) = \frac{1}{2} g^{\mu\alpha} g_{\mu\alpha,\nu}.$$

Comparing the right-hand side in the last line with Eq. (A.19), we obtain

$$\Gamma^\mu_{\mu\nu} = \frac{1}{\sqrt{-g}}\partial_\nu \sqrt{-g}$$

and hence

$$\nabla_\mu a^\mu = \partial_\mu a^\mu + \frac{1}{\sqrt{-g}} a^\nu \partial_\mu \sqrt{-g} = \frac{1}{\sqrt{-g}} \partial_\nu \left[\sqrt{-g} a^\nu\right]. \tag{A.20}$$

This simple formula is useful because it does not require computing the full set of Christoffel symbols.

A.6.5 Integration by parts

In field theory, one often needs to integrate by parts using the Gauss theorem for volume integrals, for instance,

$$\int_V \mathrm{d}^4x\, \partial_\mu u^\mu = \oint_\Sigma \mathrm{d}^3 A_\mu u^\mu, \tag{A.21}$$

where V is a 4-dimensional region where a vector field u^μ is defined, Σ is the boundary 3-surface of V, and $\mathrm{d}^3 A_\mu$ is a directed 3-area element of the hypersurface Σ. However, the identity (A.21) is valid only if we use the same fixed coordinate system at both sides of the equation. We would like to have a "covariant" formula such that both sides of the equation contain only "covariant" quantities that remain the same in any coordinate system.

To derive such a formula, we guess (heuristically) that the ordinary derivative $\partial_\mu u^\mu$ must be replaced by the covariant derivative $\nabla_\mu u^\mu$, and the volume element d^4x by the covariant volume element $\mathrm{d}^4x\sqrt{-g}$. In flat space, these replacements do not modify the original expressions. Let us now verify that this replacement indeed yields the correct results.

The left-hand side of Eq. (A.21) is replaced by

$$\int_V \mathrm{d}^4x\, \partial_\mu u^\mu \to \int_V \mathrm{d}^4x \sqrt{-g}\, \nabla_\mu u^\mu.$$

This expression stays the same in every coordinate system. Using Eq. (A.20), we find

$$\int_V \mathrm{d}^4x \sqrt{-g}\, \nabla_\mu u^\mu = \int_V \mathrm{d}^4x\, \partial_\mu \left[\sqrt{-g} u^\mu\right].$$

The above equation is valid in any coordinate system; choosing some coordinate system, we may use Eq. (A.21) and obtain (still in the same fixed coordinate system)

$$\int_V \mathrm{d}^4x\, \partial_\mu \left[\sqrt{-g} u^\mu\right] = \oint_\Sigma \mathrm{d}^3 A_\mu \sqrt{-g} u^\mu.$$

The expression on the right-hand side is interpreted as the integral of u^μ over the covariant area element $\sqrt{-g}\mathrm{d}^3 A_\mu$. This expression is the same in every coordinate system. Therefore, the "covariant Gauss theorem" is written as

$$\int_V \mathrm{d}^4x \sqrt{-g}\, \nabla_\mu u^\mu = \oint_\Sigma \mathrm{d}^3 A_\mu \sqrt{-g} u^\mu.$$

A.7 Einstein's equation

We have seen the equation of motion (A.10) of a point mass in a curved spacetime. The influence of gravity on a massive body is described by the presence of the covariant derivative in the equation of motion, rather than by a "force" of gravity appearing in the right-hand side. When solving Eq. (A.10), the metric $g_{\mu\nu}$ (and thus the Christoffel symbol $\Gamma^\lambda_{\alpha\beta}$) are considered known.

Einstein's equation describes the influence of matter on the metric $g_{\mu\nu}$:

$$R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} = -8\pi G T_{\mu\nu},$$

where $R_{\mu\nu}$ is the Ricci tensor, R is the curvature scalar, G is Newton's constant, and $T_{\mu\nu}$ is the combined energy-momentum tensor of all matter.

B How *not* to learn tensor calculus

This brief chapter is intended as a consolation to those students who have difficulty learning tensor calculus from some GR textbooks. Almost every explanation here (except for trivial statements) is flawed in one way or another, even though all the equations are formally correct. Because of this, a beginning student might become thoroughly confused and frustrated when trying to understand this material. However, the fault is with the explanations and not with the student! Better explanations are given elsewhere in this book.

B.1 Tensor algebra

Letters with Greek indices, such as A_α or $B^\lambda_{\mu\nu}$, denote arrays of numbers, and indices run over 0, ..., $N-1$, where N is the dimension of space. For brevity, we shall always use the **Einstein summation convention**: we omit the sign $\sum$ but *always* imply summation when an index is repeated in an expression, which must occur once as an upper index and once as a lower index. For example, we write

$$A_\alpha B^\alpha C_{\beta\gamma} D^\beta \equiv \sum_{\alpha=1}^{N} \sum_{\beta=1}^{N} A_\alpha B^\alpha C_{\beta\gamma} D^\beta.$$

The **Levi-Civita symbol** is defined as $\varepsilon_{123\ldots N} = 1$ and $\varepsilon_{\alpha\ldots\lambda\mu\ldots\rho} = -\varepsilon_{\alpha\ldots\mu\lambda\ldots\rho}$, i.e. $\varepsilon_{\alpha\ldots\rho}$ is totally antisymmetric in all the indices. For example, if $N = 3$ then $\varepsilon_{123} = \varepsilon_{231} = \varepsilon_{321} = 1$, $\varepsilon_{132} = \varepsilon_{213} = \varepsilon_{312} = -1$, and all the other components of $\varepsilon_{\alpha\beta\gamma}$ are equal to zero.

The **Kronecker symbol** $\delta_{\alpha\beta}$ is defined as $\delta_{\alpha\beta} = 1$ if $\alpha = \beta$ and $\delta_{\alpha\beta} = 0$ if $\alpha \neq \beta$. This symbol represents an identity matrix of dimensions $N \times N$. Sometimes, the Kronecker symbol is also written as δ^α_β.

It is easy to see that

$$\varepsilon_{\alpha_1\ldots\alpha_N}\varepsilon^{\alpha_1\ldots\alpha_N} = N!,$$
$$\varepsilon_{\lambda\alpha_1\ldots\alpha_{N-1}}\varepsilon^{\mu\alpha_1\ldots\alpha_{N-1}} = (N-1)!\delta^\mu_\lambda.$$

The **determinant** $\det A^{\alpha\beta}$ of a square matrix $A^{\alpha\beta}$ is the number defined by

$$\det A^{\alpha\beta} = \frac{1}{N!}\varepsilon_{\alpha_1\ldots\alpha_N}\varepsilon_{\beta_N\ldots\beta_N}A^{\alpha_1\beta_1}\ldots A^{\alpha_N\beta_N},$$

where the right-hand side contains N factors of $A^{\alpha\beta}$. It is well known that the determinant of the product of two matrices is equal to the product of their determinants.

An N-dimensional **contravariant vector** v^μ is an array of quantities $(v^1, \ldots, v^N)$ called **components** that transform under a change of basis according to the formula

$$\tilde{v}^\mu = v^\nu \left(S^{-1}\right)^\mu_\nu,$$

where S^μ_ν is the matrix describing the change of basis,

$$e^\mu_j \to \tilde{e}^\mu_j = S^k_j e^\mu_k, \tag{B.1}$$

and S^{-1} is the inverse matrix. Here, a **basis** is a set of N contravariant vectors $\left\{e^\mu_1, \ldots, e^\mu_N\right\}$ such that the determinant of the $N \times N$ matrix e^μ_j is nonzero. Such a matrix is called **nondegenerate**. Since the matrix e^μ_j is nondegenerate, any vector v^μ can be uniquely decomposed as a linear combination of basis vectors:

$$v^\mu = v^j e^\mu_j.$$

This is why the N numbers v^μ ($\mu = 1, \ldots, N$) are called the **components** of the vector v^μ. By convention, a *superscript* index is used for contravariant vectors.

A **metric** is a nondegenerate symmetric matrix $g_{\mu\nu} = g_{\nu\mu}$, $\det g_{\mu\nu} \neq 0$. The scalar product of vectors u^μ and v^μ is equal to $g_{\mu\nu}u^\mu v^\nu$ and can be written for brevity as

$$g_{\mu\nu}u^\mu v^\nu \equiv u^\mu v_\mu,$$

where the quantities v_μ are called the **covariant components** of the vector v^μ. Covariant components transform under a change of basis as

$$\tilde{v}_\mu = S^\nu_\mu v_\mu;$$

that is, they transform in the same way as the basis vectors do [Eq. (B.1)]. Such vectors v_μ are called **covariant vectors** or **covectors** and are written with a *subscript* index μ. Then, in every expression there can be only one pair of repeated indices, one subscript and one superscript, in agreement with the Einstein summation convention. An expression violating this rule, such as $a_\mu b_\mu c^\nu$, is written incorrectly, while $a_\mu b^\mu c^\nu$ is correct.

It is clear that the metric $g_{\mu\nu}$ can be used to transform contravariant components into covariant ones, $v_\mu = g_{\mu\nu}v^\nu$. This operation is called **lowering the index**. Likewise, the inverse metric $g^{\mu\nu}$ can be used to **raise the index**: $v^\mu = g^{\mu\nu}v_\nu$. Since raising an index after that index has been lowered should result in the same vector, it follows that

$$g_{\mu\nu}g^{\lambda\nu} = \delta^\lambda_\mu.$$

A **tensor** is a set of components with several indices, such as $A^{\alpha\beta}_\gamma$, that transforms under a change of basis (B.1) as a suitable product of components of covariant and contravariant vectors. For example, a tensor $A^{\alpha\beta}_\gamma$ transforms as the product $u^\alpha v^\beta w_\gamma$, i.e. according to the formula

$$\tilde{A}^{\alpha\beta}_\gamma = S^\nu_\gamma \left(S^{-1}\right)^\alpha_\lambda \left(S^{-1}\right)^\beta_\mu A^{\lambda\mu}_\nu.$$

This tensor is said to have **rank** (2,1). Tensors of any rank (m, n) are defined in this way; contravariant and covariant vectors are tensors of rank (1,0) and (0,1) respectively. It is clear that only tensors of equal rank can be added; for example, the expression $v^\alpha + u_\alpha$ does not transform as a vector and is therefore incorrect.

Note that $u^\alpha v^\beta$ is not the same tensor as $u^\beta v^\alpha$, but of course $v^\alpha v^\beta = v^\beta v^\alpha$ and $u^\alpha v^\beta = v^\beta u^\alpha$.

B.2 Tensor calculus

So far we considered vectors and tensors in Euclidean coordinates, but now we need to consider arbitrary, curvilinear coordinates x^μ. Note that x^μ itself is not a vector any more because the coordinates are not rectangular. However, an infinitesimal displacement δx^μ is a vector. To verify this, we consider an arbitrary change of coordinates described by arbitrary functions $\tilde{x}^\mu(x^\nu)$,

$$x^\mu \to \tilde{x}^\mu = \tilde{x}^\mu(x^\nu).$$

Then the displacement δx^μ transforms as

$$\delta\tilde{x}^\mu = \frac{\partial \tilde{x}^\mu}{\partial x^\nu}\delta x^\nu,$$

which is the correct transformation law for a vector under a change of basis (B.1) with the matrix $S^\mu_\nu = \partial x^\mu/\partial\tilde{x}^\nu$. Therefore, δx^μ is an example of a **contravariant vector field**, that is, a set of components $v^\mu(x^\nu)$ depending on the coordinates x^ν that transform under an arbitrary change of coordinates as components of a contravariant vector,

$$\tilde{v}^\mu = \frac{\partial \tilde{x}^\mu}{\partial x^\nu}v^\nu.$$

A **covariant vector field** transforms as components of a covariant vector. Analogously, a **tensor field of rank** (p,q) is defined as a set of components $T^{\alpha_1\ldots\alpha_p}{}_{\beta_1\ldots\beta_q}$ that transform under a change of coordinates as

$$\tilde{T}^{\alpha_1\ldots\alpha_p}{}_{\beta_1\ldots\beta_q} = \frac{\partial\tilde{x}^{\alpha_1}}{\partial x^{\lambda_1}}\ldots\frac{\partial\tilde{x}^{\alpha_p}}{\partial x^{\lambda_p}}\frac{\partial x^{\mu_1}}{\partial\tilde{x}^{\beta_1}}\ldots\frac{\partial x^{\mu_q}}{\partial\tilde{x}^{\beta_q}}T^{\lambda_1\ldots\lambda_p}{}_{\mu_1\ldots\mu_q}.$$

An important operation of tensor calculus is the covariant derivative. First, observe that a partial derivative of a contravariant vector field, $v^\mu{}_{,\nu} \equiv \partial v^\mu/\partial x^\nu$ does not transform as a tensor of rank (1,1). To get a derivative that transforms correctly, one adds a suitable "correction" term and thus defines the **covariant derivative**

$$v^\mu{}_{;\nu} \equiv \frac{\partial v^\mu}{\partial x^\nu} + \Gamma^\mu_{\alpha\nu}v^\alpha, \tag{B.2}$$

where the set of quantities $\Gamma^\mu_{\alpha\beta}$ is called the **Christoffel symbol**. It can be verified by a direct calculation that the covariant derivative $v^\mu{}_{;\nu}$ defined by Eq. (B.2) transforms as a tensor of rank (1,1) as long as the Christoffel symbol transforms as

$$\Gamma^\lambda_{\alpha\beta} = \tilde{\Gamma}^\mu_{\nu\rho}\frac{\partial\tilde{x}^\nu}{\partial x^\alpha}\frac{\partial\tilde{x}^\rho}{\partial x^\beta}\frac{\partial x^\lambda}{\partial\tilde{x}^\mu} + \frac{\partial^2 x^\lambda}{\partial\tilde{x}^\nu\partial\tilde{x}^\rho}\frac{\partial\tilde{x}^\nu}{\partial x^\alpha}\frac{\partial\tilde{x}^\rho}{\partial x^\beta}. \tag{B.3}$$

Therefore, we require that the transformation law (B.3) holds for the Christoffel symbol. Note that $\Gamma^\lambda_{\alpha\beta}$ is not a tensor, and a suitable change of coordinates $\{x^\mu\} \to \{\tilde{x}^\mu\}$ can make $\Gamma^\lambda_{\alpha\beta} = 0$ at any point. Such coordinates are called an **inertial coordinate system** at that point. However, $\Gamma^\lambda_{\alpha\beta} - \Gamma^\lambda_{\beta\alpha}$ is a tensor because the second term in Eq. (B.3) drops out when we write the transformation law for $\Gamma^\lambda_{\alpha\beta} - \Gamma^\lambda_{\beta\alpha}$. Since the tensor $\Gamma^\lambda_{\alpha\beta} - \Gamma^\lambda_{\beta\alpha}$ vanishes in a locally inertial coordinate system, it vanishes in all coordinate systems, hence $\Gamma^\lambda_{\alpha\beta}$ is symmetric in (α,β).

As shown in Eq. (B.2), the covariant derivative is denoted by a semicolon and an index. The covariant derivative of a *covariant* vector is defined similarly by

$$v_{\mu;\nu} = \frac{\partial v_\mu}{\partial x^\nu} - \Gamma^\lambda_{\mu\nu}v_\lambda.$$

Then it can be shown that the covariant derivative of the scalar product $a_\mu b^\mu$ coincides with the ordinary derivative, i.e.

$$\left(a_\mu b^\mu\right)_{;\alpha} = a_{\mu;\alpha}b^\mu + a_\mu b^\mu{}_{;\alpha} = \frac{\partial}{\partial x^\alpha}\left(a_\mu b^\mu\right).$$

This is, of course, to be expected for a scalar quantity $a_\mu b^\mu$.

Since tensors are quantities that transform as products of vectors and covectors, the covariant derivative can be defined on arbitrary tensors in a similar way. For each upper index there is a term with $+\Gamma^\lambda_{\alpha\beta}$, and for each lower index a term with $-\Gamma^\lambda_{\alpha\beta}$. For example,

$$A^{\alpha\beta}_{\gamma\ ;\mu} = \frac{\partial}{\partial x^\mu}A^{\alpha\beta}_\gamma + \Gamma^\alpha_{\mu\nu}A^{\nu\beta}_\gamma + \Gamma^\beta_{\mu\nu}A^{\alpha\nu}_\gamma - \Gamma^\nu_{\gamma\mu}A^{\alpha\beta}_\nu.$$

Then one can prove that the Leibnitz rule holds for arbitrary tensors,

$$\left(A^{\alpha\beta\ldots}B^{\gamma\delta\ldots}\right)_{;\mu} = A^{\alpha\beta\ldots}{}_{;\mu}B^{\gamma\delta\ldots} + A^{\alpha\beta\ldots}B^{\gamma\delta\ldots}{}_{;\mu}.$$

In General Relativity, one uses the Christoffel symbol of the form

$$\Gamma^\mu_{\alpha\nu} = \frac{1}{2}g^{\mu\beta}\left(g_{\alpha\beta,\nu} + g_{\nu\beta,\alpha} - g_{\alpha\nu,\beta}\right). \tag{B.4}$$

This expression can be derived from the property $\Gamma^\lambda_{\alpha\beta} = \Gamma^\lambda_{\beta\alpha}$. To derive Eq. (B.4), let us first prove that $g_{\mu\nu;\alpha} = 0$. For any vector X^β, we have $g_{\alpha\beta}X^\beta = X_\alpha$, hence

$$g_{\alpha\beta}\left(A^\beta\right)_{;\lambda} = \left(A_\alpha\right)_{;\lambda} \tag{B.5}$$

and therefore

$$\left(g_{\alpha\beta}A^\beta\right)_{;\lambda} = g_{\alpha\beta;\lambda}A^\beta + g_{\alpha\beta}\left(A^\beta\right)_{;\lambda}$$
$$\Rightarrow \quad g_{\alpha\beta;\lambda}A^\beta = 0.$$

The required property $g_{\alpha\beta;\lambda} = 0$ follows since A^β is an arbitrary vector. Then it is a matter of "juggling the indices" to derive Eq. (B.4). Namely, we write

$$0 = g_{\alpha\beta;\lambda} = g_{\alpha\beta,\lambda} - \Gamma^\mu_{\alpha\lambda}g_{\beta\mu} - \Gamma^\mu_{\beta\lambda}g_{\alpha\mu},$$

then exchange indices α,β,λ to get

$$0 = g_{\alpha\lambda,\beta} - \Gamma^\mu_{\alpha\beta}g_{\lambda\mu} - \Gamma^\mu_{\beta\lambda}g_{\alpha\mu},$$
$$0 = g_{\lambda\beta,\alpha} - \Gamma^\mu_{\alpha\lambda}g_{\beta\mu} - \Gamma^\mu_{\beta\alpha}g_{\lambda\mu}.$$

Adding the above two equations and subtracting the initial one, we obtain Eq. (B.4).

The **Riemann tensor** $R^\kappa_{\lambda\rho\sigma}$ is defined as the "covariant derivative" of the Christoffel symbol as follows,

$$R^\kappa_{\lambda\rho\sigma} = \partial_\rho\Gamma^\kappa_{\lambda\sigma} - \partial_\sigma\Gamma^\kappa_{\rho\lambda} + \Gamma^\kappa_{\nu\rho}\Gamma^\nu_{\sigma\lambda} - \Gamma^\kappa_{\nu\sigma}\Gamma^\nu_{\lambda\rho}.$$

It is easy to check that this formula defines a tensor (despite $\Gamma^\lambda_{\alpha\beta}$ not being a tensor), because one can notice that

$$R^\kappa_{\lambda\rho\sigma}u^\sigma = \nabla_\rho\nabla_\lambda u^\kappa - \nabla_\lambda\nabla_\rho u^\kappa, \tag{B.6}$$

which shows that $R^\kappa_{\lambda\rho\sigma}$ is manifestly a tensor. In flat space, $\nabla_\lambda \equiv \partial_\lambda$ and therefore $R^\kappa_{\lambda\rho\sigma} = 0$ because $\partial_\lambda\partial_\mu = \partial_\mu\partial_\lambda$.

B.3 Hints

In this section, I give you some hints as to why the preceding explanations are flawed.

The Levi-Civita symbol ε is not merely a collection of numbers 0, 1, and -1, arranged in a special way, but is interpreted as an antisymmetric tensor or a volume N-form. The Kronecker symbol δ^{α}_{β} is the matrix representing an identity transformation (in any basis), while the matrix $\delta_{\alpha\beta}$ represents (in an orthogonal basis) a bilinear form that may represent a scalar product in a Euclidean space. Determinants of matrices are not simply some complicated combinations of matrix elements. Determinants have a direct geometric meaning; for instance, the oriented volume of the image of the unit cube under a linear transformation T is equal to $\det T$.

Vectors are most clearly seen as geometric quantities — elements of vector spaces, rather than collections of components that mysteriously "transform themselves under a change of basis." Bases are maximal sets of linearly independent vectors, and the components of a vector are coefficients relative to a chosen basis. It is easier to regard the transformation formulas for components as consequences of the geometric picture rather than as definitions.

Covariant vectors are better seen as vectors from a *different* vector space (the dual space), rather than as a "different kind of components" of the same vector. There is no natural map between contravariant and covariant vectors unless a metric is given; and when several metrics are given, there are several such maps. The idea that there exist "covariant components" of vectors can be justified only if a metric is fixed once and for all.

The inverse metric $g^{\alpha\beta}$ is either *defined* as the inverse matrix to $g_{\alpha\beta}$, or is derived through the dual basis (in the dual space).

A tensor is an element of a tensor space (the space of tensor products and their linear combinations). As in the case of vectors, the components of a tensor with respect to a basis can be defined after the tensor space is defined. Then the complicated transformation law for the components will be a natural consequence of the construction, rather than a definition of a tensor.

An "infinitesimal displacement" δx^{μ} is a heuristic representation of a tangent vector: one imagines that a point p moves by "infinitesimal" amount along a curve which is a flow line of a given tangent vector (see Sec. 1.2.5). "Covariant" tangent vectors are elements of the dual tangent space.

A covariant derivative is not simply an old derivative with a "correction" added to it, but rather the only type of directional derivative that can be meaningfully considered as a geometric object in a curved space. (The coordinate derivative $\partial v^{\mu}/\partial x^{\nu} \equiv v^{\mu}{}_{,\nu}$ does not exist as a geometric object since it depends on the coordinate system $\{x^{\mu}\}$.) However, there exist infinitely many possible covariant derivatives since there are infinitely many possible $\Gamma^{\lambda}_{\alpha\beta}$.

The Christoffel symbol $\Gamma^{\lambda}_{\alpha\beta}$ cannot be set to zero by a change of coordinates unless $T^{\lambda}_{\alpha\beta} \equiv \Gamma^{\lambda}_{\alpha\beta} - \Gamma^{\lambda}_{\beta\alpha} = 0$ (torsion-freeness) already holds. If the torsion tensor $T^{\lambda}_{\alpha\beta}$ is nonzero, locally inertial coordinate systems do not exist; it is a *physical* assumption that they exist, rather than a mathematical property.

Covariant derivative is *defined* on arbitrary tensors by the requirements that the Leibnitz rule hold and that ∇ coincide with ∂ on scalar functions. These properties cannot be derived from the covariant transformation alone. The requirement of correct transformation of components is not sufficient to define covariant derivatives of arbitrary tensors, e.g. $A_{\alpha\beta\gamma}{}^{\mu}$, because one could in principle choose a different $\Gamma^{\lambda}_{\alpha\beta}$ for tensors of each different rank. One needs extra information to derive the fact that the *same* $\Gamma^{\lambda}_{\alpha\beta}$ is used in covariant derivatives of tensors of every rank. This extra information comes from *assuming* the Leibnitz rule and the property $\nabla_{\mu} f = \partial_{\mu} f$ for scalar functions f.

The assumption $g_{\alpha\beta;\lambda} = 0$ is a *separate* physical assumption that cannot be actually derived from general properties of vectors; the argument given above assumes $g_{\alpha\beta;\lambda} = 0$ tacitly in Eq. (B.5). The formula (B.4) indeed follows from $\Gamma^{\lambda}_{\alpha\beta} = \Gamma^{\lambda}_{\beta\alpha}$ and $g_{\alpha\beta;\lambda} = 0$.

The Riemann tensor cannot be seen as a covariant derivative of $\Gamma^{\lambda}_{\alpha\beta}$ because $\Gamma^{\lambda}_{\alpha\beta}$ is not a tensor, and in any case, the given formula does not represent a covariant derivative of a third-rank tensor. The property (B.6) is ordinarily used as a *definition* of $R^{\kappa}_{\lambda\rho\sigma}$.

C Calculations and proofs

C.1 For Chapter 1

Proof of Statement 1.2.2.1 on page 5: Suppose such a smooth map exists and maps the north pole $\{0,0,1\}$ into a point $p_0 \in \mathbb{R}^2$. Consider a very small circle around the north pole. Since the chart provides a smooth one-to-one map, this circle will be mapped into a small "image" circle around p_0 in $\mathbb{R}^2$. The circle around the north pole can be smoothly contracted into a point in two essentially different ways: either by contracting it to the north pole, or by shifting it around the sphere towards the south pole and contracting to the south pole, without ever crossing the north pole. These deformations of the circle are, by assumption, smoothly mapped into corresponding deformations of the image circle in $\mathbb{R}^2$. During the second deformation, the image circle around p_0 in $\mathbb{R}^2$ is supposedly contracted into a point without ever crossing p_0. But such a deformation is obviously impossible within $\mathbb{R}^2$. This contradicts the assumption that a smooth one-to-one map $S^2 \to \mathbb{R}^2$ exists. ■

Proof of Statement 1.2.4.1 on page 9: A smooth function can be approximated arbitrarily precisely by a polynomial, and it is technically convenient to assume that f is a polynomial,

$$f(z) = \sum_{m=0}^{m_{\max}} \frac{1}{m!} z^m \left. \frac{d^{(m)} f}{dz^m} \right|_{z=0}.$$

Since derivations are linear maps by Eq. (1.4), it is sufficient to prove the statement for $f(z) = z^m$, $m = 0, 1, 2, \ldots$ By Eq. (1.6), we have $\mathbf{v} \circ 1 = 0$. It is trivial to check the statement for $f(z) = z$. Now use induction in m, starting with $m = 1$, i.e. with $f(z) = z$. Once the statement is proved for $f(z) = z^{m-1}$, we have

$$\mathbf{v} \circ \left(g^{m-1}\right) = (m-1)\, g^{m-2} \mathbf{v} \circ g.$$

Now we consider $f(z) = z^m$ and use Eq. (1.5),

$$\begin{aligned} \mathbf{v} \circ f(g) = \mathbf{v} \circ \left(g g^{m-1}\right) &= (\mathbf{v} \circ g)\, g^{m-1} + g \mathbf{v} \circ (g^{m-1}) \\ &= m g^{m-1} \mathbf{v} \circ g, \end{aligned}$$

thus the statement is proved for $f(z) = z^m$. ■

Proof of Statement 1.2.10.2 on page 13: It is *not* necessary to introduce a local coordinate system around p_0. Introduce the points p_1 and p_2 as in Fig. 1.7. By definition of a tangent vector and its orbits, we have

$$(\mathbf{a} \circ f)|_{p_0} = \lim_{\delta\tau \to 0} \frac{f(p_1) - f(p_0)}{\delta\tau}, \tag{C.1}$$

because the point p_1 is obtained by following the orbit of $\mathbf{a}$ for an interval $\delta\tau$, while the right-hand side of Eq. (C.1) is the definition of the directional derivative along the orbit. The relation (C.1) holds for every function f, so let us apply it to the function $\mathbf{b} \circ f$ instead of f,

$$(\mathbf{a} \circ (\mathbf{b} \circ f))|_{p_0} = \lim_{\delta\tau \to 0} \frac{1}{\delta\tau} \left(\mathbf{b} \circ f(p_1) - \mathbf{b} \circ f(p_0)\right). \tag{C.2}$$

We now need to express the function $\mathbf{b} \circ f$ similarly,

$$\begin{aligned} (\mathbf{b} \circ f)|_{p_1} &= \lim_{\delta\tau' \to 0} \frac{f(p) - f(p_1)}{\delta\tau'}, \\ (\mathbf{b} \circ f)|_{p_0} &= \lim_{\delta\tau' \to 0} \frac{f(p_2) - f(p_0)}{\delta\tau'}, \end{aligned}$$

where we introduced a different variable $\delta\tau'$ for convenience. Substituting this into Eq. (C.2), we find

$$(\mathbf{a} \circ (\mathbf{b} \circ f))|_{p_0} = \lim_{\substack{\delta\tau \to 0 \\ \delta\tau' \to 0}} \frac{f(p) - f(p_1) - f(p_2) + f(p_0)}{\delta\tau\, \delta\tau'}.$$

Similarly, using the analogous relation

$$(\mathbf{a} \circ f)|_{p_2} = \lim_{\delta\tau \to 0} \frac{f(p') - f(p_2)}{\delta\tau},$$

we find

$$(\mathbf{b} \circ (\mathbf{a} \circ f))|_{p_0} = \lim_{\substack{\delta\tau \to 0 \\ \delta\tau' \to 0}} \frac{f(p') - f(p_2) - f(p_1) + f(p_0)}{\delta\tau\, \delta\tau'}.$$

It follows that

$$([\mathbf{a}, \mathbf{b}] \circ f)|_{p_0} = \lim_{\substack{\delta\tau \to 0 \\ \delta\tau' \to 0}} \frac{f(p) - f(p')}{\delta\tau\, \delta\tau'},$$

as required. ■

Calculation 1.2.11.1 on page 13: Using the family of curves $\gamma(\tau; s)$ as a coordinate grid, we may introduce a local coordinate system $\{\tau, s, t_1, \ldots, t_{n-2}\}$ such that τ and s are the first two coordinates. Then $\mathbf{v} = \partial_\tau$ and $\mathbf{c} = \partial_s$, therefore the vectors $\mathbf{c}$ and $\mathbf{v}$ are the first two vectors of the coordinate basis corresponding to the local coordinate system $\{\tau, s, t_1, \ldots, t_{n-2}\}$. Hence, these vectors commute. ■

Proof of Statement 1.2.11.2 on page 14: We may complete the vector $\mathbf{v}(p_0) \neq 0$ to an arbitrary basis $\{\mathbf{v}(p_0), \mathbf{c}_1(p_0), \mathbf{c}_2(p_0), \ldots\}$ at the initial point p_0. Then the relation $[\mathbf{v}, \mathbf{c}_j] = 0$ is a first-order differential equation for $\mathbf{c}_j$ that has a unique solution $\mathbf{c}_j(p)$ along one orbit of $\mathbf{v}$ starting from the initial point p_0, at least in some neighborhood of p_0. To see this explicitly, consider the equation $[\mathbf{c}, \mathbf{v}] = 0$ in a local coordinate system,

$$\sum_\alpha c^\alpha \frac{\partial v^\beta}{\partial x^\alpha} - \sum_\alpha v^\alpha \frac{\partial c^\beta}{\partial x^\alpha} = 0.$$

Consider one orbit $\gamma(\tau)$ of $\mathbf{v}$, such that $\mathbf{v} \equiv \dot{\gamma}$ is the tangent vector to the curve $\gamma(\tau)$. For points p on the curve γ, the second term above is equal to the derivative of the component c^β along this orbit,

$$\left. \sum_\alpha v^\alpha \frac{\partial c^\beta}{\partial x^\alpha} \right|_p = \frac{d}{d\tau} c^\beta(\tau),$$

assuming that $\gamma(\tau) = p$. So the equation $[\mathbf{c}, \mathbf{v}] = 0$ at point p becomes

$$\frac{d}{d\tau} c^\beta = \sum_\alpha c^\alpha \frac{\partial v^\beta}{\partial x^\alpha},$$

which is an ordinary differential equation for the unknown components $c^\beta(\tau)$, to be solved with an initial condition $c^\beta(\tau_0)$. This differential equation has a unique solution (by assumption, the vector field $\mathbf{v}$ is smooth everywhere). Thus we can in principle compute the vectors $\mathbf{c}_j(p)$, where p is any point along one orbit of $\mathbf{v}$. The set of vectors $\{\mathbf{v}(p), \mathbf{c}_1(p), ..., \mathbf{c}_{n-1}(p)\}$ is linearly independent at the initial point p_0 and will remain linearly independent along the orbit at least in some neighborhood of p_0. We can prove this by noting that linear independence of the vectors $\{\mathbf{v}, \mathbf{c}_1, ..., \mathbf{c}_{n-1}\}$ can be verified by computing the determinant of their components in a local coordinate system. The vectors form a basis iff the determinant is nonzero. By assumption, the determinant is nonzero at p_0, and the determinant is a smooth function along the orbit. Thus the determinant remains nonzero along the orbit at least in some neighborhood of p_0 (perhaps in a very small neighborhood, but that is all we need presently).

So far we obtained a basis of connecting vectors along one orbit of $\mathbf{v}$. Since the orbits of $\mathbf{v}$ fill at least a small patch $\mathcal{U} \subset \mathcal{M}$ of the manifold $\mathcal{M}$ around the point p_0, the same construction gives a basis of connecting vector fields along every other orbit within $\mathcal{U}$, again perhaps only in a sufficiently small neighborhood of p_0. Thus we have obtained a basis of connecting fields *locally*, in a neighborhood of p_0. ■

Proof of Statement 1.4.4.1 on page 21: **(a)** Let $\{\mathbf{e}_1, ..., \mathbf{e}_n\}$ be a basis in $\mathbb{R}^n$; we will show that $\mathbf{a}_1 \wedge ... \wedge \mathbf{a}_n$ is either equal to zero, or proportional to $\mathbf{e}_1 \wedge ... \wedge \mathbf{e}_n$ with a nonzero coefficient. It will follow that any two n-vectors are proportional.

First we show that $\mathbf{a}_1 \wedge ... \wedge \mathbf{a}_n = 0$ if the set $\{\mathbf{a}_1, ..., \mathbf{a}_n\}$ is linearly dependent. This follows from the antisymmetry of the exterior product: suppose

$$\mathbf{a}_n = \sum_{j=1}^{n-1} \lambda_j \mathbf{a}_j,$$

where some λ_j are nonzero. Since

$$\mathbf{a}_1 \wedge ... \wedge \mathbf{a}_{n-1} \wedge \mathbf{a}_j = 0 \quad \text{for } j = 1, ..., n-1,$$

we have

$$\mathbf{a}_1 \wedge ... \wedge \mathbf{a}_n = \mathbf{a}_1 \wedge ... \wedge \mathbf{a}_{n-1} \wedge \sum_{j=1}^{n-1} \lambda_j \mathbf{a}_j = 0.$$

Now suppose that the set $\{\mathbf{a}_j\}$ is linearly independent. Then every basis vector $\mathbf{e}_k$ can be expressed as a linear combination of $\{\mathbf{a}_j\}$ with some coefficients,

$$\mathbf{e}_k = \sum_j A_{kj} \mathbf{a}_j.$$

Let us now substitute these linear combinations into the nonzero multivector $\mathbf{e}_1 \wedge ... \wedge \mathbf{e}_n$, and simplify the resulting expression using the linearity of $\wedge$. The result will be a sum of terms of the form $\mathbf{a}_{j_1} \wedge ... \wedge \mathbf{a}_{j_n}$ with some coefficients. Only the terms with all different j_k will survive. Finally, we can reorder the vectors $\{\mathbf{a}_j\}$ at will (and if necessary change the sign of $\mathbf{a}_1 \wedge ... \wedge \mathbf{a}_n$), so eventually we will obtain simply $\mathbf{a}_1 \wedge ... \wedge \mathbf{a}_n$ with a nonzero coefficient. Therefore, the multivectors $\mathbf{a}_1 \wedge ... \wedge \mathbf{a}_n$ and $\mathbf{e}_1 \wedge ... \wedge \mathbf{e}_n$ are proportional.

(b) We need to show that the relation between n-dimensional volumes,

$$\lambda \text{Vol}\,(\mathbf{a}_1, ..., \mathbf{a}_n) = \text{Vol}\,(\mathbf{b}_1, ..., \mathbf{b}_n),$$

is equivalent to the relation between multivectors

$$\lambda \mathbf{a}_1 \wedge ... \wedge \mathbf{a}_n = \mathbf{b}_1 \wedge ... \wedge \mathbf{b}_n.$$

It is sufficient to consider the case when all the volumes involved are nonzero. By part **(a)**, the two multivectors are always proportional with *some* nonzero coefficient; denote that coefficient by λ. It remains to show that the volumes are related by the same factor λ.

To show this, we will transform the parallelepiped spanned by $\{\mathbf{a}_j\}$ into the parallelepiped spanned by $\{\mathbf{b}_j\}$. The transformation will be performed through a sequence of steps which preserve the relationship between the volume of the parallelepiped and the multivector $\mathbf{a}_1 \wedge ... \wedge \mathbf{a}_n$. Each step will either multiply both the volume and the multivector by the same number, or leave them both unchanged. At the end, the multivector $\mathbf{a}_1 \wedge ... \wedge \mathbf{a}_n$ will be transformed into $\mathbf{b}_1 \wedge ... \wedge \mathbf{b}_n$; hence, the volumes are related by the factor λ, which will prove the statement.

The allowed steps of the transformation are of three kinds: (i) adding a multiple of $\mathbf{a}_j$ to $\mathbf{a}_i$,

$$\mathbf{a}_i \to \mathbf{a}_i + \alpha \mathbf{a}_j,$$

where $\alpha \neq 0$ is some number; (ii) stretching a vector $\mathbf{a}_i$,

$$\mathbf{a}_i \to \alpha \mathbf{a}_i,$$

where $\alpha \neq 0$; (iii) exchanging two vectors, $\mathbf{a}_i$ with $\mathbf{a}_j$. It is clear that the multivector

$$\mathbf{a}_1 \wedge ... \wedge \mathbf{a}_n$$

remains unchanged under (i), is multiplied by α under (ii), and changes sign under (iii). Similarly, it is easy to see that the volume of the parallelepiped spanned by $\{\mathbf{a}_j\}$ remains unchanged under (i) due to the argument of Fig. 1.10, applied to the two-dimensional plane containing the vectors $\{\mathbf{a}_i, \mathbf{a}_j\}$. The volume is multiplied by α under (ii) and changes sign under (iii).

By assumption, the initial set $\{\mathbf{a}_j\}$ is linearly independent (otherwise the volume is zero); thus, every vector $\{\mathbf{b}_k\}$ can be expressed through linear combinations of $\{\mathbf{a}_j\}$. These linear combinations can be built by a finite sequence of steps (i), (ii), or (iii). In this way, the multivector $\mathbf{a}_1 \wedge ... \wedge \mathbf{a}_n$ can be transformed into $\mathbf{b}_1 \wedge ... \wedge \mathbf{b}_n$. This concludes the proof. ■

Proof of Statement 1.4.5.1 on page 21: After Statement 1.4.4.1, we only need to prove the relationship between the multivectors,

$$\hat{T}\mathbf{v}_1 \wedge \hat{T}\mathbf{v}_2 \wedge ... \wedge \hat{T}\mathbf{v}_n = (\mathbf{v}_1 \wedge \mathbf{v}_2 \wedge ... \wedge \mathbf{v}_n) \det \hat{T}.$$

Since any two n-vectors are proportional in an n-dimensional space, it remains to prove that the proportionality factor $\det \hat{T}$ is actually independent of the choice of the vectors $\{\mathbf{v}_j\}$. (It will follow that $\det \hat{T}$ coincides with the determinant defined through an orthogonal basis.)

Let us first give the proof for $n = 2$ and then generalize to any dimension n. In two dimensions, we have

$$\hat{T}\mathbf{v}_1 \wedge \hat{T}\mathbf{v}_2 = (\mathbf{v}_1 \wedge \mathbf{v}_2) \det \hat{T}. \tag{C.3}$$

The bivector $\mathbf{v}_1 \wedge \mathbf{v}_2$ is zero if $\mathbf{v}_1$ is parallel to $\mathbf{v}_2$; in this case, the statement is trivial, so let us consider the case $\mathbf{v}_1 \wedge \mathbf{v}_2 \neq 0$. Then the vectors $\{\mathbf{v}_1, \mathbf{v}_2\}$ are a basis in the plane. We would like to show that

$$\hat{T}\mathbf{u}_1 \wedge \hat{T}\mathbf{u}_2 = (\mathbf{u}_1 \wedge \mathbf{u}_2) \det \hat{T} \tag{C.4}$$

for any vectors $\mathbf{u}_1, \mathbf{u}_2$. Since $\{\mathbf{v}_1, \mathbf{v}_2\}$ is a basis, we can decompose

$$\mathbf{u}_1 = U_{11}\mathbf{v}_1 + U_{12}\mathbf{v}_2, \quad \mathbf{u}_2 = U_{21}\mathbf{v}_1 + U_{22}\mathbf{v}_2.$$

Since $\hat{T}$ is a linear map, we also have a similar decomposition

$$\hat{T}\mathbf{u}_1 = U_{11}\hat{T}\mathbf{v}_1 + U_{12}\hat{T}\mathbf{v}_2, \quad \hat{T}\mathbf{u}_2 = U_{21}\hat{T}\mathbf{v}_1 + U_{22}\hat{T}\mathbf{v}_2.$$

Then we compute

$$\begin{aligned}
\mathbf{u}_1 \wedge \mathbf{u}_2 &= (U_{11}\mathbf{v}_1 + U_{12}\mathbf{v}_2) \wedge (U_{21}\mathbf{v}_1 + U_{22}\mathbf{v}_2) \\
&= (U_{11}U_{22} - U_{12}U_{21})\, \mathbf{v}_1 \wedge \mathbf{v}_2, \\
\hat{T}\mathbf{u}_1 \wedge \hat{T}\mathbf{u}_2 &= (U_{11}\hat{T}\mathbf{v}_1 + U_{12}\hat{T}\mathbf{v}_2) \wedge (U_{21}\hat{T}\mathbf{v}_1 + U_{22}\hat{T}\mathbf{v}_2) \\
&= (U_{11}U_{22} - U_{12}U_{21})\, \hat{T}\mathbf{v}_1 \wedge \hat{T}\mathbf{v}_2.
\end{aligned}$$

Therefore, the relationship (C.4) simply follows from Eq. (C.3) through multiplication by the factor $U_{11}U_{22} - U_{12}U_{21}$.

In n dimensions, it is sufficient to consider the case when $\mathbf{v}_1 \wedge \ldots \wedge \mathbf{v}_n \neq 0$, i.e. the set $\{\mathbf{v}_j\}$ is a basis. We need to show that

$$\hat{T}\mathbf{u}_1 \wedge \ldots \wedge \hat{T}\mathbf{u}_n = (\mathbf{u}_1 \wedge \ldots \wedge \mathbf{u}_n) \det \hat{T}$$

for any other n-vector $\mathbf{u}_1 \wedge \ldots \wedge \mathbf{u}_n$. Since every $\mathbf{u}_j$ can be expressed as a linear combination of $\{\mathbf{v}_j\}$,

$$\mathbf{u}_j = \sum_k U_{jk}\mathbf{v}_k,$$

we may substitute these linear combinations into $\mathbf{u}_1 \wedge \ldots \wedge \mathbf{u}_n$ and also into $\hat{T}\mathbf{u}_1 \wedge \ldots \wedge \hat{T}\mathbf{u}_n$. Since $\hat{T}$ is a linear map,

$$\hat{T}\mathbf{u}_j = \sum_k U_{jk}\hat{T}\mathbf{v}_k,$$

so the substitution of $\hat{T}\mathbf{u}_j = \sum_k U_{jk}\hat{T}\mathbf{v}_k$ into $\hat{T}\mathbf{u}_1 \wedge \ldots \wedge \hat{T}\mathbf{u}_n$ yields terms

$$\hat{T}\mathbf{v}_1 \wedge \ldots \hat{T}\mathbf{v}_n$$

with the same coefficients as the substitution of $\mathbf{u}_j = \sum_k U_{jk}\mathbf{v}_k$ into $\mathbf{u}_1 \wedge \ldots \wedge \mathbf{u}_n$. Therefore the desired relationship for $\{\mathbf{u}_j\}$ follows. ■

Calculation 1.8.4.1 on page 48: Denote by $\tilde{\nabla}$ and ∇ the Levi-Civita connections corresponding to the metrics $\tilde{g}$ and g, and analogously the Riemann tensors $\tilde{R}(\ldots)$ and $R(\ldots)$. It is convenient to consider the fully covariant Riemann tensor,

$$\tilde{R}(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) = \tilde{g}\left([\tilde{\nabla}_\mathbf{a}, \tilde{\nabla}_\mathbf{b}]\mathbf{c} - \tilde{\nabla}_{[\mathbf{a},\mathbf{b}]}\mathbf{c}, \mathbf{d}\right).$$

Since the derivatives of the arbitrary vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$ will eventually cancel, we can simplify the calculations if we assume that all these derivatives vanish, so $\nabla_\mathbf{a}\mathbf{b} = 0$, $[\mathbf{a}, \mathbf{b}] = 0$, etc., at a point p where we are computing the Riemann tensor. (But note that $\tilde{\nabla}_\mathbf{a}\mathbf{b} \neq 0$, etc.) Since $[\mathbf{a}, \mathbf{b}] = 0$, we have

$$\begin{aligned}
e^{-2\lambda}\tilde{R}(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) &= g\left(\tilde{\nabla}_\mathbf{a}\tilde{\nabla}_\mathbf{b}\mathbf{c} - \tilde{\nabla}_\mathbf{b}\tilde{\nabla}_\mathbf{a}\mathbf{c}, \mathbf{d}\right), \\
R(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) &= g\left(\nabla_\mathbf{a}\nabla_\mathbf{b}\mathbf{c} - \nabla_\mathbf{b}\nabla_\mathbf{a}\mathbf{c}, \mathbf{d}\right),
\end{aligned}$$

We would like to transform the expression $\tilde{R}(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d})$ to something containing $R(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d})$ plus extra terms. It is clear that we need to express $\tilde{\nabla}$ everywhere through ∇. We may work as follows. The only way to express $\tilde{\nabla}$ through ∇ is by enclosing them in the metric $\tilde{g}$ and using Eq. (1.47). The difference between the connections $\tilde{\nabla}$ and ∇ is a transformation-valued 1-form $\hat{\Gamma}$, which is expressed as

$$\begin{aligned}
\hat{\Gamma}(\mathbf{a})\mathbf{b} &\equiv \tilde{\nabla}_\mathbf{a}\mathbf{b} - \nabla_\mathbf{a}\mathbf{b}, \\
g(\hat{\Gamma}(\mathbf{a})\mathbf{b}, \mathbf{c}) &= (\mathbf{a} \circ \lambda)\, g(\mathbf{b}, \mathbf{c}) + (\mathbf{b} \circ \lambda)\, g(\mathbf{a}, \mathbf{c}) \\
&\quad - (\mathbf{c} \circ \lambda)\, g(\mathbf{a}, \mathbf{b}).
\end{aligned}$$

We would like to write $\hat{\Gamma}$ more explicitly, without enclosing it in g. To this end, we introduce an auxiliary (but fixed) vector field $\mathbf{l} \equiv \hat{g}^{-1}\mathrm{d}\lambda$. Then $\mathbf{x} \circ \lambda \equiv g(\mathbf{x}, \mathbf{l})$ for every vector $\mathbf{x}$. Using the field $\mathbf{l}$, we can rewrite $\hat{\Gamma}$ more concisely,

$$\hat{\Gamma}(\mathbf{a})\mathbf{b} = g(\mathbf{a}, \mathbf{l})\mathbf{b} + g(\mathbf{b}, \mathbf{l})\mathbf{a} - g(\mathbf{a}, \mathbf{b})\mathbf{l}.$$

Now we consider the first term in the new Riemann tensor,

$$\begin{aligned}
\tilde{\nabla}_\mathbf{a}\tilde{\nabla}_\mathbf{b}\mathbf{c} &= \left(\nabla_\mathbf{a} + \hat{\Gamma}(\mathbf{a})\right)\left(\nabla_\mathbf{b} + \hat{\Gamma}(\mathbf{b})\right)\mathbf{c} \\
&= \nabla_\mathbf{a}\nabla_\mathbf{b}\mathbf{c} + \nabla_\mathbf{a}\hat{\Gamma}(\mathbf{b})\mathbf{c} + \hat{\Gamma}(\mathbf{a})\hat{\Gamma}(\mathbf{b})\mathbf{c},
\end{aligned}$$

where we omitted terms containing first derivatives of $\mathbf{b}$ and $\mathbf{c}$, since by assumption these derivatives vanish. The derivative $\nabla_\mathbf{a}\hat{\Gamma}$ can be computed as follows,

$$\nabla_\mathbf{a}\hat{\Gamma}(\mathbf{b})\mathbf{c} = g(\mathbf{b}, \nabla_\mathbf{a}\mathbf{l})\mathbf{c} + g(\mathbf{c}, \nabla_\mathbf{a}\mathbf{l})\mathbf{b} - g(\mathbf{b}, \mathbf{c})\nabla_\mathbf{a}\mathbf{l},$$

again omitting first derivatives of $\mathbf{b}, \mathbf{c}$. Finally, we simplify the last term,

$$\begin{aligned}
\hat{\Gamma}(\mathbf{a})\hat{\Gamma}(\mathbf{b})\mathbf{c} &= g(\mathbf{a}, \mathbf{l})\hat{\Gamma}(\mathbf{b})\mathbf{c} + g(\hat{\Gamma}(\mathbf{b})\mathbf{c}, \mathbf{l})\mathbf{a} - g(\mathbf{a}, \hat{\Gamma}(\mathbf{b})\mathbf{c})\mathbf{l} \\
&= g(\mathbf{a}, \mathbf{l})\left(g(\mathbf{b}, \mathbf{l})\mathbf{c} + g(\mathbf{c}, \mathbf{l})\mathbf{b} - g(\mathbf{b}, \mathbf{c})\mathbf{l}\right) \\
&\quad + \left(g(\mathbf{b}, \mathbf{l})g(\mathbf{c}, \mathbf{l}) + g(\mathbf{c}, \mathbf{l})g(\mathbf{b}, \mathbf{l}) - g(\mathbf{b}, \mathbf{c})g(\mathbf{l}, \mathbf{l})\right)\mathbf{a} \\
&\quad - \left(g(\mathbf{b}, \mathbf{l})g(\mathbf{a}, \mathbf{c}) + g(\mathbf{c}, \mathbf{l})g(\mathbf{a}, \mathbf{b}) - g(\mathbf{b}, \mathbf{c})g(\mathbf{a}, \mathbf{l})\right)\mathbf{l} \\
&= \left(2g(\mathbf{b}, \mathbf{l})g(\mathbf{c}, \mathbf{l}) - g(\mathbf{b}, \mathbf{c})g(\mathbf{l}, \mathbf{l})\right)\mathbf{a} + g(\mathbf{a}, \mathbf{l})g(\mathbf{b}, \mathbf{l})\mathbf{c} \\
&\quad + g(\mathbf{a}, \mathbf{l})g(\mathbf{c}, \mathbf{l})\mathbf{b} - \left(g(\mathbf{b}, \mathbf{l})g(\mathbf{a}, \mathbf{c}) + g(\mathbf{c}, \mathbf{l})g(\mathbf{a}, \mathbf{b})\right)\mathbf{l}.
\end{aligned}$$

Now we need to antisymmetrize the resulting expression for $\tilde{\nabla}_\mathbf{a}\tilde{\nabla}_\mathbf{b}\mathbf{c}$ in $\mathbf{a}, \mathbf{b}$. Note that

$$g(\mathbf{b}, \nabla_\mathbf{a}\mathbf{l}) = \nabla_\mathbf{a} g(\mathbf{b}, \mathbf{l}) = \mathbf{a} \circ (\mathbf{b} \circ \lambda),$$

and also $[\mathbf{a}, \mathbf{b}] = 0$ by assumption, therefore $g(\mathbf{b}, \nabla_\mathbf{a}\mathbf{l})$ is a symmetric bilinear form in $\mathbf{a}, \mathbf{b}$. This bilinear form is called the **Hessian** of the function λ; it is a tensor representing all the (covariant) second derivatives of λ. Let us denote this tensor by

$$H_\lambda(\mathbf{a}, \mathbf{b}) = H^\cdot(\mathbf{b}, \mathbf{a}) \equiv g(\nabla_\mathbf{a}\mathbf{l}, \mathbf{b});$$

in the index notation this would be $\lambda_{;\alpha\beta}$. Then

$$\nabla_\mathbf{a}\hat{\Gamma}(\mathbf{b})\mathbf{c} = H_\lambda(\mathbf{a}, \mathbf{b})\mathbf{c} + H_\lambda(\mathbf{a}, \mathbf{c})\mathbf{b} - g(\mathbf{b}, \mathbf{c})\nabla_\mathbf{a}\mathbf{l},$$

so the antisymmetrization of $\nabla_\mathbf{a}\hat{\Gamma}(\mathbf{b})\mathbf{c}$ in $\mathbf{a}, \mathbf{b}$ cancels $H_\lambda(\mathbf{a}, \mathbf{b})$ and yields

$$\begin{aligned}
\nabla_\mathbf{a}\hat{\Gamma}(\mathbf{b})\mathbf{c} - \nabla_\mathbf{b}\hat{\Gamma}(\mathbf{a})\mathbf{c} &= H_\lambda(\mathbf{a}, \mathbf{c})\mathbf{b} - g(\mathbf{b}, \mathbf{c})\nabla_\mathbf{a}\mathbf{l} \\
&\quad - H_\lambda(\mathbf{b}, \mathbf{c})\mathbf{a} + g(\mathbf{a}, \mathbf{c})\nabla_\mathbf{b}\mathbf{l}.
\end{aligned}$$

The antisymmetrization of $\hat{\Gamma}(\mathbf{a})\hat{\Gamma}(\mathbf{b})\mathbf{c}$ gives (after some cancellations)

$$\begin{aligned}
\hat{\Gamma}(\mathbf{a})\hat{\Gamma}(\mathbf{b})\mathbf{c} - \hat{\Gamma}(\mathbf{b})\hat{\Gamma}(\mathbf{a})\mathbf{c} &= g(\mathbf{l}, \mathbf{l})\left(g(\mathbf{a}, \mathbf{c})\mathbf{b} - g(\mathbf{b}, \mathbf{c})\mathbf{a}\right) \\
&\quad + g(\mathbf{a}, \mathbf{l})\left(g(\mathbf{b}, \mathbf{c})\mathbf{l} - g(\mathbf{c}, \mathbf{l})\mathbf{b}\right) + g(\mathbf{b}, \mathbf{l})\left(g(\mathbf{c}, \mathbf{l})\mathbf{a} - g(\mathbf{a}, \mathbf{c})\mathbf{l}\right).
\end{aligned}$$

Finally, putting the pieces together, we express the new Riemann tensor as follows,[1]

$$\begin{aligned} e^{-2\lambda}\tilde{R}(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d}) &= R(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d}) \\ &+ H_\lambda(\mathbf{a},\mathbf{c})g(\mathbf{b},\mathbf{d}) - H_\lambda(\mathbf{a},\mathbf{d})g(\mathbf{b},\mathbf{c}) \\ &- H_\lambda(\mathbf{b},\mathbf{c})g(\mathbf{a},\mathbf{d}) + H_\lambda(\mathbf{b},\mathbf{d})g(\mathbf{a},\mathbf{c}) \\ &+ g(\mathbf{l},\mathbf{l})\left(g(\mathbf{a},\mathbf{c})g(\mathbf{b},\mathbf{d}) - g(\mathbf{b},\mathbf{c})g(\mathbf{a},\mathbf{d})\right) \\ &+ \left(g(\mathbf{a},\mathbf{l})g(\mathbf{b},\mathbf{c}) - g(\mathbf{b},\mathbf{l})g(\mathbf{a},\mathbf{c})\right)g(\mathbf{d},\mathbf{l}) \\ &+ \left(g(\mathbf{b},\mathbf{l})g(\mathbf{a},\mathbf{d}) - g(\mathbf{a},\mathbf{l})g(\mathbf{b},\mathbf{d})\right)g(\mathbf{c},\mathbf{l}). \end{aligned} \tag{C.5}$$

Note (out of curiosity) that the last six terms of the expression above can be rewritten more concisely as the determinant of a certain 3×3 matrix,

$$\begin{aligned} e^{-2\lambda}\tilde{R}(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d}) &= R(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d}) \\ &+ H_\lambda(\mathbf{a},\mathbf{c})g(\mathbf{b},\mathbf{d}) - H_\lambda(\mathbf{a},\mathbf{d})g(\mathbf{b},\mathbf{c}) \\ &- H_\lambda(\mathbf{b},\mathbf{c})g(\mathbf{a},\mathbf{d}) + H_\lambda(\mathbf{b},\mathbf{d})g(\mathbf{a},\mathbf{c}) \\ &+ \det\begin{vmatrix} g(\mathbf{a},\mathbf{c}) & g(\mathbf{b},\mathbf{c}) & g(\mathbf{l},\mathbf{c}) \\ g(\mathbf{a},\mathbf{d}) & g(\mathbf{b},\mathbf{d}) & g(\mathbf{l},\mathbf{d}) \\ g(\mathbf{a},\mathbf{l}) & g(\mathbf{b},\mathbf{l}) & g(\mathbf{l},\mathbf{l}) \end{vmatrix}. \end{aligned}$$

Now we can compute the Ricci tensor, assuming that the spacetime has N dimensions. We use the properties

$$\begin{aligned} \mathrm{Tr}_{(\mathbf{a},\mathbf{b})} g(\mathbf{a},\mathbf{b}) &= N, \\ \mathrm{Tr}_{(\mathbf{a},\mathbf{b})} H_\lambda(\mathbf{a},\mathbf{b}) &\equiv \mathrm{Tr}_{(\mathbf{a},\mathbf{b})} g(\nabla_{\mathbf{a}}\mathbf{l},\mathbf{b}) \equiv \mathrm{div}\,\mathbf{l} \equiv \Box\lambda, \\ \mathrm{Tr}_{(\mathbf{a},\mathbf{b})} g(\mathbf{a},\mathbf{x})F(\mathbf{b},\mathbf{c},\ldots) &= F(\mathbf{x},\mathbf{c},\ldots). \end{aligned}$$

The trace Tr is always performed with respect to the *old* metric g, so the new Ricci tensor is $e^{-2\lambda}$ times the trace Tr of the new Riemann tensor. In this way we derive the required expressions for $\tilde{\mathrm{Ric}}$ and $\tilde{R}$. First we use Eq. (C.5) to obtain

$$\begin{aligned} \tilde{\mathrm{Ric}}\,(\mathbf{a},\mathbf{c}) &= e^{-2\lambda}\mathrm{Tr}_{(\mathbf{b},\mathbf{d})}\tilde{R}(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d}) \\ &= \mathrm{Ric}\,(\mathbf{a},\mathbf{c}) \\ &\quad + (N-2)\,H_\lambda(\mathbf{a},\mathbf{c}) + g(\mathbf{a},\mathbf{c})\Box\lambda \\ &\quad + g(\mathbf{l},\mathbf{l})\left(g(\mathbf{a},\mathbf{c})N - g(\mathbf{a},\mathbf{c})\right) \\ &\quad + \left(g(\mathbf{a},\mathbf{l})g(\mathbf{l},\mathbf{c}) - g(\mathbf{l},\mathbf{l})g(\mathbf{a},\mathbf{c})\right) \\ &\quad + \left(g(\mathbf{a},\mathbf{l}) - g(\mathbf{a},\mathbf{l})N\right)g(\mathbf{c},\mathbf{l}) \\ &= \mathrm{Ric}\,(\mathbf{a},\mathbf{c}) + (N-2)\,H_\lambda(\mathbf{a},\mathbf{c}) + g(\mathbf{a},\mathbf{c})\Box\lambda \\ &\quad + (N-2)\left[g(\mathbf{l},\mathbf{l})g(\mathbf{a},\mathbf{c}) - g(\mathbf{a},\mathbf{l})g(\mathbf{c},\mathbf{l})\right]. \end{aligned}$$

This is the required formula for the new Ricci tensor; the new Ricci scalar is obtained straightforwardly by taking the trace of the new Ricci tensor. ■

Proof of Statement 1.9.5.1 on page 53: We would like to lift the restrictions on the vectors $\mathbf{v},\mathbf{x},\mathbf{y}$. For a fixed $\mathbf{v}$, the function $R(\mathbf{v},\mathbf{x},\mathbf{v},\mathbf{y})$ is a symmetric bilinear form of $\mathbf{x},\mathbf{y}$. This bilinear form can be determined for spacelike $\mathbf{x},\mathbf{y}$ by a finite number of values (say, with $\mathbf{x}$ and $\mathbf{y}$ being vectors of a spacelike basis). The requirement that vectors $\mathbf{x},\mathbf{y}$ be spacelike and orthogonal to $\mathbf{v}$ is then easily lifted since $R(\mathbf{v},\mathbf{x},\ldots)$ is linear and antisymmetric in $\mathbf{v},\mathbf{x}$ and hence

$$R(\mathbf{v},\mathbf{x},\mathbf{v},\mathbf{y}) = R(\mathbf{v},\mathbf{x}+\lambda\mathbf{v},\mathbf{v},\mathbf{y}+\mu\mathbf{v})$$

for any λ,μ. Thus we can deduce $R(\mathbf{v},\mathbf{x},\mathbf{v},\mathbf{y})$ for arbitrary $\mathbf{x},\mathbf{y}$. Subsequently, we can regard $R(\mathbf{v},\mathbf{x},\mathbf{v},\mathbf{y})$ for fixed $\mathbf{x},\mathbf{y}$ as a quadratic form

$$Q_{\mathbf{x},\mathbf{y}}(\mathbf{v},\mathbf{v}) \equiv R(\mathbf{v},\mathbf{x},\mathbf{v},\mathbf{y}). \tag{C.6}$$

A standard trick in linear algebra is to recover a symmetric bilinear form $A(\mathbf{a},\mathbf{b})$ if the quadratic form $A(\mathbf{v},\mathbf{v})$ is known:

$$A(\mathbf{a},\mathbf{b}) = \frac{A(\mathbf{a}+\mathbf{b},\mathbf{a}+\mathbf{b}) - A(\mathbf{a},\mathbf{a}) - A(\mathbf{b},\mathbf{b})}{2}.$$

Therefore, we can try to recover a symmetric bilinear form $Q_{\mathbf{x},\mathbf{y}}(\mathbf{a},\mathbf{b})$ from the quadratic form $Q_{\mathbf{x},\mathbf{y}}(\mathbf{v},\mathbf{v})$ as follows,

$$\begin{aligned} Q_{\mathbf{x},\mathbf{y}}(\mathbf{a},\mathbf{b}) \equiv &\frac{1}{2}\left(R(\mathbf{a}+\mathbf{b},\mathbf{x},\mathbf{a}+\mathbf{b},\mathbf{y})\right. \\ &\left. - R(\mathbf{a},\mathbf{x},\mathbf{a},\mathbf{y}) - R(\mathbf{b},\mathbf{x},\mathbf{b},\mathbf{y})\right). \end{aligned} \tag{C.7}$$

To recover $Q_{\mathbf{x},\mathbf{y}}(\mathbf{a},\mathbf{b})$ completely, it suffices to know the values of $Q_{\mathbf{x},\mathbf{y}}(\mathbf{a},\mathbf{b})$ for arbitrary basis vectors $\mathbf{a},\mathbf{b}$; in total, we need 10 different values (for fixed $\mathbf{x},\mathbf{y}$). However, we are allowed to measure only $Q_{\mathbf{x},\mathbf{y}}(\mathbf{v},\mathbf{v})$ for future-directed and timelike $\mathbf{v}$. We note that the sum of two such vectors, $\mathbf{v}_1+\mathbf{v}_2$, is again future-directed and timelike. So we may deduce $Q_{\mathbf{x},\mathbf{y}}(\mathbf{a},\mathbf{b})$ for any two future-directed, timelike $\mathbf{a}$ and $\mathbf{b}$ using Eq. (C.7). We also note that a basis in the four-dimensional space can be found as a linearly independent set of 4 future-directed, timelike vectors. Such a basis will not be orthogonal, but one does not need an orthogonal basis to determine a bilinear form. Hence, for fixed $\mathbf{x},\mathbf{y}$ the bilinear form $Q_{\mathbf{x},\mathbf{y}}(\mathbf{a},\mathbf{b})$ can be determined by finitely many measurements. This process needs to be repeated for 10 different combinations of $\mathbf{x},\mathbf{y}$. We are thus able to deduce the values $Q_{\mathbf{x},\mathbf{y}}(\mathbf{a},\mathbf{b})$ for *arbitrary* vectors $\mathbf{a},\mathbf{b},\mathbf{x},\mathbf{y}$ if we measure $R(\mathbf{v},\mathbf{x},\mathbf{v},\mathbf{y})$ for sufficiently many future-directed timelike vectors $\mathbf{v}$ and spacelike vectors $\mathbf{x},\mathbf{y}$ (each time orthogonal to $\mathbf{v}$).

So far we recovered the bilinear form $Q_{\mathbf{x},\mathbf{y}}(\mathbf{a},\mathbf{b})$, which is symmetric in $\mathbf{a},\mathbf{b}$ and also in $\mathbf{x},\mathbf{y}$. Despite the suggestive Eq. (C.6), we must note that $Q_{\mathbf{x},\mathbf{y}}(\mathbf{a},\mathbf{b}) \neq R(\mathbf{a},\mathbf{x},\mathbf{b},\mathbf{y})$. Consider $R(\mathbf{a},\mathbf{x},\mathbf{b},\mathbf{y})$ for fixed $\mathbf{x},\mathbf{y}$ as a bilinear form of $\mathbf{a},\mathbf{b}$. This bilinear form is not necessarily symmetric in $\mathbf{a},\mathbf{b}$. In general, this bilinear form has a symmetric part,

$$\frac{1}{2}\left(R(\mathbf{a},\mathbf{x},\mathbf{b},\mathbf{y}) + R(\mathbf{b},\mathbf{x},\mathbf{a},\mathbf{y})\right),$$

and an antisymmetric part,

$$\frac{1}{2}\left(R(\mathbf{a},\mathbf{x},\mathbf{b},\mathbf{y}) - R(\mathbf{b},\mathbf{x},\mathbf{a},\mathbf{y})\right).$$

The quadratic form $Q_{\mathbf{x},\mathbf{y}}(\mathbf{v},\mathbf{v})$ is defined through Eq. (C.6) and knows only about the symmetric part of $R(\ldots)$ as shown above. Therefore, we recover precisely the symmetric part of $R(\ldots)$ when we determine the bilinear form $Q_{\mathbf{x},\mathbf{y}}(\mathbf{a},\mathbf{b})$. In other words, $Q_{\mathbf{x},\mathbf{y}}(\mathbf{a},\mathbf{b})$ is related to $R(\mathbf{a},\mathbf{x},\mathbf{b},\mathbf{y})$ through symmetrization in $(\mathbf{a},\mathbf{b})$,

$$Q_{\mathbf{x},\mathbf{y}}(\mathbf{a},\mathbf{b}) = \frac{1}{2}\left(R(\mathbf{a},\mathbf{x},\mathbf{b},\mathbf{y}) + R(\mathbf{b},\mathbf{x},\mathbf{a},\mathbf{y})\right). \tag{C.8}$$

To recover the full Riemann tensor, it remains to extract the antisymmetric part. A property of R not yet used is the first Bianchi identity (1.67), which we apply to the last term in Eq. (C.8) and obtain

$$\begin{aligned} 2Q_{\mathbf{x},\mathbf{y}}(\mathbf{a},\mathbf{b}) &= R(\mathbf{a},\mathbf{x},\mathbf{b},\mathbf{y}) - R(\mathbf{x},\mathbf{a},\mathbf{b},\mathbf{y}) - R(\mathbf{a},\mathbf{b},\mathbf{x},\mathbf{y}) \\ &= -2R(\mathbf{a},\mathbf{x},\mathbf{y},\mathbf{b}) + R(\mathbf{a},\mathbf{b},\mathbf{y},\mathbf{x}). \end{aligned}$$

[1] Note that the Landau-Lifshitz convention has the opposite sign of $R_{\alpha\beta}$ (but Ludvigsen has the same sign).

We now need to express this somehow through Q. Note that a suitable combination is

$$2Q_{\mathbf{b},\mathbf{x}}(\mathbf{a},\mathbf{y}) = R(\mathbf{a},\mathbf{b},\mathbf{y},\mathbf{x}) + R(\mathbf{a},\mathbf{x},\mathbf{y},\mathbf{b}).$$

Hence

$$2Q_{\mathbf{x},\mathbf{y}}(\mathbf{a},\mathbf{b}) = -3R(\mathbf{a},\mathbf{x},\mathbf{y},\mathbf{b}) + 2Q_{\mathbf{b},\mathbf{x}}(\mathbf{a},\mathbf{y});$$
$$R(\mathbf{a},\mathbf{x},\mathbf{b},\mathbf{y}) = \frac{2}{3}\left(Q_{\mathbf{x},\mathbf{y}}(\mathbf{a},\mathbf{b}) - Q_{\mathbf{b},\mathbf{x}}(\mathbf{a},\mathbf{y})\right).$$

Thus all the values of $R(...)$ can be deduced from the given experimental data. ■

C.2 For Chapter 3

Details for Calculation 3.1.3.3 on page 76: Suppose that the vector field $\mathbf{u}$ is the 4-velocity of the stationary observers, so

$$\mathbf{u} = \partial_t = \frac{1}{z}\mathbf{k}, \quad z = e^{Ht}.$$

Denote by $\mathbf{v}$ the 4-velocity of the particle. Then

$$g(\mathbf{v},\mathbf{v}) = g(\mathbf{u},\mathbf{u}) = 1.$$

The relative 3-velocity $\delta\vec{v}$ is related to the "relativistic gamma-factor" by

$$\gamma = \frac{1}{\sqrt{1-(\delta\vec{v})^2}} \geq 1;$$

on the other hand, $\gamma = g(\mathbf{u},\mathbf{v})$. Therefore, we need to compute the change of the "gamma-factor" γ along the worldline of the particle. We first note that for arbitrary $\mathbf{x},\mathbf{y}$ the property

$$g(\nabla_{\mathbf{x}}\mathbf{k},\mathbf{y}) + g(\nabla_{\mathbf{y}}\mathbf{k},\mathbf{x}) = (\mathcal{L}_{\mathbf{k}}g)\circ(\mathbf{x},\mathbf{y}) = 2He^{Ht}g(\mathbf{x},\mathbf{y})$$

holds. Thus, assuming $\nabla_{\mathbf{v}}\mathbf{v} = 0$, we have

$$\nabla_{\mathbf{v}}g(\mathbf{k},\mathbf{v}) = g(\nabla_{\mathbf{v}}\mathbf{k},\mathbf{v}) = He^{Ht}g(\mathbf{v},\mathbf{v}) = He^{Ht}.$$

However, we need to compute $\nabla_{\mathbf{v}}g(\mathbf{u},\mathbf{v})$, while $g(\mathbf{u},\mathbf{v})$ differs from $g(\mathbf{k},\mathbf{v})$ by a function of t only. So we need to compute the derivative

$$\nabla_{\mathbf{v}}t \equiv \mathbf{v}\circ t = g(\mathbf{v},\partial_t) = g(\mathbf{v},\mathbf{u}) \equiv \gamma.$$

Then we can write

$$\begin{aligned}\nabla_{\mathbf{v}}\gamma = \nabla_{\mathbf{v}}g(\mathbf{u},\mathbf{v}) &= \nabla_{\mathbf{v}}\left[e^{-Ht}g(\mathbf{k},\mathbf{v})\right]\\ &= \left(\nabla_{\mathbf{v}}e^{-Ht}\right)e^{Ht}\gamma + e^{-Ht}He^{Ht}\\ &= \left(1-\gamma^2\right)H.\end{aligned}$$

Denote by τ the proper time parameter along the worldline of the particle. Then we obtain the differential equation

$$\nabla_{\mathbf{v}}\gamma = \frac{d\gamma}{d\tau} = \left(1-\gamma^2\right)H.$$

The general solution, $\gamma(\tau) = \coth H(\tau-\tau_0)$, exponentially approaches $\gamma = 1$ at late times. In the Newtonian limit,

$$\gamma \approx 1 - \frac{1}{2}(\delta\vec{v})^2,$$

the equation for $\gamma(\tau)$ becomes

$$\delta\vec{v}\cdot\frac{d}{d\tau}\delta\vec{v} = -H(\delta\vec{v})^2.$$

It is easy to see that a geodesic is a straight line in the three-dimensional space. Thus the "force of friction" acting on the particle is directed opposite to the velocity $\delta\vec{v}$. Then, in the framework of Newton's second law, the force vector can be expressed as

$$\vec{F} = m\frac{d}{d\tau}\delta\vec{v} = -H\delta\vec{v}.$$

■

Proof of Statement 3.1.3.4 on page 76: We need to compute

$$\begin{aligned}\frac{d\gamma}{d\tau} \equiv \nabla_{\mathbf{v}}g(\mathbf{u},\mathbf{v}) &= g(\nabla_{\mathbf{v}}\left(z^{-1}\mathbf{k}\right),\mathbf{v})\\ &= -g(\mathbf{k},\mathbf{v})z^{-2}\nabla_{\mathbf{v}}z + z^{-1}g(\nabla_{\mathbf{v}}\mathbf{k},\mathbf{v})\\ &= \frac{g(\nabla_{\mathbf{v}}\mathbf{k},\mathbf{v}) - g(\mathbf{u},\mathbf{v})\nabla_{\mathbf{v}}z}{z}.\end{aligned}$$

Since $z \equiv \sqrt{g(\mathbf{k},\mathbf{k})}$, we have

$$\nabla_{\mathbf{v}}z = \frac{1}{2z}\nabla_{\mathbf{v}}g(\mathbf{k},\mathbf{k}) = \frac{1}{z}g(\nabla_{\mathbf{v}}\mathbf{k},\mathbf{k}).$$

So it remains to compute the bilinear form $g(\nabla_{\mathbf{x}}\mathbf{k},\mathbf{y})$. If the vector field $\mathbf{k}$ is integrable then $\hat{g}\mathbf{k}$ is an exact 1-form and the first term the Koszul formula,

$$g(\nabla_{\mathbf{x}}\mathbf{k},\mathbf{y}) = \left(\frac{1}{2}\mathrm{d}\hat{g}\mathbf{k} + \frac{1}{2}\mathcal{L}_{\mathbf{k}}g\right)\circ(\mathbf{x},\mathbf{y}),$$

cancels (here $\mathbf{x},\mathbf{y}$ are arbitrary vectors). For an integrable conformal Killing vector $\mathbf{k}$, we therefore get

$$g(\nabla_{\mathbf{x}}\mathbf{k},\mathbf{y}) = \frac{1}{2}(\mathcal{L}_{\mathbf{k}}g)\circ(\mathbf{x},\mathbf{y}) = \lambda g(\mathbf{x},\mathbf{y}).$$

Using this property, we compute

$$\begin{aligned}\frac{d\gamma}{d\tau} &= \frac{\lambda g(\mathbf{v},\mathbf{v}) - \gamma z^{-1}\lambda g(\mathbf{v},\mathbf{k})}{z}\\ &= \left(1-\gamma^2\right)\lambda z^{-1}.\end{aligned}$$

■

Proof of Statement 3.2.1.1 on page 79: **(a)** Consider a congruence of null curves in a two-dimensional spacetime, and denote by $\mathbf{u}$ the corresponding tangent vector field. Since $g(\mathbf{u},\mathbf{u}) = 0$, we have

$$g(\mathbf{u},\nabla_{\mathbf{u}}\mathbf{u}) = 0.$$

In two dimensions, there is only a one-dimensional subspace of vectors orthogonal to a null vector $\mathbf{u}$, namely the space of vectors parallel to $\mathbf{u}$ itself. Thus $\nabla_{\mathbf{u}}\mathbf{u}$ is parallel to $\mathbf{u}$ itself, so there exists a scalar function μ such that

$$\nabla_{\mathbf{u}}\mathbf{u} = \mu\mathbf{u}.$$

This means that the orbits of $\mathbf{u}$ are geodesic (but perhaps not affinely parameterized). A geodesic null vector field can be found as $\alpha\mathbf{u}$, where α is a scalar function satisfying

$$\nabla_{\mathbf{u}}\alpha + \mu\alpha = 0.$$

Such a function α always exists since it is specified by a differential equation along the orbits of $\mathbf{u}$. Since $\nabla_{\alpha\mathbf{u}}(\alpha\mathbf{u}) = 0$, so the orbits of $\alpha\mathbf{u}$ (which are the same curves as the orbits of $\mathbf{u}$ but perhaps differently parameterized) are geodesic curves.

(b) Radial null geodesics in the flat metric g are lines of either constant u or constant v (and constant θ, ϕ). Since the new metric is

$$\tilde{g} = \frac{1}{2}e^{2\lambda}\left(du \otimes dv + dv \otimes du\right) - e^{2\lambda}\frac{(u-v)^2}{4}dS^2,$$

while $dS^2 = 0$ if $\theta, \phi = \text{const}$, it is clear that lines of constant u or v (and constant θ, ϕ) will remain *null* curves also according to the metric $\tilde{g}$. Since the metric is independent of θ, ϕ, lines of constant θ, ϕ will be geodesics if they are geodesics in the two-dimensional spacetime with metric

$$\frac{1}{2}e^{2\lambda}\left(du \otimes dv + dv \otimes du\right)$$

and coordinates (u, v) at constant θ, ϕ. But in a two-dimensional spacetime *any* null curve is a null geodesic, according to part (a) of this statement. Thus, we find that the radial null curves are also null geodesics in the metric $\tilde{g}$. ■

C.3 For Chapter 6

Proof of Statement 6.1.3.1 on page 131: Let $\{\mathbf{e}_a\}$ be an orthonormal basis and $\{\theta^a\}$ the corresponding dual basis. Since the Hodge duality map $\omega \mapsto *\omega$ is linear in ω, it suffices to consider ω of the form $\omega = \chi_1 \wedge ... \wedge \chi_n$, where χ_j are some 1-forms. Moreover, it is sufficient to consider the case when all χ_j are *basis* 1-forms θ^a. Further, it is sufficient to treat the case when $\chi_1 = \theta^1, ..., \chi_n = \theta^n$. Thus we only need to consider the n-form $\omega = \theta^1 \wedge ... \wedge \theta^n$. Using the definition $\varepsilon = \theta^1 \wedge ... \wedge \theta^N$ and the fact that

$$\iota_{\mathbf{e}_n}...\iota_{\mathbf{e}_1}\left(\theta^1 \wedge ... \wedge \theta^n\right) = 1$$

is (up to permutations) the only nonzero expression of the type

$$\iota_{\mathbf{e}_a}\iota_{\mathbf{e}_b}...\iota_{\mathbf{e}_c}\left(\theta^1 \wedge ... \wedge \theta^n\right),$$

we can compute $**\omega$ explicitly:

$$\begin{aligned}
*\omega &= *(\theta^1 \wedge ... \wedge \theta^n) = \eta^{11}...\eta^{nn}\left(\iota_{\mathbf{e}_n}...\iota_{\mathbf{e}_1}\varepsilon\right) \\
&= \eta^{11}...\eta^{nn}\theta^{n+1} \wedge ... \wedge \theta^N; \\
**\omega &= *(\theta^{n+1} \wedge ... \wedge \theta^N)\eta^{11}...\eta^{nn} \\
&= \eta^{11}...\eta^{NN}\left(\iota_{\mathbf{e}_N}...\iota_{\mathbf{e}_{n+1}}\varepsilon\right) = (\det\eta_{ab})\,(-1)^{(N-n)n}\,\omega.
\end{aligned}$$

The sign factor $(-1)^{(N-n)n}$ appears in the identity

$$\begin{aligned}
\iota_{\mathbf{e}_N}...\iota_{\mathbf{e}_{n+1}}\varepsilon &= \iota_{\mathbf{e}_N}...\iota_{\mathbf{e}_{n+1}}\left(\theta^1 \wedge ...\theta^n \wedge \theta^{n+1} \wedge ... \wedge \theta^N\right) \\
&= (-1)^{(N-n)n}\,\theta^1 \wedge ... \wedge \theta^n
\end{aligned}$$

due to the necessity to carry $(N-n)$ vectors $\{\mathbf{e}_N, ..., \mathbf{e}_{n+1}\}$ through the initial group of n 1-forms $\left\{\theta^1, ..., \theta^n\right\}$. ■

Proof of Statement 6.1.3.2 on page 131: **(a)** We will use the definition of $*\omega$ through an orthonormal basis $\{\mathbf{e}_a\}$. We compute the N-form

$$\omega_1 \wedge *\omega_2 = \sum_{k_1,...,k_n}\frac{\eta^{k_1k_1}...\eta^{k_nk_n}}{n!}\left(\iota_{\mathbf{e}_{k_n}}...\iota_{\mathbf{e}_{k_1}}\omega_2\right)\omega_1 \wedge \left(\iota_{\mathbf{e}_{k_n}}...\iota_{\mathbf{e}_{k_1}}\varepsilon\right);$$

the first task is to show that this expression is symmetric with respect to the exchange $\omega_1 \leftrightarrow \omega_2$. We use Eq. (6.6) to compute the Hodge dual of this N-form,

$$*(\omega_1 \wedge *\omega_2) = (\det\eta_{ab})\,(\omega_1 \wedge *\omega_2) \circ (\mathbf{e}_1, ..., \mathbf{e}_N).$$

As an intermediate step, we need to compute

$$\begin{aligned}
&\left[\omega_1 \wedge \left(\iota_{\mathbf{e}_{k_n}}...\iota_{\mathbf{e}_{k_1}}\varepsilon\right)\right] \circ (\mathbf{e}_1, ..., \mathbf{e}_N) \\
&= \sum_\sigma \frac{(-1)^{|\sigma|}}{n!\,(N-n)!}\omega_1(\mathbf{e}_{\sigma(1)}, ..., \mathbf{e}_{\sigma(n)})\varepsilon(\mathbf{e}_{k_1}, ..., \mathbf{e}_{k_n}, \mathbf{e}_{\sigma(n+1)}, ..., \mathbf{e}_{\sigma(N)}).
\end{aligned} \tag{C.9}$$

In the expression (C.9), the summation is performed over all permutations σ of the set $\{1, ..., N\}$. We note that

$$\varepsilon(\mathbf{e}_{k_1}, ..., \mathbf{e}_{k_n}, \mathbf{e}_{\sigma(n+1)}, ..., \mathbf{e}_{\sigma(N)}) = 0$$

unless the set $\{k_1, ..., k_n, \sigma(n+1), ..., \sigma(N)\}$ is also a permutation of the same kind. It follows that a nonzero contribution is possible only if the set $\{k_1, ..., k_n\}$ is a permutation of the set $\{\sigma(1), ..., \sigma(n)\}$. The expression for $*(\omega_1 \wedge *\omega_2)$ contains a sum over the indices k_1, ..., k_n as well as a sum over all permutations σ,

$$\begin{aligned}
*\left(\omega_1 \wedge *\omega_2\right) &= (\det\eta_{ab})\sum_\sigma\sum_{k_1,...,k_n}\frac{\eta^{k_1k_1}...\eta^{k_nk_n}}{n!}\omega_2(\mathbf{e}_{k_1}, ..., \mathbf{e}_{k_n}) \\
&\times \frac{(-1)^{|\sigma|}}{n!\,(N-n)!}\omega_1(\mathbf{e}_{\sigma(1)}, ..., \mathbf{e}_{\sigma(n)}) \\
&\times \varepsilon(\mathbf{e}_{k_1}, ..., \mathbf{e}_{k_n}, \mathbf{e}_{\sigma(n+1)}, ..., \mathbf{e}_{\sigma(N)}).
\end{aligned} \tag{C.10}$$

For a given permutation σ, there are $n!$ sets $\{k_1, ..., k_n\}$ that give a nonzero contribution because they are permutations of $\{\sigma(1), ..., \sigma(n)\}$. The contributions of these $n!$ sets to Eq. (C.10) are all equal, hence the sum over $\{k_1, ..., k_n\}$ can be replaced by an extra factor $n!$, while the indices k_1, ..., k_n may be relabeled as $\sigma(1)$, ..., $\sigma(n)$. So we find

$$\begin{aligned}
*\left(\omega_1 \wedge *\omega_2\right) &= (\det\eta_{ab})\sum_\sigma\frac{(-1)^{|\sigma|}}{n!\,(N-n)!}\eta^{\sigma(1)\sigma(1)}...\eta^{\sigma(n)\sigma(n)} \\
&\times \omega_2(\mathbf{e}_{\sigma(1)}, ..., \mathbf{e}_{\sigma(n)})\omega_1(\mathbf{e}_{\sigma(1)}, ..., \mathbf{e}_{\sigma(n)})\varepsilon(\mathbf{e}_{\sigma(1)}, ..., \mathbf{e}_{\sigma(N)}).
\end{aligned}$$

The expression above is manifestly symmetric under the exchange $\omega_1 \leftrightarrow \omega_2$. This already proves the first part of the statement **(a)**, but it is useful to derive a simplified form of the expression $*(\omega_1 \wedge *\omega_2)$. We note that

$$(-1)^{|\sigma|}\,\varepsilon(\mathbf{e}_{\sigma(1)}, ..., \mathbf{e}_{\sigma(N)}) = 1$$

for any permutation σ, and hence

$$\begin{aligned}
*\left(\omega_1 \wedge *\omega_2\right) &= (\det\eta_{ab})\sum_\sigma\frac{\eta^{\sigma(1)\sigma(1)}...\eta^{\sigma(n)\sigma(n)}}{n!\,(N-n)!} \\
&\times \omega_2(\mathbf{e}_{\sigma(1)}, ..., \mathbf{e}_{\sigma(n)})\omega_1(\mathbf{e}_{\sigma(1)}, ..., \mathbf{e}_{\sigma(n)}).
\end{aligned}$$

Since the expression above is independent of the permutation elements $\{\sigma(n+1), ..., \sigma(N)\}$, we may sum over them, introducing an extra factor $(N-n)!$. Finally, we relabel the indices $\sigma(1), ..., \sigma(n)$ as $k_1, ..., k_n$ and obtain the formula

$$*(\omega_1 \wedge *\omega_2) = \frac{(\det \eta_{ab})}{n!} \sum_{k_1,...,k_n} \eta^{k_1 k_1} ... \eta^{k_n k_n} \times \omega_1(\mathbf{e}_{k_1}, ..., \mathbf{e}_{k_n}) \omega_2(\mathbf{e}_{k_1}, ..., \mathbf{e}_{k_n}). \tag{C.11}$$

Using this explicit formula, it is straightforward to show that the basis n-forms $\theta^{a_1} \wedge ... \wedge \theta^{a_n}$ comprise an orthonormal basis with respect to the scalar product $*(\omega_1 \wedge *\omega_2)$.

(b) We rewrite the formula (C.11) without an explicit choice of basis. We note that a summation over basis vectors $\{\mathbf{e}_k\}$ multiplied by η^{kk} is equivalent to taking the trace, e.g.

$$\mathrm{Tr}_{(\mathbf{a},\mathbf{b})} X(\mathbf{a}, \mathbf{b}) = \sum_k \eta^{kk} X(\mathbf{e}_k, \mathbf{e}_k).$$

Hence, Eq. (C.11) becomes

$$* (\omega_1 \wedge *\omega_2) = \frac{(\det \eta_{ab})}{n!} \mathrm{Tr}_{(\mathbf{a}_1,\mathbf{b}_1)...(\mathbf{a}_n,\mathbf{b}_n)} \omega_1(\mathbf{a}_1, ..., \mathbf{a}_n) \omega_2(\mathbf{b}_1, ..., \mathbf{b}_n).$$

This is equivalent to Eq. (6.7).

(c) Using the explicit formula (6.7), we find

$$\omega_1 \wedge *\omega_2 = \mathrm{Vol}\, \mathrm{Tr}_{(\mathbf{a},\mathbf{b})} \omega_1(\mathbf{a}) \omega_2(\mathbf{b}).$$

The trace is computed as

$$\mathrm{Tr}_{(\mathbf{a},\mathbf{b})} \omega_1(\mathbf{a}) \omega_2(\mathbf{b}) = \mathrm{Tr}_{(\mathbf{a},\mathbf{b})} g(\mathbf{v}_1, \mathbf{a}) g(\mathbf{v}_2, \mathbf{b}) = g(\mathbf{v}_1, \mathbf{v}_2).$$

(d) We use Eq. (6.7) and obtain

$$\begin{aligned} \Omega \wedge *(\omega_1 \wedge \omega_2) &= \frac{\mathrm{Vol}}{2!} \mathrm{Tr}_{(\mathbf{a}_1,\mathbf{b}_1)(\mathbf{a}_2,\mathbf{b}_2)} \Omega(\mathbf{a}_1, \mathbf{a}_2) (\omega_1 \wedge \omega_2) \circ (\mathbf{b}_1, \mathbf{b}_2) \\ &= \mathrm{Vol}\, \mathrm{Tr}_{(\mathbf{a}_1,\mathbf{b}_1)(\mathbf{a}_2,\mathbf{b}_2)} \Omega(\mathbf{a}_1, \mathbf{a}_2) \omega_1(\mathbf{b}_1) \omega_2(\mathbf{b}_2) \\ &= \mathrm{Vol}\, \mathrm{Tr}_{(\mathbf{a}_1,\mathbf{b}_1)(\mathbf{a}_2,\mathbf{b}_2)} \Omega(\mathbf{a}_1, \mathbf{a}_2) g(\mathbf{v}_1, \mathbf{b}_1) g(\mathbf{v}_2, \mathbf{b}_2) \\ &= \mathrm{Vol}\, \Omega(\mathbf{v}_1, \mathbf{v}_2). \end{aligned}$$

■

Statement 6.1.3.2 on page 131: Since $\omega \wedge *\Omega$ is a 3-form, there exists a 1-form ϕ such that

$$\omega \wedge *\Omega = *\phi.$$

Consider an arbitrary 1-form χ and the expression $\chi \wedge *\phi$. We will use Statement 6.1.3.2. On the one hand,

$$\chi \wedge *\phi = g^{-1}(\chi, \phi)\, \mathrm{Vol}.$$

On the other hand,

$$\begin{aligned} \chi \wedge *\phi = \chi \wedge \omega \wedge *\Omega &= \Omega \wedge * (\chi \wedge \omega) \\ &= \Omega(\hat{g}^{-1}\chi, \hat{g}^{-1}\omega)\, \mathrm{Vol} = -(\iota_{\mathbf{v}} \iota_{\mathbf{x}} \Omega)\, \mathrm{Vol}, \end{aligned}$$

where $\mathbf{x} \equiv \hat{g}^{-1}\chi$ and $\mathbf{v} \equiv \hat{g}^{-1}\omega$. We obtain the relationship

$$g^{-1}(\chi, \phi) = \Omega(\hat{g}^{-1}\chi, \hat{g}^{-1}\omega)$$

that holds for all 1-forms χ, or equivalently

$$\iota_{\mathbf{x}} \phi = -\iota_{\mathbf{x}} \iota_{\mathbf{v}} \Omega$$

that holds for all vectors $\mathbf{x}$. It follows that $\phi = -\iota_{\mathbf{v}} \Omega$, so $\omega \wedge *\Omega = - * (\iota_{\mathbf{v}} \Omega)$. ■

Proof of Statement 6.1.6.1 on page 134: The formula for $\chi_{a\,bc}$ can be guessed in the following way. One notes that $\chi_{a\,bc}$ should be a linear combination of $A_{a\,bc}$, with some indices permuted. Due to the antisymmetry of $A_{a\,bc}$, there are only three possible permutations (namely with either a, b, or c as the first index). Moreover, the antisymmetry of $\chi_{a\,bc}$ in (a, b) means that the only possibility is

$$\chi_{a\,bc} = (A_{a\,bc} - A_{b\,ac})\, \lambda + A_{c\,ab} \mu,$$

where λ, μ are unknown constants. A direct substitution then yields $\lambda = -\mu = \frac{1}{2}$.

It remains to show that this solution is unique. Suppose there exist two different solutions $\chi_{a\,b}$ and $\tilde{\chi}_{a\,b}$. The difference $\mathbf{X}_{a\,b} \equiv \tilde{\chi}_{a\,b} - \chi_{a\,b}$ satisfies the homogeneous equations

$$\sum_b \mathbf{X}_{a\,b} \wedge \theta^b = 0, \quad \mathbf{X}_{a\,b} = -\mathbf{X}_{b\,a}.$$

The 1-forms $\mathbf{X}_{a\,b}$ can be expanded in the dual frame basis $\{\theta^c\}$ as

$$\mathbf{X}_{a\,b} = \sum_c X_{a\,bc} \theta^c, \quad X_{a\,bc} = -X_{b\,ac}.$$

Then we can write

$$\sum_b \mathbf{X}_{a\,b} \wedge \theta^b = \sum_{b,c} X_{a\,bc} \theta^b \wedge \theta^c = 0.$$

It follows that $X_{a\,bc} = X_{a\,cb}$, i.e. $X_{a\,bc}$ is symmetric in b, c. However, it is impossible that a nonzero quantity $X_{a\,bc}$ is symmetric in b, c and at the same time antisymmetric in a, b. Namely,

$$X_{a\,bc} = -X_{b\,ac} = -X_{b\,ca} = X_{c\,ba} = X_{c\,ab} = -X_{a\,cb} = -X_{a\,bc},$$

so $X_{a\,bc} = 0$ for any a, b, c. Therefore, the difference between $\chi_{a\,b}$ and $\tilde{\chi}_{a\,b}$ equals zero, i.e. the solution $\chi_{a\,b}$ is unique. ■

Proof of Statement 6.1.6.4 on page 135: The overall shape of Eq. (6.26) can be guessed up to the values of the numerical coefficients. First we verify that the given expression for $\boldsymbol{B}_{a\,b}$ solves the required equations, checking the equivalence of the two given formulas, and then we investigate the uniqueness of the solution.

Let us compute $\sum_b \theta^b \wedge \boldsymbol{B}_{b\,a}$, where $\boldsymbol{B}_{b\,a}$ is given by Eq. (6.26). Using Statement 6.1.6.2 for n-forms, we find

$$\frac{1}{n} \sum_b \theta^b \wedge \iota_{\mathbf{e}_b} \boldsymbol{A}_a = \boldsymbol{A}_a,$$

$$\begin{aligned} \frac{1}{n} \sum_b \theta^b \wedge \iota_{\mathbf{e}_b} \left(\iota_{\mathbf{e}_a} \sum_c \theta^c \wedge \boldsymbol{A}_c \right) &= \iota_{\mathbf{e}_a} \sum_c \theta^c \wedge \boldsymbol{A}_c \\ &= \boldsymbol{A}_a - \sum_c \theta^c \wedge \iota_{\mathbf{e}_a} \boldsymbol{A}_c. \end{aligned}$$

Then it is straightforward to see that Eq. (6.25) holds. The equivalence with Eq. (6.27) follows by virtue of Eq. (6.24), which also holds for arbitrary n-forms $\boldsymbol{A}_a$.

We can demonstrate the non-uniqueness of the solution $\boldsymbol{B}_{b\,a}$. One may modify a given solution $\boldsymbol{B}_{b\,a}$ by adding an arbitrary $(n-1)$-form $\boldsymbol{X}_{b\,a}$ as long as

$$\sum_a \theta^a \wedge \boldsymbol{X}_{b\,a} = 0, \quad \boldsymbol{X}_{b\,a} = -\boldsymbol{X}_{a\,b}.$$

Unlike the case $n = 2$, there exist nontrivial $\boldsymbol{X}_{b\,a} \neq 0$ satisfying these conditions. Note that $\boldsymbol{X}_{b\,a}$ is an $(n-1)$-form while $n \geq 3$, so we may write for example

$$\boldsymbol{X}_{b\,a} = \boldsymbol{\xi}_b \wedge \boldsymbol{\xi}_a \wedge \boldsymbol{h},$$

where $\boldsymbol{h}$ is an arbitrary $(n-3)$-form and $\boldsymbol{\xi}_b$ is a 1-form such that $\sum_b \boldsymbol{\theta}^b \wedge \boldsymbol{\xi}_b = 0$. An example of such $\boldsymbol{\xi}_b$ is

$$\boldsymbol{\xi}_b \equiv \sum_c F_{bc}\boldsymbol{\theta}^c, \quad F_{bc} = F_{cb},$$

where F_{bc} is an arbitrary but symmetric array of coefficients. One can also add several such terms $\boldsymbol{X}_{b\,a}$ with different F_{bc} and $\boldsymbol{h}$ (this is not equivalent to adding one such term). Thus there is a considerable remaining freedom in selecting the $(n-1)$-forms $\boldsymbol{B}_{b\,a}$. Note that this freedom does not exist if $n = 2$. ■

D Comments on literature

D.1 Comments on Ludvigsen's *General Relativity*

The book is M. Ludvigsen, *General relativity: a geometric aproach* (Cambridge University Press, 1999). Many explanations in that book are outstandingly clear, and I benefited greatly by reading it. Nevertheless, there are some minor gaffes:

1) On p. 91, eq. 9.27 is supposedly the same as eq. 9.20 when "written in full." However, these equations actually differ by the choice of the permuted indices. The relation 9.27 can be obtained from 9.20 only if one assumes the identity $R_{abcd} = R_{cdab}$, which Ludvigsen actually never mentions in the book. This well-known standard identity is a consequence of 9.19 and 9.20.

2) On p. 103, eq. 10.14 contains $\nabla_a \nabla^a \phi = R...$, while the preceding (unnumbered) equation on p.102 contains $-\nabla_e \nabla^a \phi = R....$ A minus sign has materialized from nowhere! The answer (10.17) is correct, and the extra minus sign is actually needed to compensate for an error made earlier. In the last paragraph on p. 101, Ludvigsen writes (in 3-dimensional notation) $\mathbf{a} = \nabla\phi$ whereas in fact $\mathbf{a} = -\nabla\phi$ in Newtonian gravity. (The acceleration points down, the potential grows upwards.) So the correct calculation starts by introducing $a_a = \nabla_a \phi$ and not $-\nabla_a \phi$.

3) On p. 108, top line, "Using the fact that $l_a n^a = 1$ and $\nabla_a l_b = \nabla_b l_a$, we have..." — actually, the same result follows with merely the assumption that l_a is a null geodesic. It is not necessary to assume that l_a is integrable.

4) On p. 109, top line, — one cannot actually derive Eq. (11.10) as claimed. By contracting the top equation with $l^a m^b m^c$ and using Eq. (11.8), which already assumes that l is integrable, one gets

$$D\sigma = R(l, m, l, m) + 2l^a \bar{m}_b (\nabla_a m^b).$$

Now it is unclear how to show that the last term is equal to $2\rho\sigma$. There is a significant freedom in ∇m since m is chosen simply to be orthogonal to n, l and this is not sufficient to fix ∇m. In fact, the null tetrad can be changed by the transformation $m \to e^{i\lambda} m$. Then $\sigma \to e^{2i\lambda}\sigma$, as shown at top of page 110. If λ is a function of position then the equation $D\sigma = ...$ will be changed after the transformation! Thus this equation really depends on the choice of the tetrad, and some choices are better than others. The equation (11.10) is perhaps obtained with a suitable choice of the tetrad, but this is not discussed in the book.

Extra topics: Derivation of Schwarzschild uniquely from symmetry (Ludvigsen?) Definition of mass in spherically symmetric spacetime (see Frolov-Kofman?) ADM and Bondi mass? Positivity of energy theorem? Energy-momentum pseudotensors?

E License for this text

E.1 Author's position on commercial publishing

Thanks to modern technology, one can prepare an entire book electronically on a personal computer, in ready-to-print form. Sending an electronic book across the world takes at most a few minutes and costs about as much as a cup of tea. The author encourages everyone interested in reading the text to download and/or print it, in whole or in part. The two-column formatting of the text is designed to require the least possible amount of paper when printed. Everyone is also entitled to commission a print shop to produce bound copies of the text, in which case the single-column formatting may be preferred. The cost of one bound copy may be estimated as 10 to 20 US dollars.

A commercial publisher may want to offer professionally printed and bound copies for sale. Since this book is distributed with complete source, it will be a matter of minutes to reformat the book according to the taste or constraints of a particular publisher even without the author's assistance. The author welcomes commercial printing of the text, as long as the publisher adheres to the conditions of the license (the GNU FDL). Since the FDL disallows granting exclusive distribution rights, the author cannot sign a standard exclusive-rights contract with a publisher. However, the author will consider signing any publishing contract that leaves intact the conditions of the FDL.

E.2 GNU Free Documentation License

Version 1.2, November 2002

Preamble

The purpose of this License is to make a manual, textbook, or other functional and useful document free in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

E.2.0 Applicability and definitions

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "Document", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "you". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "Modified Version" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "Secondary Section" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "Invariant Sections" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "Cover Texts" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "Transparent" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "Opaque".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "Title Page" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

A section "Entitled XYZ" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "Acknowledgements", "Dedications", "Endorsements", or "History".) To "Preserve the Title" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

E.2.1 Verbatim copying

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section E.2.2.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

E.2.2 Copying in quantity

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

E.2.3 Modifications

You may copy and distribute a Modified Version of the Document under the conditions of sections E.2.1 and E.2.2 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.

B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.

C. State on the Title page the name of the publisher of the Modified Version, as the publisher.

D. Preserve all the copyright notices of the Document.

E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.

F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.

G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.

H. Include an unaltered copy of this License.

I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the

Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.

J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.

K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.

L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.

M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.

N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.

O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties—for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

Combining documents

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements."

Collections of documents

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

Aggregation with independent works

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section E.2.2 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

Translation

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section E.2.3. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warrany Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (sec-

tion E.2.3) to Preserve its Title (section E.2.0) will typically require changing the actual title.

Termination

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

Future revisions of this license

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See http://www.gnu.org/copyleft/.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.

ADDENDUM: How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

Copyright (c) <year> <your name>. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License".

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with...Texts." line with this:

with the Invariant Sections being <list their titles>, with the Front-Cover Texts being <list>, and with the Back-Cover Texts being <list>.

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.

Copyright

Bibliography

[1] V. I. Arnold, *Mathematical methods of classical mechanics* (Springer, NY, 1997). 16, 28

[2] P. Bamberg, S. Sternberg. A course in mathematics for students of physics (Cambridge, 1990). 23, 26

[3] A. Borde, A. Vilenkin, and A. H. Guth. *Inflationary spacetimes are not past-complete*. Phys. Rev. Lett. **90**, 151301 (2003); online preprint gr-qc/0110012 (2001). 92

[4] R. Bott and J. Milnor. *On the parallelizability of the spheres*. Bull. Amer. Math. Soc. **64**, 87 (1958). 141

[5] R. Bousso. *The Holographic Principle*. Rev. Mod. Phys. **75**, 825 (2002); online preprint hep-th/0203101 (2002). 101

[6] T. Eguchi, P. B. Gilkey, and E. J. Hanson. *Gravitation, gauge theories and differential geometry*. Phys. Rep. **66**, 213 (1980).

[7] J. Frauendiener. *Conformal infinity*. Living Rev. Rel. **7**, 1 (2004). Online article: cited Nov. 2005, www.livingreviews.org/lrr-2004-1 . 87

[8] R. P. Geroch. *Singularities*, in: *Relativity*, edited by M. Carmeli, S. I. Fickler, and L. Witten (Plenum Press, NY, 1970), p. 259. 91

[9] R. P. Geroch. *Asymptotic structure of spacetime*. In: Proc. Symp. Cincinnati, Ohio, June 14-18, 1976, ed. by F. P. Esposito and L. Witten (Plenum Press, NY, 1977), p. 1. 80, 87

[10] Ø. Grøn and S. Hervik. *Einstein's theory of relativity* (book draft, online at www.fys.uio.no/~sigbjorh/GRbook.html, dated December 2004). v

[11] S. Gudmundsson. *An introduction to Riemannian geometry* (book draft, online at www.matematik.lu.se/matematiklu/personal/sigma/, dated 2006). 1

[12] S. W. Hawking and G. F. R. Ellis. *The large-scale structure of space-time* (Cambridge, 1973). v, 91, 95, 99

[13] A. Hebecker. *Allgemeine Relativität* (lecture notes for the University of Heidelberg, online at res.kaelteschaden.de/ART.pdf, dated July 2007). 129

[14] R. Jackiw. *Constrained quantization without tears*. Online preprint arxiv:hep-th/9306075 (1993). 115

[15] A. Kempf. *General Relativity for cosmology* (lecture notes for AMATH875, online at www.math.uwaterloo.ca/~akempf/amath875.shtml, dated 2005). 139

[16] S. Lang. Introduction to differentiable manifolds (Springer, 2002). 23

[17] J. M. Lee. *Introduction to smooth manifolds* (Springer, 2003). 1

[18] J. Lee and R. M. Wald. *Local symmetries and constraints*. J. Math. Phys. **31** , 725 (1990). 115

[19] M. Ludvigsen. *General relativity: a geometric approach* (Cambridge, 1999). v, vi, 1

[20] P. W. Michor. *Topics in differential geometry* (lecture notes of a course given in Vienna, online at www.mat.univie.ac.at/~michor/dgbook.pdf, draft dated April 2007). 135

[21] C. Misner, K. Thorne, and J. Wheeler. *Gravitation* (Freeman, 1973). v, 45, 91, 127

[22] B. O'Neill. *Semi-Riemannian geometry* (Academic Press, 1983). 1

[23] R. Penrose. *Asymptotic properties of fields and spacetimes*. Phys. Rev. Lett. **10**, 66 (1963). 87

[24] R. Penrose. *Gravitational collapse and space-time singularities*. Phys. Rev. Lett. **14**, 57 (1965). 99

[25] R. Penrose. *Structure of spacetime*. In: *Battelle Rencontres*, ed. by C. M. DeWitt and J. A. Wheeler (Benjamin, NY, 1968), p. 121. 40

[26] R. Penrose. *Techniques of differential topology in relativity* (SIAM, Philadelphia, 1972). 91

[27] R. Penrose and W. Rindler. *Spinors and space-time* (Cambridge, 1988). 143, 155, 156

[28] E. Poisson. *A relativist's toolkit* (Cambridge, 2004). v, vi, 113

[29] B. F. Schutz. *A first course in general relativity* (Cambridge, 1985). v, 159

[30] N. Steenrod. *Topology of fibre bundles* (Princeton, 1951). 141

[31] S. Sternberg. *Semi-riemannian geometry and GR* (book draft, online at www.math.harvard.edu/~shlomo, dated September 2003).

[32] J. Stewart. *Advanced general relativity* (Cambridge, 1991). v, 143, 155

[33] N. Straumann. *General relativity and relativistic astrophysics* (Springer, 1984). v, 16, 23, 46, 127

[34] A. Trautman. *Conservation laws in general relativity*. In: *Gravitation: an introduction to current research*, ed. by L. Witten (Wiley, 1962), p. 169. 103

[35] A. Vilenkin. *Interpretation of the wave function of the universe*. Phys. Rev. D **39**, 1116 (1989). 122, 123

[36] R. M. Wald. *General relativity* (University of Chicago, 1984). v, 76, 80, 87, 91, 113, 115, 127, 154

[37] H. Whitney. *Differentiable manifolds*. Ann. of Math. **37**, 645 (1936). 6

[38] S. Winitzki. *Drawing conformal diagrams for a fractal landscape*. Phys. Rev. D **71**, 123523 (2005); online preprint arxiv.org/abs/gr-qc/0503061 (2005).

Index