General Relativity

— Lecture Notes —

Preliminary Version

February 4, 2020

Prof. Dr. Haye Hinrichsen

Lehrstuhl für Theoretische Physik III
Fakultät für Physik und Astronomie
Universität Würzburg, Germany

Summer term 2018

Contents

1 Mathematical foundations

1.1 Elements of group theory

1.1.1 Groups

A *group* is a set G equipped with an operation $* : G \times G \to G$ with the following properties:

(i) The execution order (the arrangement of parenthesis) does not matter, i.e., the map is *associative*:
$(a * b) * c = a * (b * c) \quad \forall a, b, c \in G$.

(ii) There exists a *neutral element* $e \in G$ such that $e * g = g * e = g$ for all $g \in G$.

(iii) For every $g \in G$ there exists an *inverse element* g^{-1} with $g * g^{-1} = e$.

(iv) If, in addition, the map is symmetric ($a * b = b * a$), the group is said to be *commutative* or *Abelian*.

A subset $U \subset G$ is called a *subgroup* of G if U in itself forms again a group. This means that for all $u, v \in U$ the result of $u * v$ is again in U, so the subgroup is 'closed' with respect to the operation '$*$'.

A group is called

- *finite* if it has a finite number of elements.
- *discrete* if the elements are countable.
- *continuous* if the elements can be parameterized continuously.
- a *Lie group* if it is continuous and in a still-to-be-defined sense Taylor-expandable around its neutral element.

In physics, we usually regard the elements of a group as *transformations* that can be executed sequentially, denoted by the symbol '$\circ$'. If a set of transformations leaves the considered physical system invariant, we speak of a *symmetry transformation* forming a *symmetry group*. Typical examples are translations and rotations.

> **Example:** A simple example of a *finite* group is the reflection group Z_2. It consists of only two elements $\{e, r\}$, where the neutral element e does nothing while r carries out a reflection. The reflection is an involution, i.e., it is inverse to itself: $r^2 = r \circ r = e$. Another example is the group P_n of the permutations of n objects, which is known to have $n!$ elements.
>
> A finite group is always discrete, as its elements can be enumerated, but the reverse is not

always true. For example, the addition of integers $\mathbb{Z}$ is a discrete but infinite group. Here the neutral element is 0, and the inverse element of n is $-n$.

An illustrative example of a continuous group is the group of rotations in n dimensions, known as the 'group of special orthogonal transformations in 3 dimension' and abbreviated by $SO(n)$. Here the group elements can be parametrized by continuously varying rotation angles. Clearly, a continuous group is always infinite. In addition, as we will discuss below, one can consider infinitesimal rotations. This allows us to Taylor-expand the group elements around identity, hence the $SO(n)$ is a Lie group.

1.1.2 Cosets

If you map all elements of a subgroup $U \subset G$ by means of the group operation '$*$' with another group element $a \in G$ that does *not* belong to this subgroup, you will be led out of the subgroup. The corresponding image set is called a *coset* of U with respect to a. Depending on whether a is applied to the elements of the subgroup from the left or from the right, we have to distinguish two types of cosets:

$$\begin{aligned} &\text{Left coset:} && aU = \{a * u \mid u \in U\} \\ &\text{Right coset:} && Ua = \{u * a \mid u \in U\} \end{aligned}$$

Note that different group elements $a, b \in G$ may generate the same coset $aU = bU$ or $Ua = Ub$; in this case one can easily show that $a * b^{-1} \in U$. Note that cosets (with the exception of $eU = Ue = U$) are not subgroups of G because they have no neutral element.

Example: As an example let us consider the group of translations in two dimensions, represented by a displacement vector $(\Delta x, \Delta y)$. The set of all translations in x-direction $u = (\Delta x, 0)$ forms a certain subgroup U. The group element $a = (0, 3)$, which shifts three units in y-direction, induces the coset aU of all shifts of the form $(\Delta x, 3)$ parallel to the x-axis. Since translations are commutative, right and left cosets are identical. As noted above, different group elements can produce the same coset, e.g., the group elements $a = (0, 3)$ and $b = (5, 3)$ induces the same coset $aU = bU$. In this case we can easily verify that $a * b^{-1} = a - b = (-5, 0) \in U$.

1.1.3 Normal subgroups

A subset $N \subset G$ is called a *normal subgroup* (german: Normalteiler) if its left and right cosets are identical, meaning that $gN = Ng$ holds for all $g \in G$. In this case we use the notation $N \lhd G$ or $N \trianglelefteq G$. Normal subgroups are invariant under *conjugation*, i.e., $N = gNg^{-1}$ or $[N, g] = 0$ for all $g \in G$. Note that if G is a commutative group, all subgroups are automatically normal subgroups.

Example: As an example, consider the non-commutative group of orthogonal transformation in three dimensions $O(3)$, which is known to include rotations and reflections. As one can easily see, the rotations around the z-axis form a subgroup, but this is not a normal subgroup because rotations around different axes do not commute. The Z_2-isomorphic subgroup of reflections, on the other hand, is a normal subgroup since reflections commute with rotations.

1.1.4 Quotient groups

The set of all cosets of a normal group $N \trianglelefteq G$ again forms a group equipped with the operation $aN * bN = (a * b)N$ and the neutral element $eN = N$. This group of cosets is called *quotient group* or *factor group*, denoted by a division sign G/N. For finite groups, the number of elements in a factor group is $|G/N| = |G|/|N|$.

> **Example:** In the previous example the quotient group $O(3)/Z_2 = SO(3)$ is the group of all orthogonal transformations without reflections, which are just the rotations in 2D. A further example is the permutation group P_n where the cyclic shifts form a subgroup Z_n, e.g.,
>
> $$P_3 = \{123, 132, 213, 231, 312, 321\}, \quad Z_3 = \{123, 231, 312\} \;\Rightarrow\; P_3/Z_3 = \{\langle 123\rangle, \langle 132\rangle\}.$$
>
> Since $Z_n \lhd P_n$is a normal subgroup, there exists the quotient group P_n/Z_n consisting of all rearrangements without cyclic shifts. Obviously we have $P_n/Z_n \cong P_{n-1}$.

1.1.5 Homomorphisms

A *homomorphism* (*homomorphic* = 'similarly shaped') is a map which in the broadest sense preserves the structure of the mapped object. [1] There are several special cases of homomorphisms:

- An *isomorphism* is a bijective (i.e., invertible) homomorphism.
- An *endomorphism* is a homomorphism of an object onto itself.
- An *automorphism* is an isomorphism of an object onto itself.

Iso- or automorphisms thus preserve the structure without loss of information, while homo- and endomorphism only partially preserve the structure, comparable to a projection. Depending on the type of objects to be mapped, there are various different types of homomorphisms.

For our purposes the most important type of homomorphisms is a *group homomorphism*. A group homomorphism is a map $h : G \to H$ from a group G into another group H so that for all $a, b \in G$ the following relation holds:

$$h(a *_G b) = h(a) *_H h(b). \tag{1.1}$$

Here the symbols $*_G$ and $*_H$ denote the group operations in G and H, respectively. The above definition states that h maps the group operation of G faithfully onto the group operation of H in such a way that the group structure of G is (partially) preserved. This equation also implies that a group homomorphism commutes with the operation of taking the inverse, i.e.,

$$h(a^{-1}) = h(a)^{-1}.$$

Two groups are called *isomorphic* if they are related by an isomorphism, commonly expressed by the notation

$$G \cong H.$$

[1] I should be noted that 'homeomorphism' means something different, namely, that the map preserves the *topology* of an object rather than its group structure.

In contrast to isomorphisms, which map the group structure without any loss of information, a homomorphism can partially or completely project away the group structure contained in G. An extreme case is the trivial mapping $h : G \to H : a \to e \in H$, which maps every element of G to the neutral element of H, meaning that each group is homomorphic to the trivial subgroup containing only the identity:

$$G \cong \{e\}.$$

Examples:

(a) **Modulo operation:** The group of real numbers equipped with the group operation $+$ is *homomorphic* to the group $SO(2)$ of rotations around the origin in a plane, where the homomorphism maps the real number to the rotation angle. Note that this mapping is not invertible, since rotation angles are only specified modulo 2π.

(b) **Exponential function:** The group of real numbers equipped with the group operation $+$ is *isomorphic* to the group of positive real numbers equipped with the operation $\cdot$, where the exponential function is the (invertible) isomorphism between the two groups. Here we write $(\mathbb{R}, +) \cong (\mathbb{R}^+, \cdot)$.

(c) **Permutation parity:** Let P_n be the set of permutations of n objects, where the identical permutation $e \in P_n$ is the neutral element. Furthermore let $s \in P_n$ with $s \neq e$ be an arbitrarily chosen transposition swapping two elements, i.e., $s \circ s = e$. Then the map $h : P_n \to P_n$ with

$$h(p) = \begin{cases} e & \text{if } sign(p) = 1 \\ s & \text{if } sign(p) = -1 \end{cases}$$

is an *endomorphism* of P_n onto itself. This endomorphism preserves only the sign of the permutation and maps it to a certain transposition, forming the subgroup $P_2 \cong Z_2$.

(d) **Reciprocal:** The group of real numbers except zero $\mathbb{R}\backslash\{0\}$ equipped with the operation of multiplication is mapped onto itself by taking its reciprocal. Since computing the reciprocal is invertible, this mapping is an *automorphism*.

1.1.6 Kernel, image, and the corresponding quotient groups

In order to characterize the information loss of a homomorphism $G \to H$ let us define

- the *kernel* $\ker(h) = \{a \in G \,|\, h(a) = e \in H\}$ as the subset of G which is "projected away" (mapped to zero) by the homomorphism, and
- the *image* $\mathrm{img}(h) = \{h(a) \in H \,|\, a \in G\}$ as the subset of H which is actually reached by the homomorphism.

The kernel of a homomorphism is a normal subgroup because for all $b \in G$ we have

$$\begin{aligned} b * \ker(h) &= \{b * c \,|\, c \in G \wedge h(c) = e\} \\ &= \{a \in G \,|\, h(b^{-1} * a) = e\} \\ &= \{a \in G \,|\, h(a * b^{-1}) = e\} \\ &= \{c * b \,|\, c \in G \wedge h(c) = e\} = \ker(h) * b\,, \end{aligned}$$

implying that $b * \ker(h) = \ker(h) * b$. Therefore the kernel is a normal subgroup:

$$b * \ker(h) = \ker(h) * b \quad \Rightarrow \quad \ker(h) \trianglelefteq G\,. \tag{1.2}$$

Thus every homomorphism automatically induces a corresponding quotient group $G/\ker(h)$.

Examples: In the preceding examples, the corresponding quotient groups $G/\ker(h)$ are

(a)	$(\mathbb{R},+)/(2\pi\mathbb{Z}) \cong SO(2)$	(shifts by multiples of2π)
(b)	$(\mathbb{R},+)/\{0\} = (\mathbb{R},+)$	(isomorphisms do not destroy an information)
(c)	$P_n/\{e,s\} \cong P_n/Z_2 \cong P_n^+$	(set of all positive permutations)
(d)	$(\mathbb{R}\backslash\{0\},\cdot)/\{1\} = (\mathbb{R}\backslash\{0\},\cdot)$	(automorphisms do not destroy information)

Conversely, for every given normal subgroup of a group there exists a corresponding homomorphism whose kernel is precisely this normal subgroup.

1.2 Vector spaces

1.2.1 Fields

A *field* K (german *Körper*) is a set of elements which can both be added $(+)$ and multiplied $(\cdot)$ with each other. More specifically, a field $(K,+,\cdot)$ involves two operations $+$ and $\cdot$, obeying the following axioms:

(i) $(K,+)$ is a commutative group.

(ii) $(K\backslash\{0\},\cdot)$ is a commutative group.

(iii) Addition and multiplication are compatible with one another in the sense that one can apply the well-known "point before line calculation" rule, i.e., the two operations obey the *distributive laws*

$$\lambda\cdot(\mu+\nu) = \lambda\cdot\mu + \lambda\cdot\nu, \qquad (\lambda+\mu)\cdot\nu = \lambda\cdot\nu + \mu\cdot\nu \qquad \forall_{\lambda,\mu,\nu\in K}.$$

Examples are the rational numbers $\mathbb{Q}$, the real numbers $\mathbb{R}$ and the complex numbers $\mathbb{C}$. A counterexample is the set of integers $\mathbb{Z}$, because here the multiplication by an integer has no inverse. Note that both operations have different neutral elements (addition 0, multiplication 1) and that in the second axiom the element zero has to be excluded in order to avoid division by zero, meaning that '0' has no inverse with respect to multiplication.

Remark: Numerical non-vectorial values $\lambda \in K$ are often referred to as *scalars* with K being the corresponding *scalar field*. Scalars are quantities that are independent of the selected coordinate system, that is, they are invariant under coordinate transformations. For example, the temperature of a solid body is a scalar while its velocity is not.

1.2.2 Vector space axioms

A linear *vector space* V based on a scalar field K is a set of elements called *vectors* which can be scaled (varied in their length) and combined linearly (added). The axioms of a vector space read

(i) Vectors form a commutative group with respect to addition $(V,+)$.

(ii) Vectors $\mathbf{u}, \mathbf{v} \in V$can be multiplied by scalars $\lambda, \mu \in K$ from the left. This type of scalar multiplication has the following properties:

- Homogeneity: $\lambda(\mu\mathbf{v}) = (\lambda\mu)\mathbf{v}$
- Linearity in V: $\lambda(\mathbf{u} + \mathbf{v}) = \lambda\mathbf{u} + \lambda\mathbf{v}$
- Linearity in K: $(\lambda + \mu)\mathbf{v} = \lambda\mathbf{v} + \mu\mathbf{v}$
- Neutral element $1 \in K$: $1\mathbf{v} = \mathbf{v}$.

The following should be noted:

a) Since $(V, +)$ is a commutative group, a vector space singles out a special neutral element, namely, the zero vector $\vec{0}$. However, it is clear that the physical position space does not single out a particular position and therefore, strictly speaking, it is not a vector space in the mathematical sense (cf. Sect. 1.2.3 on page 6).

b) The vector space axioms tell us nothing about the length of a vector. This requires a norm or a metric, so to say as 'addons'.

c) Likewise, the vector space axioms tell us nothing about the angle between two vectors. To specify angles one needs to define a scalar product which is an additional concept.

> **Example:** The complex 3×3-matrices can be added and scaled, hence they form a vector space over $\mathbb{C}$. However, there is no natural notion of 'distance' or 'angle' between two matrices motivated by everyday experience. But as we shall see further below, it is indeed possible to define these notions in a meaningful manner.

1.2.3 Affine spaces

An affine space is a set of points connected by vectors.

> **More precisely:** An affine space A based on a vector space V is a set of points $p, q \in \mathbf{A}$ equipped with a map $A \times A \to V : p, q \to \overrightarrow{pq}$ such that:
>
> - The distance vectors are additive, i.e., for all $p, q, r \in A$ one has: $\overrightarrow{pq} + \overrightarrow{qr} = \overrightarrow{pr}$.
> - An affine space has no boundaries, hence for all $p \in A$ and $\mathbf{v} \in V$ there exists a $q \in A$ such that $\mathbf{v} = \overrightarrow{pq}$.

An affine space thus has essentially the same structure as the underlying vector space, but in contrast it does not single out a particular neutral element, i.e., it does not have an origin (see Fig. 1.1 c). This is achieved by using the vectors to no longer describe absolute positions but only differences between positions. The transition from a vector space to the corresponding affine space is often described as "forgetting the origin." Physical spaces such as the position space in Newtonian mechanics are affine because they do not single out a particular origin.

1.2.4 Representation of vectors

The scalar multiplication allows one to form linear combinations $\mathbf{v} = \sum_i \lambda^i \mathbf{v}_i$ of vectors $\mathbf{v}_i \in V$ with linear factors $\lambda^i \in K$. It is customary to use lower indices for the vectors

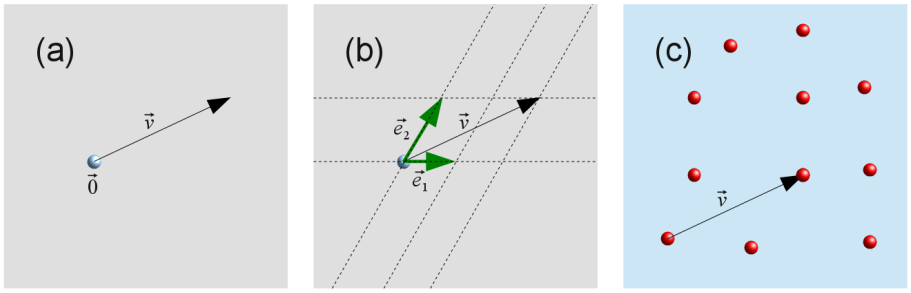


Figure 1.1: Vector space $\mathbb{R}^2$: (a) In a **vector space**, the zero vector (blue dot) is singled out since it plays the role of the neutral element. (b) Using a **basis** $\{\mathbf{e}_1, \mathbf{e}_2\}$ the vector $\mathbf{v}$ can be represented in the present example by $\mathbf{v} = 2\mathbf{e}_1 + \mathbf{e}_2$. (c) An **affine space** consists of points (some of which are shown in red here) connected by vectors of a vector space. The affine space has essentially the same structure as the vector space, but it does not single out a particular neutral element - all points are on equal footing, just as in nature.

and upper indices for the linear factors for reasons which will be discussed in more detail below.

A set $\{\mathbf{v}_i\}$ of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots$ is called *linearly dependent* if there is a non-trivial linear combination that yields the null vector, otherwise they are said to be *linearly independent*. A linearly independent set $\{\mathbf{e}_i\}$ of vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots$ is called a *basis* of V if all vectors $v \in V$ can be expressed as a linear combination

$$\mathbf{v} = \sum_i v^i \mathbf{e}_i \,. \tag{1.3}$$

This means that the entire vector space is spanned by linear combinations of the basis vectors. The number of basis vectors is the dimension $d = dim(V)$ of the vector space. The linear coefficients $v^i \in K$, which usually carry an upper index, are called *coordinates* or *components* of the vector. Usually they are written in the form of a *column vector*, e.g.

$$\mathbf{v} = \begin{pmatrix} v^1 \\ v^2 \\ v^3 \end{pmatrix} \,. \tag{1.4}$$

Note that the components refer to the arbitrarily chosen basis and allow the vector to be represented as a set of numbers $\{v^i\}$ (see Fig. 1.1). Since the choice of the basis is not unique, there are generally infinitely many possible representations for a given vector, each of them having numerically different vector components.

Remark: It is important to understand the difference between (a) the physical reality, (b) the abstract mathematical object used for modeling, and (c) its mathematical *representation*. The physical position space, for example, is modeled in the context of Newtonian mechanics by the vector space $\mathbb{R}^3$. This vector space and the vectors contained therein are abstract mathematical objects. To deal with them, one needs to represent them by coordinates, e.g., in terms of Cartesian coordinates or polar coordinates. For a given abstract mathematical object, there is usually a large variety of possible representations. Their classification is the subject of an independent branch of mathematics, called *representation theory*.

The mathematical structure of a problem becomes particularly transparent when formulated without referring to a particular representation. Especially mathematical proofs are particularly elegant if they do not rely on a specific representation. However, to work out something concretely in terms of actual numbers, it is usually inevitable to invoke a repre-

sentation at some point.

1.3 Linear maps

In Sect. 1.1.5 on page 3 we discussed the concept of homomorphisms as structurally faithful maps and discussed the example of group homomorphisms. Analogously, *vector space homomorphisms* are mappings that preserve the structure of a vector space, i.e., in addition, multiplication, and the distributive laws. As will be shown in the following, vector space homomorphisms are *linear* maps.

1.3.1 Definition and properties

Let V, W be two vector spaces. A map $\mathbf{A} : V \to W$ is called *linear* if for all $\mathbf{u}, \mathbf{v} \in V$ and for all $\lambda, \mu \in K$ the map obeys the linearity property

$$\mathbf{A}(\lambda\mathbf{u} + \mu\mathbf{v}) = \lambda\mathbf{A}(\mathbf{u}) + \mu\mathbf{A}(\mathbf{v}). \tag{1.5}$$

Remark: In complex vector spaces, there exists also another class of vector space homomorphisms, namely,m so-called *antilinear* maps, also known as *conjugate-linear* or *semilinear* maps. Antilinear maps have the property that scalars, when pulled out in front of the function by means of the laws of linearity, become complex conjugated, i.e.,

$$\mathbf{A}(\lambda\mathbf{u} + \mu\mathbf{v}) = \lambda^*\mathbf{A}(\mathbf{u}) + \mu^*\mathbf{A}(\mathbf{v})\,, \tag{1.6}$$

where the star stands for complex conjugation. While antilinear mappings are frequently encountered in quantum theory, they play a rather subordinate role in the theory of relativity, because here we primarily deal with real vector spaces.

Similar to group homomorphisms, linear maps in their capacity as vector space homomorphisms have a certain *kernel* (set of input vectors that are mapped to zero) and a certain *image* (set of output vectors that are reached by the map):

$$\ker(\mathbf{A}) \;=\; \{\mathbf{v} \in V \,|\, \mathbf{A}(\mathbf{v}) = 0\} \;\subseteq\; V \tag{1.7}$$

$$\mathrm{img}(\mathbf{A}) \;=\; \{\mathbf{A}(\mathbf{v}) \,|\, \mathbf{v} \in V\} \;\subseteq\; W\,. \tag{1.8}$$

The kernel and the image are linear subspaces of V and respectively W which obey the so-called *rank–nullity theorem* (german: Dimensionssatz)

$$\dim\big(\ker(\mathbf{A})\big) + \dim\big(\mathrm{img}(\mathbf{A})\big) \;=\; \dim(V)\,. \tag{1.9}$$

The dimension of the image is called the *rank* of the linear map:

$$\mathrm{rank}(\mathbf{A}) \;:=\; \dim\big(\mathrm{img}(\mathbf{A})\big) \tag{1.10}$$

For the rank one can show that the following inequality holds:

$$\mathrm{rank}(\mathbf{A}) \;\leq\; \min(\dim V, \dim W)\,. \tag{1.11}$$

A linear map is said to have *full rank* if $\mathrm{rank}(\mathbf{A}) = \dim V$. Obviously, this requires that $\dim W \geq \dim V$.

1.3.2 Representation of linear maps

Let us now consider the *representation* of a linear map which, as pointed out before, must not be confused with the abstract map itself. Let $\mathbf{A} : V \to W$ be a linear map and let $\{\mathbf{e}_i\}$ be a basis of V and let $\{\mathbf{f}_j\}$ a basis of W. Using these basis systems we can represent any a vector $\mathbf{v} \in V$ and its image $\mathbf{w} = \mathbf{A}(\mathbf{v}) \in W$ by

$$\mathbf{v} = v^i \mathbf{e}_i \qquad \text{and} \qquad \mathbf{w} = w^j \mathbf{f}_j \,, \tag{1.12}$$

where the components v^i and w^j are given by $v^i = \mathbf{e}^i(\mathbf{v})$ and $w^j = \mathbf{f}^j(\mathbf{w})$. Because of the linearity of $\mathbf{A}$ we have

$$\mathbf{A}(\mathbf{v}) \;=\; \mathbf{A}(v^i \mathbf{e}_i) \;=\; \mathbf{A}(\mathbf{e}_i) v^i \,, \tag{1.13}$$

where $\mathbf{A}(\mathbf{e}_i) \in W$. Being an element of W, the vectors $\mathbf{A}(\mathbf{e}_i)$ can also be represented in the basis $\{\mathbf{f}_j\}$ by certain components A^j which will of course depend on $\mathbf{e}_i$, i.e.,

$$\mathbf{A}(\mathbf{e}_i) \;=\; \underbrace{A^j(\mathbf{e}_i)}_{=:\, A^j{}_i} \mathbf{f}_j \,, \tag{1.14}$$

or in short: $\mathbf{A}(\mathbf{v}) = A^j{}_i v^i \mathbf{f}_j$. Comparing this expression with $\mathbf{A}(\mathbf{v}) = \mathbf{w} = w^j \mathbf{f}_j$ one can easily see that the components of the vectors are related by

$$\boxed{w^j \;=\; A^j{}_i v^i} \,. \tag{1.15}$$

This relation can be recast in form of a matrix multiplication, e.g. a linear map $\mathbb{R}^3 \to \mathbb{R}^2$ would be represented by a 2×3-matrix

$$\begin{pmatrix} w^1 \\ w^2 \end{pmatrix} \;=\; \begin{pmatrix} A^1{}_1 & A^1{}_2 & A^1{}_3 \\ A^2{}_1 & A^2{}_2 & A^2{}_3 \end{pmatrix} \begin{pmatrix} v^1 \\ v^2 \\ v^3 \end{pmatrix} \tag{1.16}$$

using the well-known calculation rule of multiplying "line by column". Therefore, in a given basis a linear map can always be represented as a matrix whose actual entries depend on the choice of the basis. This illustrates that any representation carries a man-made redundancy due to the choice of the basis.

1.3.3 Basis transformations

In the broadest sense a *transformation* is an isomorphism that transforms the object to be described in a reversible way. There are basically two types of transformations:

- *Active transformations* which manipulate the object physically, for example, by rotating or moving it.
- *Passive transformations*, which mediate between different representations, while the object itself remains unaffected.

In linear algebra, we are particularly interested in *linear transformations*. Vector space automorphisms, that is, linear maps described by square matrices, are *active* transformations because they map given vectors onto different vectors and thus they directly manipulate the object under consideration. This is to be distinguished from *passive* linear transformations, which mediate between representations of the same object in different coordinate systems. Since a representation is determined by the choice of the basis, such transformations are also called *basis transformations*.

In the following we consider a passive linear transformation which maps a given basis $\{\mathbf{e}_i\}$ to another 'ticked' basis $\{\mathbf{e}'_i\}$. Clearly, it is possible to represent the new basis vectors in terms of linear combinations of the old basis vectors, i.e. we can write

$$\boxed{\mathbf{e}_i \;\rightarrow\; \mathbf{e}'_i = \mathbf{e}_k \tilde{M}^k{}_i\,,} \tag{1.17}$$

where $\tilde{M}$ is the corresponding transformation matrix whose determinant must be non-zero in order to ensure that the transformation is reversible. While a vector $\mathbf{v} \in V$ in itself remains invariant under the action of such a (passive) basis transformation, this does of course not apply to its components. Because of

$$\mathbf{v} = v^k \mathbf{e}_k = v'^i \mathbf{e}'_i = v'^i \mathbf{e}_k \tilde{M}^k{}_i$$

a comparison of coefficients yields $v^k = v'^i \tilde{M}^k{}_i$. Multiplying with the inverse matrix from the left, we can now deduce the transformation law for the components

$$\boxed{v^i \;\rightarrow\; v'^i \;=\; M^i{}_j v^j\,,} \tag{1.18}$$

where $M = \tilde{M}^{-1}$.

> **Remember:** If the vector components v^i transform with the matrix M, the basis vectors transform with the matrix M^{-1} and vice versa.

Behavior of matrices under basis transformations:
Likewise the representation of the linear map $\mathbf{A} : V \rightarrow V$ in terms of a matrix $A^j{}_i$ changes under basis transformations. Starting with the equation $\mathbf{w} = \mathbf{A}\mathbf{v}$, the representation (1.15) written in the original and the ticked basis reads

$$w^j = A^j{}_i v^i\,, \qquad\qquad w'^j = A'^j{}_i v'^i\,, \tag{1.19}$$

where $A^j{}_i$ and $A'^j{}_i$ are square matrices. Because of $v'^i = M^i{}_j v^j$ and $w'^i = M^i{}_j w^j$ one obtains the transformation law

$$\boxed{A'^j{}_i \;=\; M^j{}_k A^k{}_\ell \tilde{M}^\ell{}_i} \tag{1.20}$$

Again it should be noted that the basis transformation does not change the linear map as such, but only its representation in terms of a matrix.

1.4 Composite vector spaces

Two vector spaces U and V over the same scalar field K can be combined to form a new vector space in various different ways. In this context it is particularly important to understand the difference between a *direct sum* and a *tensor product*. In addition, we will see that there are several variants of tensor products which differ in their symmetry properties.

1.4.1 Direct sum $\oplus$

The *direct sum* $U \oplus V$ (also called *exterior direct sum*) is defined as the set of all *ordered* pairs $(\mathbf{u}, \mathbf{v}) =: \mathbf{u} \oplus \mathbf{v}$ of vectors $\mathbf{u} \in U$ and $\mathbf{v} \in V$ equipped with the operation of addition[2]

$$(\mathbf{u}_1 + \mathbf{u}_2) \oplus (\mathbf{v}_1 + \mathbf{v}_2) \;=\; (\mathbf{u}_1 \oplus \mathbf{v}_1) + (\mathbf{u}_2 \oplus \mathbf{v}_2) \tag{1.21}$$

and scalar multiplication

$$(\lambda \mathbf{u}) \oplus (\lambda \mathbf{v}) \;=\; \lambda(\mathbf{u} \oplus \mathbf{v})\,. \tag{1.22}$$

It is easy to verify that the direct sum defined above is also a vector space, which is usually denoted as $U \oplus V$. As can be seen, the dimension of this vector space is just the sum of the individual dimensions:

$$\dim(U \oplus V) = \dim(U) + \dim(V). \tag{1.23}$$

The concept of the direct sum can easily be extended to linear maps. Suppose that $\mathbf{A} : U \to U$ and $\mathbf{B} : V \to V$ are linear maps. Then we can define the linear map $\mathbf{A} \oplus \mathbf{B}$ on $U \oplus V$ by means of

$$(\mathbf{A} \oplus \mathbf{B})(\mathbf{u} \oplus \mathbf{v}) = (\mathbf{A}\mathbf{u}) \oplus (\mathbf{B}\mathbf{v}). \tag{1.24}$$

In low dimensions it is very easy and intuitive to visualize a direct sum. For example, the vector space $\mathbb{R}^3$ can be thought of as the direct sum of $\mathbb{R}^2$ (z.B. xy-plane) and $\mathbb{R}$ (z-axis).

1.4.2 Representation of direct sums

If $\{\mathbf{e}_i\}$ is a basis of U and likewise $\{\mathbf{f}_j\}$ a basis of V, then the canonical basis vectors of the direct sum $U \oplus V$ are given by combining the basis vectors of one constituting vector space with the neutral element of the other, i.e.,

$$\{(\mathbf{e}_1, 0),\, (\mathbf{e}_2, 0), \ldots,\, (0, \mathbf{f}_1),\, (0, \mathbf{f}_2),\, , \ldots\} \tag{1.25}$$

where '0' denotes the respective zero vectors. This construction confirms that the dimensions of the individual spaces simply add up when forming their direct sum.

Now suppose that $\mathbf{u} \in U$ and $\mathbf{v} \in V$ are represented in the respective basis systems, i.e., $\mathbf{u} = u^i \mathbf{e}_i$ and $\mathbf{v} = v^i \mathbf{f}_i$. Usually, the components of vectors are arranged in the form

[2] Formally, the direct sum $U \oplus V$ is similar to a Cartesian product $U \times V$, but goes beyond in so far as it gives the resulting space the structure of a vector space.

of column vectors, e.g.

$$\mathbf{u} = \begin{pmatrix} u^1 \\ u^2 \\ u^3 \end{pmatrix}, \qquad \mathbf{v} = \begin{pmatrix} v^1 \\ v^2 \end{pmatrix}. \tag{1.26}$$

When forming the direct sum, the components of the individual vectors are simply concatenated:

$$\mathbf{u} \oplus \mathbf{v} = \begin{pmatrix} u^1 \\ u^2 \\ u^3 \end{pmatrix} \oplus \begin{pmatrix} v^1 \\ v^2 \end{pmatrix} = \begin{pmatrix} u^1 \\ u^2 \\ u^3 \\ v^1 \\ v^2 \end{pmatrix}. \tag{1.27}$$

A similar rule can be formulated for matrices: If $\mathbf{A} : U \to U$ and $\mathbf{B} : V \to V$ are two linear maps represented by the square matrices $A^j{}_i$ and $B^j{}_i$, then the direct sum $\mathbf{A} \oplus \mathbf{B}$ represented in the canonical basis has a block-diagonal structure:

$$\mathbf{A} \oplus \mathbf{B} = \begin{pmatrix} A^1{}_1 & A^1{}_2 & A^1{}_3 \\ A^2{}_1 & A^2{}_2 & A^2{}_3 \\ A^3{}_1 & A^3{}_2 & A^3{}_3 \end{pmatrix} \oplus \begin{pmatrix} B^1{}_1 & B^1{}_2 \\ B^2{}_1 & B^2{}_2 \end{pmatrix} = \begin{pmatrix} A^1{}_1 & A^1{}_2 & A^1{}_3 & 0 & 0 \\ A^2{}_1 & A^2{}_2 & A^2{}_3 & 0 & 0 \\ A^3{}_1 & A^3{}_2 & A^3{}_3 & 0 & 0 \\ 0 & 0 & 0 & B^1{}_1 & B^1{}_2 \\ 0 & 0 & 0 & B^2{}_1 & B^2{}_2 \end{pmatrix}. \tag{1.28}$$

Here one can see immediately that it is not always possible to write any linear map $\mathbf{C} : U \oplus V \to U \oplus V$ in the form $\mathbf{A} \oplus \mathbf{B}$, but only such maps which possess a block-diagonal structure.

1.4.3 Tensor product $\otimes$

The *tensor product* $U \otimes V$ (also known as the *exterior product*) is also defined in terms of the *ordered* pairs $(\mathbf{u}, \mathbf{v})$ of vectors $\mathbf{u} \in U$ and $\mathbf{v} \in V$, but in contrast to the direct sum the vector space structure is implemented here in a completely different way in order to equip it with a multiplicative structure. In particular, we can deal with a tensor product in the same way as with a regular product by 'multiplying out' as follows:

$$(\mathbf{u}_1 + \mathbf{u}_2) \otimes (\mathbf{v}_1 + \mathbf{v}_2) \;=\; (\mathbf{u}_1 \otimes \mathbf{v}_1) + (\mathbf{u}_1 \otimes \mathbf{v}_2) + (\mathbf{u}_2 \otimes \mathbf{v}_1) + (\mathbf{u}_2 \otimes \mathbf{v}_2)\,. \tag{1.29}$$

Moreover, the tensor product is bilinear under scalar multiplication in both arguments:

$$(\lambda \mathbf{u}) \otimes (\mu \mathbf{v}) \;=\; \lambda\mu(\mathbf{u} \otimes \mathbf{v})\,. \tag{1.30}$$

> **Remark:** These definition properties are significantly different from those for the direct sum in Eqs. (1.21)-(1.22). In the case of the direct sum, the components are simply concatenated, while in the case of the tensor product, 'everyone is multiplied with everyone' in all possible combinations, giving additional in mixed terms.

With the tensor product defined in this way, let us consider the set of so-called *product vectors*

$$P_{UV} \;:=\; \{\mathbf{u} \otimes \mathbf{v} \,|\, \mathbf{u} \in U, \mathbf{v} \in V\}\,. \tag{1.31}$$

At this point it is very important to note that this set does not yet form a vector space because not every linear combination of two product vectors $\mathbf{u}_1 \otimes \mathbf{v}_1$ and $\mathbf{u}_2 \otimes \mathbf{v}_2$ can again be written as product vector $\mathbf{u}_3 \otimes \mathbf{v}_3$. The *tensor product space* $U \otimes V$ is therefore defined as the *span* of P, that is, the set of all of product vectors *plus* all their possible linear combinations:[3]

$$U \otimes V \;:=\; \langle P_{UV} \rangle \,. \tag{1.32}$$

In the spanned space $U \otimes V$ the vector space axioms are fulfilled and it is easy to see that the dimension of this space is equal to the *product* of the individual dimensions:

$$\dim(U \otimes V) = \dim(U)\dim(V). \tag{1.33}$$

The constituting vector spaces U and V are called the *tensor components* of the tensor product $U \otimes V$. An illustrative interpretation of the tensor product is difficult since the first nontrivial case $\mathbb{R}^2 \otimes \mathbb{R}^2$ already challenges us with imaging a four-dimensional space.

Since a tensor product is based on ordered pairs, we have in general $U \otimes V \neq V \otimes U$, i.e., the tensor product is not commutative and the tensor factors (the so-called *sots* of the tensor) thus have an individual distinct identity.

> **Remark:** Probably you have already encountered tensor products in quantum theory. For central potentials, for example, the eigenfunctions $\psi_n(\mathbf{r})$ factorize into a radial component $u(r)$ and a spherical surface function $Y_{lm}(\theta, \phi)$. However, instead of correctly writing $|\psi_n\rangle = |u\rangle \otimes |Y_{lm}\rangle$ most textbooks use the shortcut notation $|\psi_n\rangle = |u\rangle|Y_{lm}\rangle$. Similarly, for systems with several spins one uses the shortcut notation $|\uparrow\uparrow\rangle = |\uparrow\rangle|\uparrow\rangle$ instead of $|\uparrow\rangle \otimes |\uparrow\rangle$.

1.4.4 Calculation rules for tensor products

The tensor product of two linear maps is defined in such a way that the two maps act separately on each tensor component, i.e.,

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{u} \otimes \mathbf{v}) \;=\; (\mathbf{A}\mathbf{u}) \otimes (\mathbf{B}\mathbf{v}) \,. \tag{1.34}$$

A linear map $\mathbf{C} : U \otimes V \to U \otimes V$, which can be written in the form $\mathbf{C} = \mathbf{A} \otimes \mathbf{B}$, is called *factorizable*. Such maps act on the two tensor factors separately without generating any correlation between them. The transpose of a factorizable linear map can be computed by individually transposing all tensor components, i.e.,

$$(\mathbf{A} \otimes \mathbf{B})^T = \mathbf{A}^T \otimes \mathbf{B}^T \tag{1.35}$$

and the same holds for the adjoint of $\mathbf{A} \otimes \mathbf{B}$ in a complex vectors space:

$$(\mathbf{A} \otimes \mathbf{B})^\dagger = \mathbf{A}^\dagger \otimes \mathbf{B}^\dagger \,. \tag{1.36}$$

[3] Both the direct sum $U \oplus V$ and the tensor product $U \otimes V$ are based on ordered pairs of vectors taken from U and V. Apart from the different addition rule, one significant difference is that in the case of the direct sum the ordered pairs already encompass the entire sum space, whereas in the tensor product space they represent only a specific subset (namely, that of the product vectors) and that the entire vector space can only be defined consistently by including all their linear combinations.

> **Remark:** In contrast to concatenated maps (represented by matrix products), where the order of the factors is reversed under transposition by $(\mathbf{A}_1\mathbf{A}_2)^T = \mathbf{A}_2^T\mathbf{A}_1^T$, the order of the tensor factors in a tensor product does not change under transposition. Be careful in situations where both types of products are mixed, we have for example $[(\mathbf{A}_1\mathbf{A}_2)\otimes(\mathbf{B}_1\mathbf{B}_2\mathbf{B}_3)]^T = (\mathbf{A}_2^T\mathbf{A}_1^T)\otimes(\mathbf{B}_3^T\mathbf{B}_2^T\mathbf{B}_1^T)$.

The tensor product of two scalars $\lambda, \mu \in \mathbb{C}$ is formally interpreted as multiplication in $\mathbb{C}$:

$$\lambda \otimes \mu \equiv \lambda\mu\,. \tag{1.37}$$

The trace of a tensor product is defined as the product of the traces, that is

$$\mathrm{Tr}(\mathbf{A}\otimes\mathbf{B}) \;=\; \mathrm{Tr}\,\mathbf{A}\;\mathrm{Tr}\,\mathbf{B}\,. \tag{1.38}$$

Contrarily, the determinant of a tensor product is the product of the determinant of each factors raised to the respective power of the dimension of the other factors:

$$\det(\mathbf{A}\otimes\mathbf{B}) \;=\; (\det\mathbf{A})^{\dim(V)}\,(\det\mathbf{B})^{\dim(U)}\,. \tag{1.39}$$

> **Remember:** Pitfall: The determinant of the tensor product is *not* the product of the determinants of its factors, instead one has to raise each factor to the power of the dimension of the other tensor slots, respectively.

Tensor products can easily be performed multiple times. For example, a vector space V can be defined as a triple tensor product $V = V_1 \otimes V_2 \otimes V_3$ of three individual vector spaces V_1, V_2, V_3. Note that the tensor product is associative, i.e., there is no necessity to use parenthesis in expressions with more than two tensor factors.

1.4.5 Representation of the tensor product

If $\{\mathbf{e}_i\}$ is a basis of U and $\{\mathbf{f}_j\}$ is a basis of V, we obtain a natural basis of the tensor product space by creating all possible tensor products of the basis vectors:

$$\begin{aligned}\{&\mathbf{e}_1\otimes\mathbf{f}_1,\ \mathbf{e}_1\otimes\mathbf{f}_2,\ \mathbf{e}_1\otimes\mathbf{f}_3,\ldots\\ &\mathbf{e}_2\otimes\mathbf{f}_1,\ \mathbf{e}_2\otimes\mathbf{f}_2,\ \mathbf{e}_2\otimes\mathbf{f}_3,\ldots\\ &\mathbf{e}_3\otimes\mathbf{f}_1,\ \mathbf{e}_3\otimes\mathbf{f}_2,\ \mathbf{e}_3\otimes\mathbf{f}_3,\ldots\\ &\ldots,\}\,.\end{aligned} \tag{1.40}$$

This is the so-called *canonical product basis*, where one uses the convention to order the factors lexicographically. As one can easily verify, the dimension of the product space is given by $\dim(U\otimes V) = \dim(U)\dim(V)$.

If a product vector is represented in the canonical product basis, its components can be computed by multiplying the components of the constituting vectors in all lexicographically ordered combinations. Let us, for example, consider the vectors $\mathbf{u} = u^i\mathbf{e}_i$ and $\mathbf{v} = v^j\mathbf{f}_j$. Then we have $\mathbf{u}\otimes\mathbf{v} = \sum_{i,j} u^iv^j\,(\mathbf{e}_i\otimes\mathbf{f}_j)$, which in the notation of column

vectors can be written as

$$\mathbf{u} \otimes \mathbf{v} = \begin{pmatrix} u^1 \\ u^2 \\ u^3 \end{pmatrix} \otimes \begin{pmatrix} v^1 \\ v^2 \end{pmatrix} = \begin{pmatrix} u^1 v^1 \\ u^1 v^2 \\ u^2 v^1 \\ u^2 v^2 \\ u^3 v^1 \\ u^3 v^2 \end{pmatrix}. \tag{1.41}$$

Thus, in contrast to the direct sum, where the two column vectors are simply concatenated, tensor products of vectors are formed by computing the products of all possible combinations of the tensor components in lexicographical order.

It is important to note that in the above example, a six-component vector is obtained on the right-hand side, although it is formed out of only five independent components on the left hand side. This reflects the circumstance mentioned before that the product vectors cover actually only a subset of the entire six-dimensional space. Only when we add all their linear combinations, we obtain the full vector space.

Remark: Note that when forming a tensor product, the corresponding physical units are also multiplied. For example, if $\mathbf{u}$ and $\mathbf{v}$ represent distances with the unit of a length, the tensor product has the dimension of an area. In contrast to the direct sum, the tensor product thus allows us to multiply vectors with different physical units and different dimensions.

The tensor product of two linear maps represented by matrices is constructed as shown in the following example::

$$\begin{aligned} \mathbf{A} \otimes \mathbf{B} &= \begin{pmatrix} A^1{}_1 & A^1{}_2 & A^1{}_3 \\ A^2{}_1 & A^2{}_2 & A^2{}_3 \\ A^3{}_1 & A^3{}_2 & A^3{}_3 \end{pmatrix} \otimes \begin{pmatrix} B^1{}_1 & B^1{}_2 \\ B^2{}_1 & B^2{}_2 \end{pmatrix} \\ &= \begin{pmatrix} A^1{}_1 B^1{}_1 & A^1{}_1 B^1{}_2 & A^1{}_2 B^1{}_1 & A^1{}_2 B^1{}_2 & A^1{}_3 B^1{}_1 & A^1{}_3 B^1{}_2 \\ A^1{}_1 B^2{}_1 & A^1{}_1 B^2{}_2 & A^1{}_2 B^2{}_1 & A^1{}_2 B^2{}_2 & A^1{}_3 B^2{}_1 & A^1{}_3 B^2{}_2 \\ A^2{}_1 B^1{}_1 & A^2{}_1 B^1{}_2 & A^2{}_2 B^1{}_1 & A^2{}_2 B^1{}_2 & A^2{}_3 B^1{}_1 & A^2{}_3 B^1{}_2 \\ A^2{}_1 B^2{}_1 & A^2{}_1 B^2{}_2 & A^2{}_2 B^2{}_1 & A^2{}_2 B^2{}_2 & A^2{}_3 B^2{}_1 & A^2{}_3 B^2{}_2 \\ A^3{}_1 B^1{}_1 & A^3{}_1 B^1{}_2 & A^3{}_2 B^1{}_1 & A^3{}_2 B^1{}_2 & A^3{}_3 B^1{}_1 & A^3{}_3 B^1{}_2 \\ A^3{}_1 B^2{}_1 & A^3{}_1 B^2{}_2 & A^3{}_2 B^2{}_1 & A^3{}_2 B^2{}_2 & A^3{}_3 B^2{}_1 & A^3{}_3 B^2{}_2 \end{pmatrix} \end{aligned} \tag{1.42}$$

Remark: In practice it is extremely useful to implement the tensor product $\otimes$ on algebra computer system. For example, the simple *Mathematica*® function, which can handle both vectors and matrices of arbitrary dimensionality according to the rules outlined above, takes only two lines of code:

```
Attributes[CircleTimes] = {Flat, OneIdentity};
CircleTimes[a_List, b_List] := KroneckerProduct[a, b];
```

Using this code snippet the tensor product $|c\rangle = |a\rangle \otimes |b\rangle$ can be written as

```
c = {a1,a2,a3} ⊗ {b1,b2}
```

where the symbol $\otimes$is obtained by typing the sequence ESC c * ESC. Note that the tensor product defined above works not only for vectors but also for matrices.

1.4.6 Multiple tensor products

Since $U \otimes V$ is again a vector space, it is of course possible to tensorize it with another vector space W, giving the vector space $(U \otimes V) \otimes W$. It can be shown that the tensor product, although being non-commutative, is an associative product, i.e.,

$$(U \otimes V) \otimes W = U \otimes (V \otimes W) \tag{1.43}$$

so that it makes sense to write a multiple tensor product without brackets as $U \otimes V \otimes W$. The associativity can be confirmed by inspection of the components in a given basis. For example, if we compute the tensor product of three vectors from a two-dimensional vector space, we obtain an 8-dimensional vector, namely,

$$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \otimes \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \otimes \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} a_1 b_1 c_1 \\ a_1 b_1 c_2 \\ a_1 b_2 c_1 \\ a_1 b_2 c_2 \\ a_2 b_1 c_1 \\ a_2 b_1 c_2 \\ a_2 b_2 c_1 \\ a_2 b_2 c_2 \end{pmatrix}, \tag{1.44}$$

where it obviously does not matter whether the left or the right tensor product is evaluated first.

Multiple tensor products play an important role in quantum mechanics. For example, if we investigate a system with N fermionic spins are described by a Hilbert space of N tensor factors $\mathcal{H} = \mathbb{C}^2 \otimes \mathbb{C}^2 \otimes \ldots \otimes \mathbf{C}^2$. In such a case, where all tensor factors have the same structure (here $\mathbb{C}^2$), it is common to write the tensor product as a *tensor power*

$$V^{\otimes N} := V \otimes V \otimes \ldots \otimes V \qquad (N \text{ times}) \tag{1.45}$$

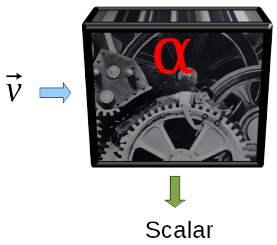
1.5 Multilinear forms

1.5.1 1-forms

Let V be a vector space based on the field K. A *linear form* or, using the parlance of differential forms, a so-called *1-form* is a map $\alpha : V \to K$ with the following properties:

- Additivity: $\alpha(\mathbf{u} + \mathbf{v}) = \alpha(\mathbf{u}) + \alpha(\mathbf{v})$ for all $\mathbf{u}, \mathbf{v} \in V$.
- Homogeneity: $\alpha(\lambda \mathbf{v}) = \lambda \alpha(\mathbf{v})$ for all $\mathbf{v} \in V$ and $\lambda \in K$.

Simply put, a 1-form is a linear *black box* that maps a vector to a scalar (a real number).



You can add two 1-forms α, τ by simply adding their results. Similarly, one can multiply a 1-form with a scalar $\lambda \in K$ by multiplying its result by the scalar. In formulas this can be

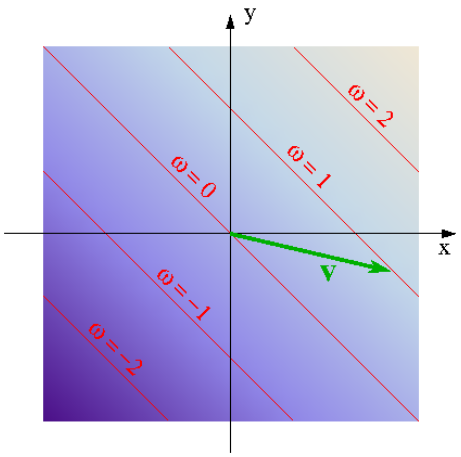


Figure 1.2: Visualization of a 1-form in $\mathbb{R}^2$. A 1-form α can be interpreted as a kind of scalar field living on the vector space, which is visualized in the figure by the intensity of the blue coloration. Since the 1-form is linear, the *equipotential lines* of this field must be parallel straight lines. If you apply a 1-form to a vector $\mathbf{v}$, then the result $\alpha(\mathbf{v} =$ is just the value of the field at this point targeted by $\mathbf{v}$, so in the present example we have $\alpha(\mathbf{v}) = 1$.

expressed as follows:

$$(\alpha + \beta)(\mathbf{v}) := \alpha(\mathbf{v}) + \beta(\mathbf{v}), \tag{1.46}$$

$$(\lambda\alpha)(\mathbf{v}) := \lambda\alpha(\mathbf{v}). \tag{1.47}$$

It is easy to show that $\alpha + \beta$ and $\lambda\alpha$ are again additive homogeneous maps, thus they are again 1-forms. Therefore the set of all 1-forms in itself also constitutes a vector space. This vector space of 1-forms is denoted by the symbol V^* and called the *co-vectorspace* or *dual vector space* or simply the *dual space* of V.[4]

Remember: A 1-form is a linear machine that maps vectors of a vector space V to numbers. The set of all 1-forms constitutes a new vector space which is referred to as the dual space V^*.

Note that 1-forms are not suitable for defining the length of a vector. For example, the length of a vector should always be positive, while a 1-form can also give negative results because of $\alpha(-\mathbf{v}) = -\alpha(\mathbf{v})$.

1.5.2 Representation of 1-forms

Let $\alpha \in V^*$ be a 1-form and let $\{\mathbf{e}_i\}$ be a basis of V. Since every vector $\mathbf{v} \in V$can be written as a linear combination $\mathbf{v} = \sum_i v^i \mathbf{e}_i$ and since α is linear, the 1-form can be represented as

$$\alpha(\mathbf{v}) = \alpha\left(\sum_i v^i \mathbf{e}_i\right) = \sum_i v^i \alpha(\mathbf{e}_i)\,. \tag{1.48}$$

Hence, in order to define a 1-form, it suffices to specify how it acts on the basis vectors, meaning that a 1-form is completely and uniquely characterized by the numbers $\alpha(\mathbf{e}_i) \in K$.

[4] More rigorously, the concept of the dual space requires continuity as an additional property, which can be relevant in infinite-dimensional spaces.

This observation allows us to construct a special set $\{\mathbf{e}^j\}$ of 1-forms with the property

$$\boxed{\mathbf{e}^j(\mathbf{e}_i) \;=\; \delta^j_i} \tag{1.49}$$

were the indices j and i belong to the same set of indices and where

$$\delta^j_{\ i} \;=\; \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{1.50}$$

is the Kronecker symbol. Since the Kronecker symbol is symmetric under transposition, the order of the indices does not matter, which is the reason why the indices are positioned vertically on top of each other. As we will see below, the Kronecker symbol is in fact the only multiindex-object where the indices are positioned in this way.

The set of 1-forms $\{\mathbf{e}^j\}$ constitutes the basis of the dual vector space V^*. In fact, it is easy to see that *every* 1-form $\alpha \in V^*$ can be expressed as a linear combination of these dual basis forms:

$$\boxed{\alpha = \sum_j \alpha_j \mathbf{e}^j \qquad \text{where} \quad \alpha_j = \alpha(\mathbf{e}_j)\,.} \tag{1.51}$$

Proof: In order to show this we investigate how the 1-form α acts on the basis vectors. Using linearity one obtains $\alpha(\mathbf{e}_i) = \sum_j \alpha_j \mathbf{e}^j(\mathbf{e}_i) = \sum_j \alpha_j \delta^j_i = \alpha_i$, hence the linear coefficients are uniquely determined, and so it is clear that *every* 1-form can be represented in this way.

Hence for every basis $\{\mathbf{e}_i\}$ of V there is an associated well-defined basis $\{\mathbf{e}^j\}$ of V^* obeying the important property

$$\mathbf{e}^j \mathbf{e}_i = \delta^j_i.$$

This special basis is usually referred to as the *dual basis* of the co-vector space V^*.

In a finite dimensions, it is clear that V and V^* have exactly the same number of basis vectors and thus both spaces must have the same dimension. It is common to label the usual basis vectors of V with lower indices and the dual basis 1-forms in V^* with upper indices. Conversely, the components of vectors are written with upper indices while components the 1-forms carry lower indices, that is, oppositely. The summation is always carried out over pairs of upper and lower indices. According to the *Einstein sum convention*, it is customary to suppress the sum sign and to perform the sum over all pairs of oppositely positioned indices automatically.

Remember: A vector $\mathbf{v}$ can be represented as $\mathbf{v} = v^i \mathbf{e}_i$ with the components $v^i = \mathbf{e}^i(\mathbf{v})$.
A 1-form α can be represented as $\alpha = \alpha_i \mathbf{e}^i$ with the components $\alpha_i = \alpha(\mathbf{e}_i)$.

If we apply a 1-form α to a vector $\mathbf{v}$ this can be expressed in the representation as

$$\boxed{\alpha(\mathbf{v}) = \alpha_i v^i} \tag{1.52}$$

The result $\alpha(\mathbf{v})$ is called the *contraction* of α with $\mathbf{v}$ which is also denoted as $\iota_{\mathbf{v}}\alpha$ in the mathematical literature. There is also a variety of alternative notations that will be discussed below in Sect. 1.5.10 on page 24.

The summation over a pair of opposing indices reminds us of a scalar product, but it

is actually not a scalar product, because it is defined as a map acting on two different vector spaces V and V^*, whereas a scalar product acts on $V \times V$.

> **Remark:** in high school mathematics the representation of 1-forms is already introduced through the back door in the form of *line vectors*. For example, in textbooks one often writes
>
> $$\alpha(\mathbf{v}) \;=\; (\alpha_1, \alpha_2, \ldots) \begin{pmatrix} v^1 \\ v^2 \\ \ldots \end{pmatrix},$$
>
> performing the contraction according to the "row times column" rule. Also in quantum mechanics we are already familiar with 1-forms, which in the Dirac notation are written as as *bra vectors* $\langle\phi| \in \mathcal{H}^*$, and which are contracted with the state vectors $|\psi\rangle \in \mathcal{H}$ to a scalar $\langle\phi|\psi\rangle \in \mathbb{C}$. In many textbooks, such contractions are referred to as scalar products, which, strictly speaking, this is not correct because it is a contraction rather than a scalar product. We will return to this point later (see Sect. 1.6.1 on page 26).

We have introduced 1-forms as linear machines that are applied to vectors and yield as an output a number. But we can also view the vectors in the same way: Since the 1-forms by themselves form a dual vector space V^*, vectors by themselves be regarded as linear machines which, when applied to a 1-form, yield a number. It is therefore customary in differential geometry to treat both points of view synonymously by writing:

$$\mathbf{v}(\alpha) \;:=\; \alpha(\mathbf{v}). \tag{1.53}$$

This attitude of treating the vector space and its co-vector space on equal footing is quite unusual for beginners.

> **Remember:** A vector space V and the associated co-vector space V^* are closely related and have essentially the same structure. The high degree of symmetry a 1-form acting on a vector can also be interpreted as a vector acting on a 1-form.

1.5.3 How 1-forms change under basis transformations

Under a basis transformation $\mathbf{e}_i \;\to\; \mathbf{e'}_i \;=\; \mathbf{e}_k \tilde{M}^k{}_i$ (see Sect. 1.3.3 on page 9) the basis vectors of the dual space have to be transformed oppositely with the matrix $M = (\tilde{M})^{-1}$ by

$$\boxed{\mathbf{e}^j \;\to\; \mathbf{e'}^j = M^j{}_\ell \mathbf{e}^\ell}, \tag{1.54}$$

because only then that the new basis obeys the definition property $\mathbf{e'}^j(\mathbf{e'}_i) \;=\; \delta^j_i$ (cf. page 17).

> **Proof:** We can easily verify that $\mathbf{e'}^j(\mathbf{e'}_i) \;=\; M^j{}_\ell \tilde{M}^k{}_i \mathbf{e}^\ell(\mathbf{e}_k) \;=\; M^j{}_k \tilde{M}^k{}_i \;=\; \delta^j_i$.

Unlike the components of a vector, for which the transformation law

$$v^i \;\to\; v'^i = M^i{}_j v^j \tag{1.55}$$

applies, the components of the 1-form $\alpha = \alpha_j \mathbf{e}^j$ transform inversely according to

$$\boxed{\alpha_j \;\to\; \alpha'_j = \alpha_k \tilde{M}^k{}_j}. \tag{1.56}$$

In this expression one has to take care of the specific order of the indices.

Proof: Indeed we have $\alpha'_j \mathbf{e}'^j = \tilde{M}^k{}_j M^j{}_\ell \alpha_k \mathbf{e}^\ell = \alpha_\ell \mathbf{e}^\ell = \alpha$.

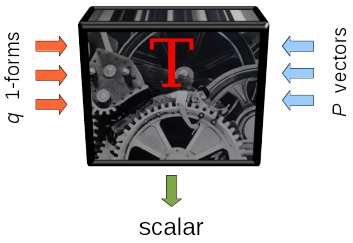
Remark: We may rewrite this transformation law in the usual notation as a matrix acting on α. Note that in the above expression α_k is contracted with the first index of $\tilde{M}$. This means that the components of α have to be written as a line vector on the left side of the matrix $\tilde{M}$.

1.5.4 Tensors

As we have seen above, for a given vector space V, the associated 1-forms constitute another vector space, namely, the so-called dual vector space V^*. This vector space V^* has exactly the same dimension as can be seen in the construction of the dual basis. Here we can already recognize that there is a high degree of symmetry between V and V^*, although at this point no isomorphism is available with which one could map vectors to 1-forms and vice versa. This symmetry becomes apparent in the more general concept of tensors of any order, which we will discuss below.

Let V be a linear vector space over the scalar field K and let V^* be the corresponding dual vector space of 1-forms. A so-called *tensor* is a *multilinear map* (i.e., a map which is linear each of its arguments)

$$\mathbf{T} : (V^*)^{\otimes q} \otimes (V)^{\otimes p} \to K,$$



Tensor of rank (q, p) as a 'black box'.

or roughly speaking a linear *black box* as shown in the figure which maps q 1-forms and p vectors to a number, i.e., a tensor has $q + p$ inputs and a single output. In order to be able to distinguish the enumeration of the inputs from the indexing of components, we write enumerate the arguments in parentheses in the opposite place, writing

$$\alpha^{(1)}, \ldots, \alpha^{(q)}, \mathbf{v}_{(1)}, \ldots, \mathbf{v}_{(p)} \mapsto \mathbf{T}(\alpha^{(1)}, \ldots, \alpha^{(q)}; \mathbf{v}_{(1)}, \ldots, \mathbf{v}_{(p)}).$$

The two parameters (q, p) are denoted as the *rank*, the *order*, or the *degree* of the tensor.[5] We already know a few special cases:

- A tensor of rank (0,0) that maps nothing to a number is a **scalar**.
- A tensor of rank (1,0) that maps a 1-form to a number is a **vector**.
- A tensor of rank (0,1), which maps a vector to a number, is a **1-form**.

If a tensor has more than one input slot, multilinearity means that the map is linear in each of the inputs. Depending on the number of inputs, one classifies the following types of tensors:

- ***Covariant** tensors* map exclusively vectors.
- ***Contravariant** tensors* map exclusively 1-forms.

[5] Some authors define the rank as the sum of the two numbers $p + q$.

- ***Mixed** tensors* map both vectors and 1-forms.

Just like 1-forms, tensors of the same rank can be added by simply adding their results:

$$\Big(\mathbf{T}_1 + \mathbf{T}_2\Big)\big(\{\alpha^{(i)}; \mathbf{v}_{(j)}\}\big) \;:=\; \mathbf{T}_1\big(\{\alpha^{(i)}; \mathbf{v}_{(j)}\}\big) + \mathbf{T}_2\big(\{\alpha^{(i)}; \mathbf{v}_{(j)}\}\big). \tag{1.57}$$

Likewise, they can be multiplied by a scalar by simply multiplying their result scalarly:

$$\Big(\lambda\mathbf{T}\Big)\big(\{\alpha^{(i)}; \mathbf{v}_{(j)}\}\big) \;:=\; \lambda\mathbf{T}\big(\{\alpha^{(i)}; \mathbf{v}_{(j)}\}\big). \tag{1.58}$$

The set of tensors of rank (q, p) equipped with these operations thus forms another vector space, denoted as

$$\bigotimes^{(q,p)} V\,. \tag{1.59}$$

Since tensors are defined as linear maps $(V^*)^{\otimes q} \otimes (V)^{\otimes p} \to K$, they are by definition elements of dual vector space of $(V^*)^{\otimes q} \otimes (V)^{\otimes p}$, hence

$$\bigotimes^{(q,p)} V \;=\; \Big((V^*)^{\otimes q} \otimes (V)^{\otimes p}\Big)^* = (V)^{\otimes q} \otimes (V^*)^{\otimes p}\,. \tag{1.60}$$

Please note that the star is on the right side of the tensor product. In other words, $\mathbf{T}$ is an element of $(V)^{\otimes q} \otimes (V^*)^{\otimes p}$, mapping elements of $(V^*)^{\otimes q} \otimes (V)^{\otimes p}$ to numbers.

1.5.5 Representation of tensors

Because of their linearity, tensors can be represented in a given basis $\{\mathbf{e}_i\}$ of V and the associated dual basis $\{\mathbf{e}^j\}$ of V^* in terms of components. To this end let a tensor of rank (q, p) act on q 1-forms and p vectors. In order to be able to distinguish their enumeration from the indexing of the components, we write the number of the argument in parentheses:

$$\begin{aligned}
&\mathbf{T}(\alpha^{(1)}, \ldots, \alpha^{(q)}\,;\, \mathbf{v}_{(1)}, \ldots, \mathbf{v}_{(p)}) \\
&= \mathbf{T}\Big(\sum_{j_1} \alpha^{(1)}_{j_1}\mathbf{e}^{j_1}, \ldots, \sum_{j_q} \alpha^{(q)}_{j_q}\mathbf{e}^{j_q}\,;\, \sum_{i_1} v^{i_1}_{(1)}\mathbf{e}_{i_1}, \ldots, \sum_{i_p} v^{i_p}_{(p)}\mathbf{e}_{i_p}\Big) \\
&= \sum_{j_1,\ldots,j_q,i_1,\ldots,i_p} \alpha^{(1)}_{j_1} \cdots \alpha^{(q)}_{j_q} v^{i_1}_{(1)} \cdots v^{i_p}_{(p)} \underbrace{\mathbf{T}(\mathbf{e}^{j_1}, \ldots, \mathbf{e}^{j_q}; \mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_p})}_{=:\, T^{j_1\ldots j_q}{}_{i_1\ldots i_p}}\,.
\end{aligned} \tag{1.61}$$

Using the Einstein sum convention we can write this more compactly as

$$\mathbf{T}(\alpha^{(1)}, \ldots, \alpha^{(q)}\,;\, \mathbf{v}_{(1)}, \ldots, \mathbf{v}_{(p)}) \;=\; T^{j_1\ldots j_q}{}_{i_1\ldots i_p}\, \alpha^{(1)}_{j_1} \cdots \alpha^{(q)}_{j_q}\, v^{i_1}_{(1)} \cdots v^{i_p}_{(p)} \tag{1.62}$$

According to the nomenclature introduced for 1-forms and vectors, the upper indices are called *contravariant*, while the lower ones are called *covariant*. The numbers

$$\boxed{T^{j_1\ldots j_q}{}_{i_1\ldots i_p} \;=\; \mathbf{T}(\mathbf{e}^{j_1}, \ldots, \mathbf{e}^{j_q}; \mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_p})} \tag{1.63}$$

are the *components* of the tensor $\mathbf{T}$.

1.5.6 Induced basis in the space of tensors $\bigotimes^{(q,p)} V$

Using these components, the tensor $\mathbf{T}$ can be represented as a linear combination of all products of the basis vectors

$$\mathbf{T} \;=\; T^{j_1\ldots j_q}{}_{i_1\ldots i_p}\, \mathbf{e}_{j_1} \otimes \ldots \otimes \mathbf{e}_{j_q} \otimes \mathbf{e}^{i_1} \otimes \ldots \otimes \mathbf{e}^{i_p} \,. \tag{1.64}$$

The tensor product of the basis vectors on the right hand side are the basis vectors of the vector space $\bigotimes^{(q,p)} V$ (see Eq. (1.60)) and shall be denoted by

$$\mathbf{E}_{k_1\ldots k_q}{}^{l_1\ldots l_p} := \mathbf{e}_{k_1} \otimes \ldots \otimes \mathbf{e}_{k_q} \otimes \mathbf{e}^{l_1} \otimes \ldots \otimes \mathbf{e}^{l_p} \,. \tag{1.65}$$

As they should, these basis vectors obey the property

$$\mathbf{E}_{k_1\ldots k_q}{}^{l_1\ldots l_p}\left(\mathbf{e}^{j_1}, \ldots, \mathbf{e}^{j_q}\,;\, \mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_p}\right) \;=\; \delta^{\,j_1,\ldots,j_q,l_1,\ldots,l_p}_{k_1,\ldots,k_q,i_1,\ldots,i_p} \,. \tag{1.66}$$

Using this notation the representation of $\mathbf{T}$ can be written in a compact form as

$$\boxed{\mathbf{T} \;=\; T^{j_1\ldots j_q}{}_{i_1\ldots i_p}\mathbf{E}_{j_1\ldots j_q}{}^{i_1\ldots i_p}\,,} \tag{1.67}$$

i.e., a tensor can always be represented as a linear combination of these basis vectors, where the linear coefficients are just the tensor components defined in Eq. (1.63):

The theory of relativity works with tensors up to a total rank of $q + p = 4$, i.e., no object has more then four indices. In a representation-dependent formulation of the theory, the "index gymnastics" that is already visible here quickly becomes confusing and error-prone, and indices can even obscure the physical meaning of a formula. This is another reason why it is advisable to aim for a presentation-independent formulation. Of course, any concrete calculation, for example on a computer, requires in most cases the use of an appropriate representation.

1.5.7 Tensors versus matrices

It is often said that tensors are like matrices and thus like linear maps. Strictly speaking, this is not correct, since a linear map $\mathbf{A} : V \to V$ maps one vector onto another vector, whereas a tensor $\mathbf{T}$ maps vectors and 1-forms onto a number. Nevertheless, there is a close connection. For example, let $\mathbf{A}$ be a linear map of a vector space V on itself, represented by a quadratic matrix $A^j{}_i$. Although this matrix looks like a mixed tensor of rank (1,1), this map returns a vector instead of a number. However, we could contract this resulting vector with another 1-form α in order to obtain a number. This would give us a tensor

$$\mathbf{T}_\mathbf{A}(\alpha, \mathbf{v}) \;:=\; \alpha(\mathbf{A}\mathbf{v}) \tag{1.68}$$

of rank (1,1), whose components $\mathbf{T_A}(\mathbf{e}^j, \mathbf{e}_i)$ are just the matrix elements $A^j_{\ i}$. In that broader sense, a tensor of rank (1,1) can indeed be interpreted as a linear map.

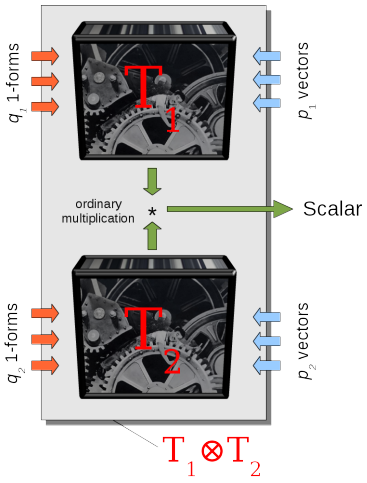
> **Remark:** This relationship is well-known in quantum theory: On the one hand, an operator $\mathbf{H}$ can be understood as a linear map $\mathcal{H} \to \mathcal{H}$ on the Hilbert space $\mathcal{H}$, but on the other hand it can also be considered as a bilinear map that maps a ket vector $|\psi\rangle \in \mathcal{H}$ and a bra vector $\langle\phi| \in \mathcal{H}^*$ to a number $\langle\phi|\mathbf{H}|\psi\rangle \in \mathbb{C}$.

1.5.8 Tensor product of tensors

The tensor product $\otimes$ combines two tensors to a new higher-rank tensor by simply multiplying their results, as shown in the figure. The new tensor $\mathbf{T}_1 \otimes \mathbf{T}_2$ has as many arguments as the original tensors together, i.e., the ranks of $\mathbf{T}_1$ and $\mathbf{T}_2$ simply add up. The tensor product can thus be understood as a map

$$\left(\bigotimes^{(q_1,p_1)} V\right) \otimes \left(\bigotimes^{(q_2,p_2)} V\right) \to \bigotimes^{(q_1+q_2,p_1+p_2)} V$$

and allows one to construct tensors of higher rank.



The tensor product of two tensors is again a tensor.

As an example we consider two linear forms α and β, which are tensors of rank (0,1). These tensors can be used to construct a bilinear form $\gamma = \alpha \otimes \beta$, which is a tensor of rank (0,2), by simply multiplying their results:

$$\gamma = \alpha \otimes \beta : \qquad \mathbf{v}_1, \mathbf{v}_2 \mapsto \gamma(\mathbf{v}_1, \mathbf{v}_2) \;=\; \alpha(\mathbf{v}_1)\beta(\mathbf{v}_2)\,. \tag{1.69}$$

Important: At this point it is important to note that *not all* tensors of rank (0,2) can be written in the form $\alpha \otimes \beta$. To see this, recall that the two 1-forms acting on $\mathbb{R}^3$ each have three degrees of freedom, so together six degrees of freedom, while a tensor of rank (0,2) has nine degrees of freedom. Therefore, the tensors of the form $\alpha \otimes \beta$ form only a subset of $\bigotimes^{(0,2)} V$, namely, the subset of the *factorizable tensors*. also called *product tensors*. Only by including all linear combinations of factorizable tensors does one obtain the entire vector space (see Sect. 1.4.3 on page 12). This means that any non-factorizable tensor can be written as a linear combination of product tensors.

By carrying out the tensor product several times, it is possible to successively generate tensors of arbitrarily high rank. So the tensor product is an algebraic operation that allows you to construct higher-rank tensors from elementary ones.

1.5.9 Representation of the tensor product

In a given basis $\{\mathbf{e}_i\}$ of V and $\{\mathbf{e}^i\}$ of V^*, the tensor product of two tensors is simply formed by multiplying the corresponding components in all possible combinations. For example, if $\alpha = \alpha_i \mathbf{e}^i$ and $\beta = \beta_j \mathbf{e}^j$, then their tensor product is simply represented by

$\alpha \otimes \beta = \alpha_i \beta_j (\mathbf{e}^i \otimes \mathbf{e}^j)$. Consequently, the bilinear form $\gamma = \alpha \otimes \beta$ is represented by

$$\boxed{\gamma = \alpha \otimes \beta \quad \Leftrightarrow \quad \gamma_{ij} = \alpha_i \beta_j.} \tag{1.70}$$

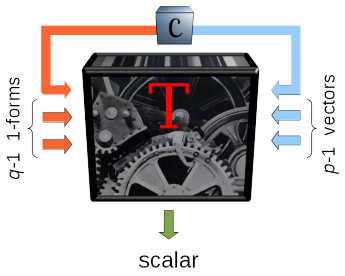
Analogously, the representation of a tensor product $\mathbf{C} = \mathbf{A} \otimes \mathbf{B}$ of two tensors of rank (q_1, p_1) and (q_2, p_2) can be computed by simply multiplying the components in all possible combinations, i.e.,

$$C^{i_1 \ldots i_{q_1} k_1 \ldots k_{q_2}}{}_{j_1 \ldots j_{p_1} \ell_1 \ldots \ell_{p_2}} = A^{i_1 \ldots i_{q_1}}{}_{j_1 \ldots j_{p_1}} B^{k_1 \ldots k_{q_2}}{}_{\ell_1 \ldots \ell_{p_2}}. \tag{1.71}$$

As one can easily see, we actually get a tensor of rank $(q_1 + q_2, p_1 + p_2)$, confirming that the ranks are additive under the tensor product.

1.5.10 Contraction

A *contraction* $\mathcal{C}$ (german "Verjüngung"), is a map that reduces the tensor's rank, so in comparison with the tensor product it works in opposite direction. Roughly speaking, a contraction can be thought of as short-circuiting two input channels of a tensor and tracing it out. Obviously, only contravariant with covariant inputs can be short-circuited in pairs in this way. Each contraction thus reduces the rank from (q, p) to $(q-1, p-1)$.



Contraction by short-circuiting opposite tensor slots.

As a simple example, let us consider a factorizable tensor $\mathbf{T}$ of rank (1,1) which has been constructed by the tensor product $\mathbf{T} = \mathbf{v} \otimes \alpha$ of a 1-form $\alpha \in V^*$ and a vector $\mathbf{v} \in V$. In this case, the contraction is simply carried out by applying the 1-form α to the vector $\mathbf{v}$, yielding a (0,0) tensor, that is, a scalar:

$$\mathcal{C}(\mathbf{v} \otimes \alpha) = \alpha(\mathbf{v}) \tag{1.72}$$

Similarly it is possible to contract non-factorizing tensors. For example, any non-factorizing tensor of rank (1,1) can always be written as a linear combination of factorizing tensors

$$\mathbf{T} = \sum_\mu \lambda_\mu \mathbf{v}_{(\mu)} \otimes \alpha^{(\mu)}, \tag{1.73}$$

where we used parenthesis in order to indicate that μ enumerates different vectors and forms rather than indexing their components.[6] Since the contraction is a linear operation which can be "passed through" to the summands, it is therefore possible to contract such non-factorizing tensor in the same way:

$$\mathcal{C}(\mathbf{T}) = \sum_\mu \lambda_\mu \mathcal{C}(\mathbf{v}_{(\mu)} \otimes \alpha^{(\mu)}) = \sum_\mu \lambda_\mu \alpha^{(\mu)}(\mathbf{v}_{(\mu)}) \tag{1.74}$$

The advantage of this definition is that it is completely representation-independent. An

[6] Note that in this expression μ is a label rather than an index. In particular, the sum may have an arbitrary number of summands.

alternative definition that is more useful for practical purposes is

$$\boxed{\mathcal{C}(\mathbf{T}) \;=\; \mathbf{T}(\mathbf{e}^i, \mathbf{e}_i)\,,} \tag{1.75}$$

where $\{\mathbf{e}^i\}$ and $\{\mathbf{e}_i\}$ are the basis vectors of V^* and V obeying $\mathbf{e}^i(\mathbf{e}_j) = \delta^i_j$ and where the summation is carried out as usual over the index i. Although this definition makes explicit use of a basis, one can show that it is also representation-independent since the result of the contraction is a (representation-independent) scalar.

Sketch of a proof: Having a look at the previous example $\mathbf{T} = \alpha \otimes \mathbf{v}$ we can easily convince ourselves that the two definitions are equivalent because of

$$\mathcal{C}(\mathbf{T}) = (\mathbf{v} \otimes \alpha)(\mathbf{e}^i, \mathbf{e}_i) = \alpha(\mathbf{e}_i)\mathbf{e}^i(\mathbf{v}) = \alpha_i v^i = \alpha(\mathbf{v}).$$

For higher-rank tensors, it is important to specify which of the inputs are actually short-circuited. In addition, it is possible to perform multiple contractions at the same time, that is, to short-circuit several pairs of inputs. The variety of possibilities has led to a somewhat confusing variety of notations in the literature. Here are a few examples:

$\mathcal{C}^k_\ell$ — This notation generalizes the above definition $\mathcal{C} = \mathcal{C}^1_1$, indicating that the k^{th} contravariant tensor component is contracted with the ℓ^{th} covariant tensor component. This means that in a representation the sum is carried out over the k^{th} upper index and the ℓ^{th} lower index.

$\mathcal{C}^{k_1\ldots k_m}_{\ell_1\ldots\ell_m}$ — Analogous notation for multiple contractions k_1-ℓ_1, k_2-ℓ_2, $\ldots$, k_m-ℓ_m.

$\langle \beta, \mathbf{A} \rangle$ — Complete p-fold contraction of a p-form with a p-vector in all slots. Although this notation is Dirac-like, it must not be confused with a scalar product. For this reason we will not use this notation.

$\iota_{\mathbf{A}}\beta$ — Another which uses the Greek letter 'iota'. This type of contraction is used for antisymmetric tensors and will be introduced in Sect. 2.1.10 on page 48. It differs from the notations listed above by additional combinatorial factors.

Remark: From a mathematical point of view, the Dirac bracket $\langle\phi|\psi\rangle$ used in quantum theory is not a scalar product but a contraction. The reason why the Dirac brackets behave effectively like a scalar product is a consequence of the so-called *musical isomorphism* to be discussed below.

1.5.11 Representation of a contraction

For a tensor of rank (1,1), the contraction can be represented in a given basis by

$$\mathcal{C}(\mathbf{T}) \;=\; \mathbf{T}(\mathbf{e}^i, \mathbf{e}_i) \;=\; T^i{}_i\,. \tag{1.76}$$

This means that a contraction is nothing but the *trace* over a pair of oppositely positioned indices. In the case of higher-rank tensors, in principle, any contravariant tensor component can be contracted with any covariant tensor component, and multiple contractions can be performed at one go by tracing over multiple pairs.

Examples:

- The contraction $\mathcal{C}_2^2\mathbf{T}$ of a tensor of rank $(2,2)$ is it represented in components by $T^{ik}{}_{jk}$, giving a tensor of rank (1,1).
- The contraction $\beta = \langle \alpha, \mathbf{X} \rangle$ of a vector $\mathbf{X}$ with a 2-form α gives a 1-form represented in covariant components by $\beta_j = X^i \alpha_{ij}$.
- The complete contraction $\langle \omega, \mathbf{T} \rangle = \mathcal{C}_{12}^{12}(\mathbf{T} \otimes \omega)$ of a contravariant tensor $\mathbf{T}$ of rank $(2,0)$ with the 4-form ω is represented by $T^{ij}\omega_{ijk\ell}$.
- Let $\mathbf{A}$ and $\mathbf{B}$ be mixed tensors of rank (1,1). Then the contraction $\mathcal{C}_1^2(\mathbf{A} \otimes \mathbf{B}) = A^i{}_j B^j{}_k$ gives again a tensor of rank (1,1) and looks formally like a matrix multiplication. In fact, a matrix multiplication is nothing but a contraction of the last index of the first matrix with the first index of the second matrix.

1.5.12 Tensor algebra

As already mentioned before, the tensors of the rank (q,p) are elements of a vector space $\bigotimes^{(q,p)} V$. The tensor product and the contraction, which allow us to increase and decrease the rank of tensors, connects these hierarchy of vectors spaces. It is therefore customary to combine these vector spaces by forming a direct sum in order to integrate all of them in a common vector space

$$\bigotimes V := \bigoplus_{q,p} \bigotimes^{(q,p)} V \,. \tag{1.77}$$

This total vector space of all tensors together with the rules for carrying out tensor products and contractions is referred to as the *tensor algebra*. An important feature of this algebra is that it does not close, i.e., by applying the tensor product repeatedly one can generate tensors of arbitrarily high rank (represented by objects with arbitrarily many indices). In this respect it differs significantly from the algebra of antisymmetric tensors to be discussed below which, as we will see, closes by itself.

> **Remark:** Note that in contrast to the tensor product, which can combine completely unrelated vector space (see Sect. 1.4.3 on page 12), the tensor algebra is built on a *single* vector space V and its co-vector space V^*, because only then is a contraction possible.

1.6 Metric

1.6.1 Metric tensor and scalar product

An important additional mathematical structure is the *scalar product*, also called *inner product*. A scalar product is a bilinear map $\mathbf{g} : V \times V \to K$ with the following properties:

1. Linearity (right): $\mathbf{g}(\mathbf{u}, \lambda\mathbf{v} + \mu\mathbf{w}) = \lambda\mathbf{g}(\mathbf{u}, \mathbf{v}) + \mu\mathbf{g}(\mathbf{u}, \mathbf{w})$
2. Symmetry: $\mathbf{g}(\mathbf{u}, \mathbf{v}) = \mathbf{g}(\mathbf{v}, \mathbf{u})$ for real vector spaces ($K = \mathbb{R}$) or
   $\mathbf{g}(\mathbf{u}, \mathbf{v}) = \mathbf{g}(\mathbf{v}, \mathbf{u})^*$ for complex vector spaces ($K = \mathbb{C}$).

3. A scalar product is positive definite, i.e., $\mathbf{g}(\mathbf{u},\mathbf{u}) \geq 0$ and $\mathbf{g}(\mathbf{u},\mathbf{u}) = 0$ if and only if $\mathbf{u} = 0$.

The linearity in the right argument together with the symmetry property directly implies the linearity (or antininearity in the complex case) in the left argument.

The best known example is the ordinary Euclidean scalar product in $\mathbb{R}^n$. In high school we learned that the scalar product of two vectors is the product of their lengths times the cosine of the included angle:

$$\mathbf{g}(\mathbf{u},\mathbf{v}) \;=\; \mathbf{u}\cdot\mathbf{v} \;=\; uv\cos\left[\sphericalangle(\mathbf{u},\mathbf{v})\right]. \tag{1.78}$$

Thus the scalar product has something to do with *lengths* of objects and *angles* between different objects. In is interesting to note that these concepts are not yet contained in the definition of a vector space, but they are only brought to life if we define a scalar product. Thus the scalar product is some kind of 'add on', but an important one because only then we are able to do *geometry*.

Recall that a (positively definite) scalar product always induces a *norm*

$$||\mathbf{v}|| \;:=\; \sqrt{\mathbf{g}(\mathbf{v},\mathbf{v})} \tag{1.79}$$

and that this norm induces in turn a *distance measure*[7]

$$d(\mathbf{u},\mathbf{v}) \;:=\; ||\mathbf{u}-\mathbf{v}||. \tag{1.80}$$

Recall: A norm $||\cdot||$ defined on the real or complex vector space V is a map $V \to \mathbb{R}$ with the following properties: For all $\mathbf{u},\mathbf{v} \in V$ and $\lambda \in K$ we have

- $||\mathbf{u}|| \geq 0\,; \qquad ||\mathbf{u}|| = 0 \;\Rightarrow \mathbf{u} = 0$
- $||\lambda\mathbf{u}|| = |\lambda|\;||\mathbf{u}||$
- $||\mathbf{u}+\mathbf{v}|| \leq ||\mathbf{u}|| + ||\mathbf{v}||$

A metric d is a distance measure $V \times V \to \mathbb{R}^+$ with the following properties:

- $d(\mathbf{u},\mathbf{u}) = 0\,; \qquad d(\mathbf{u},\mathbf{v}) = 0 \;\Rightarrow \mathbf{u} = \mathbf{v}$
- $d(\mathbf{u},\mathbf{v}) = d(\mathbf{v},\mathbf{u})$
- $d(\mathbf{u},\mathbf{w}) \leq d(\mathbf{u},\mathbf{v}) + d(\mathbf{v},\mathbf{w})$

In real vector spaces, to which we will restrict ourselves in the theory of relativity, the scalar product is a bilinear map that maps two vectors onto a real number. $\mathbf{g}$ is therefore a covariant symmetric tensor of rank $(0,2)$, which is referred to as *metric tensor*.

As we shall see, in the theory of relativity the inner product $\mathbf{g}$ or η is no longer positive definite, but one distinguishes *spacelike vectors* with $\mathbf{g}(\mathbf{u},\mathbf{u}) > 0$ and *timelike vectors* $\mathbf{g}(\mathbf{u},\mathbf{u}) < 0$ as well as the *light cone* $\mathbf{g}(\mathbf{u},\mathbf{u}) = 0$ which separates space- and timelike regions. Since the third postulate is violated, the metric of the 3+1-dimensional spacetime is strictly speaking not a metric in the mathematical sense, rather it is some kind of *pseudometric*. Nevertheless, in physics we continue to call $\mathbf{g}$ the scalar product, the metric tensor, or simple the metric.

[7] In mathematics, such a distance measure is called a 'metric'. In general relativity, however, the term 'metric' often refers to the scalar product g itself or the representation of the 'metric tensor'. As so often, physicists are not very careful with the nomenclature.

1.6.2 Representation of the metric tensor

Like any bilinear map represented in a given basis $\{\mathbf{e}_i\}$, the *metric tensor* $\mathbf{g}$ is completely defined by specifying how it acts on the basis vectors, i.e., in a given basis it is fully given in terms of the symmetric matrix

$$g_{ij} = g_{ji} = \mathbf{g}(\mathbf{e}_i, \mathbf{e}_j)\,. \tag{1.81}$$

A basis $\{\mathbf{e}_i\}$ is called *orthogonal* with respect to the metric $\mathbf{g}$, if $\mathbf{g}$ vanishes on different basis vectors so that the representation matrix g_{ij} is diagonal:

$$\mathbf{e}_i \cdot \mathbf{e}_j \;=\; g_{ij} \;=\; 0 \quad \forall i \neq j\,. \tag{1.82}$$

Furthermore, a basis is called *orthonormal* if

$$|\mathbf{e}_i \cdot \mathbf{e}_i| \;=\; |g_{ii}| \;=\; 1\,, \tag{1.83}$$

meaning that the diagonal elements of g_{ij} are ± 1. It is customary to sort the basis vectors by permutation in such a way that we start in the left upper corner with negative entries and then continue with the positive entries, i.e.,

$$g_{ij} \;=\; \mathrm{diag}(-1,\ldots,-1,+1,\ldots,+1). \tag{1.84}$$

This specific sequence of the signs is the so-called *signature* of the metric. In the theory of relativity, we are mainly dealing with the following signatures:

$(1,1,1,\ldots)$	*Riemann metric*
$(-1,1,1,\ldots)$	*Lorentzian metric*.[8]

For example, the three-dimensional spatial space $\mathbb{R}^3$ has a Riemann metric, while the four-dimensional space-time $\mathbb{R}^{1+3}$ used in special relativity is equipped with a Lorentz metric.

1.6.3 Examples

The ordinary scalar product in $\mathbb{R}^n$ is known as the *Euclidean metric*. In Cartesian coordinates, it is simply given by a unit matrix:

$$\mathbf{g}(\mathbf{e}_i, \mathbf{e}_j) = \delta_{ij}\,. \tag{1.85}$$

It has a positive signature and is therefore of Riemann-type.

In the special theory of relativity we will soon get into touch with the *Minkowski metric*, which is also denoted by the special symbol η because of its particular significance in frames where gravity is absent. The Minkowski metric is a *Lorentzian pseudometric* that is not positive definite. In the following we will use the "mostly plus" convention

[8] In these lecture notes we work with the *mostly plus* convention (-1,1,1,1). Many textbooks also work with the *mostly minus* convention (1,-1,-1,-1). Which to take is basically a matter of taste.

and define:

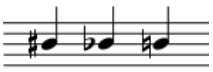
$$\eta_{ij} = g_{ij} = \text{diag}(-1, +1, +1, +1)\,. \tag{1.86}$$

Of course it is also possible to use the "mostly minus" convention

$$\Big[\eta_{ij} = g_{ij} = \text{diag}(+1, -1, -1, -1).\Big]$$

Both conventions are almost equally frequent in the literature. Thus, when reading, we always have to pay attention to which convention is actually used.[9]

1.6.4 Musical isomorphism $V \leftrightarrow V^*$



The symbols 'flat', 'sharp', and 'natural'.

The funny thing about the musical isomorphisms is that they have nothing to do with music at all. They are only called "musical" because they are denoted by the symbols 'sharp' (♯) and 'flat' (♭), which are used in music to raise or lower the pitch of a note by a half tone. In the present context they stand for raising or lowering indices.

♭-map:

Choosing a vector $\mathbf{u}$ as the left argument of $\mathbf{g}$, the resulting expression $\mathbf{g}(\mathbf{u}, \cdot)$ can be interpreted as a linear map from the vector plugged into the right slot onto a number, i.e., as a linear 1-form and thus as an element of the dual space V^*. The metric therefore induces a map

$$\flat : V \to V^* : \mathbf{u} \mapsto \mathbf{u}^\flat \tag{1.87}$$

which associates with each vector $\mathbf{u} \in V$ a corresponding unique 1-form $\mathbf{u}^\flat \in V^*$ with the property

$$\mathbf{u}^\flat(\mathbf{v}) = \mathbf{g}(\mathbf{u}, \mathbf{v}) \qquad \forall \mathbf{v} \in V\,. \tag{1.88}$$

Note that this map does **not necessarily map** a basis vector $\mathbf{e}_i$ to the corresponding dual 1-form $\mathbf{e}^i$, i.e., unless we consider special cases, we have in general:

$$\mathbf{e}_i^\flat \neq \mathbf{e}^i.$$

This can be seen by the fact that the dual basis is defined by $\mathbf{e}^i(\mathbf{e}_j) = \delta^i_j$ (see page 18) whereas $\mathbf{e}_i^\flat(\mathbf{e}_j) = \mathbf{g}(\mathbf{e}_i, \mathbf{e}_j)$.

> **Remark:** In high school mathematics, the ♭-map is already implicitly introduced by the transposition rule that converts a column vector into the corresponding row vector. In quantum mechanics, the ♭-map corresponds to converting a ket vector into a bra vector by transposition and complex conjugation.

Induced metric on V^*:

The map ♭ induces a dual scalar product $\mathbf{g}^*$ in V^*, which maps two 1-forms onto a

[9] In general relativity, a majority of researchers uses the "mostly plus" convention while the "mostly minus" convention is more popular in special relativity and related topics.

number. It is defined as a linear map $V^* \otimes V^* \mapsto K$ with the property

$$\mathbf{g}^*(\mathbf{u}^\flat, \mathbf{v}^\flat) \;:=\; \mathbf{g}(\mathbf{u}, \mathbf{v})\,, \tag{1.89}$$

where $\mathbf{u}^\flat, \mathbf{v}^\flat$ are the 1-forms which correspond to the vectors $\mathbf{u}, \mathbf{v}$. It is easy to verify that $\mathbf{u}^\flat, \mathbf{v}^\flat$ are unique and that $\mathbf{g}^*$ satisfies indeed the axioms of a scalar product on V^*.

Note that the dual norm $\mathbf{g}^*$ introduces the notion of length and angles for forms, which seems to be strange for beginners, but it works!

♯-map:

As with $\mathbf{g}$, which induces the map $\flat : V \to V^*$, the dual metric $\mathbf{g}^*$ induces another map in opposite direction

$$\sharp : V^* \to V : \alpha \mapsto \alpha^\sharp\,, \tag{1.90}$$

in that it associates with each 1-form $\alpha \in V^*$ a vector $\alpha^\sharp \in V$ with the property

$$\beta(\alpha^\sharp) = \mathbf{g}^*(\alpha, \beta) \qquad\qquad \forall \beta \in V^*\,. \tag{1.91}$$

One can show that $\mathbf{u}^{\flat\sharp} = \mathbf{u}$ and $\alpha^{\sharp\flat} = \alpha$, meaning that both maps a mutually inverse to each other:

$$\boxed{\flat\sharp = \sharp\flat = \mathbb{1}.} \tag{1.92}$$

Proof: The proof is straightforward:
$\mathbf{v}^\flat(\mathbf{u}^{\flat\sharp}) = \mathbf{g}^*(\mathbf{u}^\flat, \mathbf{v}^\flat) = \mathbf{g}(\mathbf{v}, \mathbf{u}) \;=\; \mathbf{g}(\mathbf{u}, \mathbf{v}) = \mathbf{v}^\flat(\mathbf{u}) \quad \forall \mathbf{v}^\beta \in V^* \qquad \Leftrightarrow \qquad \mathbf{u}^{\flat\sharp} = \mathbf{u}.$

Therefore the metric provides an isomorphism

$$V \underset{\sharp}{\overset{\flat}{\rightleftarrows}} V^* \tag{1.93}$$

which is called *canonical isomorphism* or, because of the idiosyncratic notation, also *musical isomorphism*. This isomorphism is the basis for the technique of *raising and lowering indices* in the theory of relativity, as will be discussed in the following section.

1.6.5 Representation of ♭ and ♯: Raising and lowering of indices

Let us now investigate the question how a vector $\mathbf{u} = u^j \mathbf{e}_j$ and the associated 1-form $\mathbf{u}^\flat = u_k \mathbf{e}^k$ manifest themselves in components.[10] To this end we apply $\mathbf{u}^\flat$ given in Eq. (1.88) to an arbitrary vector $\mathbf{v} = v^i \mathbf{e}_i$:

$$\begin{aligned} \mathbf{u}^\flat(\mathbf{v}) \;&=\; u_k \mathbf{e}^k(v^i \mathbf{e}_i) \;=\; u_k v^i \mathbf{e}^k(\mathbf{e}_i) = u_k v^i \delta^k_i = u_i v^i \\ &=\; \mathbf{g}(u^j \mathbf{e}_j, v^i \mathbf{e}_i) = u^j v^i \mathbf{g}(\mathbf{e}_j, \mathbf{e}_i) = u^j g_{ji} v^i = g_{ij} u^j v^i \end{aligned} \tag{1.94}$$

[10] It is customary to omit the ♭- and ♯-sign for the components since the position of the indices already tells us unambiguously whether we are dealing with a vector or the corresponding form, i.e., we write $u_i := u^\flat_i$ and likewise $\alpha^j := (\alpha^\sharp)^j$.

Since this equation holds for all $\mathbf{v} \in V$, a comparison of the coefficients on the right hand sides implies that

$$\boxed{\flat: \quad u_i = g_{ij} u^j} \tag{1.95}$$

i.e., g_{ij} is the transformation matrix which maps contravariant to covariant components, thus *lowering the index* of the vector components in order to obtain the components of the corresponding form.

The induced metric tensor $\mathbf{g}^*$ defined on the dual space V^* is represented by components

$$g^{ij} = g^{ji} = \mathbf{g}^*(\mathbf{e}^i, \mathbf{e}^j) \tag{1.96}$$

with upper indices. One can show that both tensors are inverse with respect to each other, i.e.

$$\boxed{g_{ij}\, g^{jk} = \delta_i^k} \tag{1.97}$$

which is a very important relationship that we should keep in mind.

> **Proof:** To prove this relation, we first examine the linear forms $\mathbf{e}_i^\flat$ which are associated with the basis vectors $\mathbf{e}_i$. As already outlined above, in general $\mathbf{e}_i^\flat \neq \mathbf{e}^i$, but on the other hand it is certainly possible to represent $\mathbf{e}_i^\flat$ as linear combinations of the basis vectors $\mathbf{e}^k \in V^*$, i.e., to find coefficients c_{ik} such that $\mathbf{e}_i^\flat = c_{ik}\mathbf{e}^k$. Applying both sides to the basis vector $\mathbf{e}_j \in V$, we can easily show that the coefficients are just given by $c_{ij} = \mathbf{e}_i^\flat(\mathbf{e}_j) = \mathbf{g}(\mathbf{e}_i, \mathbf{e}_j) = g_{ij}$. Consequently we have
>
> $$\mathbf{e}_i^\flat = g_{ij}\mathbf{e}^j\,.$$
>
> This leads us to the chain of equations $g_{ij} = \mathbf{g}(\mathbf{e}_i, \mathbf{e}_j) = \mathbf{g}^*(\mathbf{e}_i^\flat, \mathbf{e}_j^\flat) = g_{ik}g_{jm}\mathbf{g}^*(\mathbf{e}^k, \mathbf{e}^m) = g_{ik}g_{jm}g^{km}$, hence we have $g_{jm}g^{km} = \delta_j^k$, which completes the proof.

In the same way we can find out how the vector $\alpha^\sharp = \alpha^k \mathbf{e}_k$ associated with the form $\alpha = \alpha_j \mathbf{e}^j$ looks like in terms of its contravariant components. One obtains:

$$\boxed{\sharp: \quad \alpha^i = g^{ij}\alpha_j\,.} \tag{1.98}$$

With this calculation rule, an index is *raised*. For components the musical isomorphism is thus implemented as a very simple formal rule for raising and lowering indices. Now we understand why we use lower and upper indices in the theory of relativity.

Remember: Rules for lowering and raising indices:

- Vectors $\mathbf{v} = v^i \mathbf{e}_i$ are represented by **contra**variant components with **upper** indices.
- 1-forms $\alpha = \alpha_i \mathbf{e}^i$ are represented by **co**variant components with **lower** indices.
- Indices can be lowered by $v_i = g_{ij}v^j$ and raised by $v^i = g^{ij}v_j$.
- The indices of tensors can be raised or lowered separately. If several indices are moved, one needs a corresponding number of g-matrices.
  Example: $A^i_{\ jk} = g^{il}g_{jm}g_{kn}A_l^{\ mn}$.

1.6.6 Application of the musical operators to tensors

The musical isomorphisms can also be applied to tensors of higher rank. However, one has to specify which tensor component is to be raised or lowered. Also, it is possible to raise or lower several tensor components successively. While in the index representation this is quite easy, the abstract notation with $\flat$ and $\sharp$ turns out to be somewhat cumbersome.

An exception are purely contravariant and purely covariant tensors, i.e., tensors with exclusively upper or lower indices. For such tensors we will use the convention that the musical operators lower or raise **all** components simultaneously:

$$\begin{aligned} \text{purely contravariant:} \quad & \flat : \mathbf{T}^{i_1 \ldots i_q} \to \mathbf{T}_{i_1 \ldots i_q} \\ \text{purely covariant:} \quad & \sharp : \mathbf{T}_{i_1 \ldots i_p} \to \mathbf{T}^{i_1 \ldots i_p} \end{aligned} \tag{1.99}$$

1.6.7 Transformation behavior of the metric

Under a basis transformation, the metric tensor $\mathbf{g}$ defined on V and likewise its dual counterpart $\mathbf{g}^*$ on V^*, interpreted as abstract bilinear forms, remain unchanged, but their representation in terms of components changes according to

$$g_{ij} \;\to\; g'_{ij} = g_{k\ell} \tilde{M}^k{}_i \tilde{M}^\ell{}_j \tag{1.100}$$

and

$$g^{ij} \;\to\; g'^{ij} = M^i{}_k M^j{}_\ell g^{k\ell} \,. \tag{1.101}$$

If $\mathfrak{g}$ and $\mathfrak{g}^*$ denote the matrices g_{ij} and g^{ij}, these transformation laws can be written in a compact form as

$$\mathfrak{g}' = \tilde{M}^T \mathfrak{g} \tilde{M} \,, \qquad \mathfrak{g}^{*\prime} = M \mathfrak{g}^* M^T \,. \tag{1.102}$$

Remark: It can be verified easily, that $\mathfrak{g}$ and $\mathfrak{g}^*$ are inverse with respect to each other, both in the original and also in the new coordinates marked by the tics. This is because

$$g'^{ij} g_{jn} = M^i{}_k \underbrace{M^j{}_\ell \tilde{M}^r{}_j}_{=\delta^r_\ell} \tilde{M}^s{}_n g^{k\ell} g_{rs} = M^i{}_k \tilde{M}^s{}_n \underbrace{g^{k\ell} g_{\ell s}}_{=\delta^k_s} = m^i{}_k \tilde{M}^k_{spcn} = \delta^i_n$$

or in short $\mathfrak{g}' \mathfrak{g}^{*\prime} = \tilde{M}^T \mathfrak{g} \tilde{M} M \mathfrak{g}^* M^T = \mathbb{1}$.

1.6.8 Determinant of the metric

Since the metric tensor represented in coordinates can be viewed as a $d \times d$ matrix, where $d = \dim(V)$, it is near at hand that the determinant of this matrix plays a crucial role. This quantity is called g and it is defined as the determinant of $g_{i,j}$ with covariant indices:

$$\boxed{g = \det[\mathfrak{g}] = \det[(g_{ij})].} \tag{1.103}$$

Since the metric in contravariant components (g_{ij}) is the inverse of (g_{ij}), the corresponding determinant of the dual metric is the inverse (reciprocal) of g:

$$g^* = g^{-1} = \det[(g^{ij})]\,. \tag{1.104}$$

The sign of g^* depends on the signature of the metric. For example, in Riemann geometries such as the $\mathbb{R}^n$, g is positive, while it is negative in Lorentzian geometries.

In linear algebra we have learned that the determinant of a matrix is invariant under basis transformations. Not so in this case! In fact, it turns out that g is *representation-dependent* and transforms according to

$$g \;\to\; g' = \frac{g}{(\det M)^2} \tag{1.105}$$

where M is the transformation matrix (see Sect. 1.17 on page 10). Mathematically this is due to the fact that **g** is a (0,2)-form rather than a (1,1)-form, where the determinant would indeed by invariant. The transformation behavior can also be understood geometrically: If we rescale all coordinates by a factor λ, the matrix (g_{ij}) has to scale like λ^{-2} since only then the scalar product remains invariant. Hence g will be rescaled by λ^{-2d}, in accordance with the formula given above.

Interestingly, an infinitesimal volume element $\mathrm{d}^d x$, for example $\mathrm{d}V = \mathrm{d}^3 r = \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$ in $\mathbb{R}^3$, scales similarly but in opposite direction and without the square, namely

$$d^d x \;\to\; d^d x' = d^d x\,(\det M)\,. \tag{1.106}$$

This means that $d^d x$ is basis-dependent. However, we can neatly combine both quantities, g and $d^d x$, in order to define a representation-independent volume element $\mathrm{d}\omega$. namely,

$$\mathrm{d}\omega := \sqrt{|g|}\,\mathrm{d}^d x. \tag{1.107}$$

In general relativity, where g is negative, many authors prefer to write $\sqrt{-g}\,\mathrm{d}^d x$. As can be verified easily, the transformation factors cancel out under a basis transformation, proving that $\mathrm{d}\omega$ is in fact representation-independent. This is the reason why integrals over a 4-volume in general relativity, as we shall see later, typically looks like this:

$$\int \mathrm{d}^d x \sqrt{-g}\;\cdots \text{some integrand}\,,$$

since only then the value of the integral has a basis-independent meaning.

Remember:

- g is the determinant of the matrix of the metric tensor with (lower) indices.
- In the theory of relativity, the signature of the metric implies that g is always negative.
- g is not invariant under basis transformations, instead it scales $\det(M)^{-2}$.
- g can be used to define a representation-independent volume element $\mathrm{d}^d x\sqrt{-g}$.

1.6.9 Differentiating the determinant g with respect to g^{ij} or g_{ij}.

In general relativity, the components of the metric tensor are the elementary degrees of freedom of the gravitational field, i.e. they encode gravity. For this reason, one often encounters partial derivatives of these components in the theory. To ensure invariance under a coordinate transformations, the expressions to be differentiated are often 'decorated' with factors $\sqrt{-g}$, and using the product rule one often has to differentiate the determinant g with respect to one of the components g^{ij} or g_{ij}.

In order to solve this problem, we first have to understand how the determinant of a general matrix can be differentiated partially with respect to one of its components. Let A^{ij} be such a matrix and let

$$\det A = \epsilon_{k_1,\ldots,k_n} A^{k_1 1} \cdots A^{k_n n} \tag{1.108}$$

the corresponding determinant, where $\epsilon_{k_1,\ldots,k_n}$ denotes the fully antisymmetric Lev-Civita symbols. Taking now the derivative with respect to the component A^{ij}, one can show that

$$\frac{\partial}{\partial A^{ij}} \det A \;=\; (\det A) A^{-1}_{ji}\,, \tag{1.109}$$

where A^{-1} is the inverse of A, i.e., $A^{ij}A^{-1}_{jk} = \delta^i_k$.

Proof: Partially differentiating Eq. (1.108) with respect to A^{ij}, the only surviving products on the right hand side are those which preserve these components; it is then removed from the product by differentiating. The result can be written as

$$\frac{\partial}{\partial A^{ij}} \det A \;=\; \epsilon_{k_1,\ldots,k_{j-1},i,k_{j+1},\ldots,k_n}\, A^{k_1 1} \cdots A^{k_{j-1}\, j-1}\; /\!\!/\!\!/ \not{A}^{ij} /\!\!/\!\!/\; A^{k_{j+1}\, j+1} \cdots A^{k_n n}\,.$$

This equation is now multiplied on both sides by A^{ir} by summing over i. Using this trick one obtains the determinant on the right side:

$$\begin{aligned}\left[\frac{\partial}{\partial A^{ij}} \det A\right] A^{ir} \;&=\; \epsilon_{k_1,\ldots,k_{j-1},i,k_{j+1},\ldots,k_n}\, A^{k_1 1} \cdots A^{k_{j-1}\, j-1} A^{ir} A^{k_{j+1}\, j+1} \cdots A^{k_n n} \\ &=\; (\det A)\delta^r_j\,.\end{aligned}$$

This expression is then again multiplied on both sides by A^{-1}_{rs} while summing over r:

$$\left[\frac{\partial}{\partial A^{ij}} \det A\right] \delta^i_s \;=\; (\det A)\, A^{-1}_{js}\,,$$

hence

$$\frac{\partial}{\partial A^{ij}} \det A \;=\; (\det A)\, A^{-1}_{ji}\,.$$

Analogously, for a matrix B_{ij} with lower indices and the determinant $\det B$ formed from its components, the following formula applies:

$$\frac{\partial}{\partial B_{ij}} \det B \;=\; (\det B)[B^{-1}]^{ji}\,. \tag{1.110}$$

We now apply this result to the metric tensor by setting $B_{ij} = g_{ij}$. Since g_{ij} is inverse to

g^{ij} and these matrices are symmetric, one arrives at

$$\boxed{\frac{\partial}{\partial g_{ij}} g = g\, g^{ij}\,.} \tag{1.111}$$

Differentiating with respect to the upper components, we get $\frac{\partial}{\partial g^{ij}} g^{-1} = g^{-1}\, g_{ij}$ in an analogous way. On the other hand, $\frac{\partial}{\partial g^{ij}} g^{-1} = -g^{-2} \frac{\partial}{\partial g^{ij}} g$. Thus we get

$$\boxed{\frac{\partial}{\partial g^{ij}} g = -g\, g_{ij}\,.} \tag{1.112}$$

Remark: Note that it is impossible to get from Eq. (1.111) to Eq. (1.112) by simply lowering the indices, because there is an additional minus sign in the second equation. This is due to the fact that both equations are *not* valid tensor equations since g is not a scalar but representation-dependent number.

2 Differential forms

$$\wedge \qquad \iota \qquad \star \qquad \mathrm{d} \qquad \mathrm{d}^\dagger$$

This chapter focuses on the five symbols shown above: the outer product $\wedge$, the inner product ι, the Hodge star operator $\star$, and the outer derivative d; these operations are the essential building blocks in the theory of differential forms. They all operate on antisymmetrized tensors and form the so-called *exterior algebra*.

2.1 Exterior algebra

2.1.1 Exterior product (wedge product)

As pointed out in the previous chapter, the tensor product is not commutative. In fact, it is easy to verify that for two different vectors $\mathbf{v}, \mathbf{w} \in V$, the components of the two tensor products can be distinguished, i.e., $\mathbf{v} \otimes \mathbf{w} \neq \mathbf{w} \otimes \mathbf{v}$. In many situations, however, certain symmetry properties are required under permutation of the tensor slots. For example, we could symmetrize or antisymmetrize the tensor product, just in the same way as multi-particle wave functions in quantum mechanics are symmetrized (bosons) or antisymmetrized (fermions). However, the exchange of tensor slots under permutations requires that the objects living on the tensor slots are of the same kind, i.e., they have to be elements of the same vector space. In particular, it is impossible to exchange vectors with linear forms. Symmetrized or antisymmetrized tensor products are therefore either purely contravariant (formed only from vectors, only upper indices) or purely covariant (formed only from linear forms, only lower indices).

In the theory of relativity, antisymmetric tensors play an important role because they generalize the antisymmetric vector product (the cross product) in the position space $\mathbb{R}^3$. As we shall see, antisymmetric tensors provide the basis for many physical notions such as e.g. field lines. In fact, it seems that Nature somehow prefers antisymmetric tensors. Why this is so, it is an open question to be discussed below.

To construct such tensors, one defines a special antisymmetric variant of the tensor product, the so-called *outer* product, also known as *wedge product* since the product is denoted by the symbol '$\wedge$'. For two vectors $\mathbf{v}_1, \mathbf{v}_2 \in V$ the wedge product is defined by

$$\mathbf{v}_1 \wedge \mathbf{v}_2 \;:=\; \mathbf{v}_1 \otimes \mathbf{v}_2 - \mathbf{v}_2 \otimes \mathbf{v}_1. \tag{2.1}$$

More generally, the wedge product of n vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n \in V$ is defined as the sum over all permutations of the vectors combined with the corresponding signs of the per-

mutation:

$$\mathbf{v}_1 \wedge \mathbf{v}_2 \wedge \ldots \wedge \mathbf{v}_n \;:=\; \sum_{\sigma \in P_n} \operatorname{sgn}(\sigma) \bigotimes_{k=1}^{n} \mathbf{v}_{\sigma_k}\,, \tag{2.2}$$

For example, for $n = 3$ one obtains and antisymmetric rank-3 tensor consisting of $3! = 6$ summands:

$$\begin{aligned} \mathbf{v}_1 \wedge \mathbf{v}_2 \wedge \mathbf{v}_3 \;=\; & \mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \mathbf{v}_3 + \mathbf{v}_2 \otimes \mathbf{v}_3 \otimes \mathbf{v}_1 + \mathbf{v}_3 \otimes \mathbf{v}_1 \otimes \mathbf{v}_2 - \\ & \mathbf{v}_2 \otimes \mathbf{v}_1 \otimes \mathbf{v}_3 - \mathbf{v}_3 \otimes \mathbf{v}_2 \otimes \mathbf{v}_1 - \mathbf{v}_1 \otimes \mathbf{v}_3 \otimes \mathbf{v}_2\,. \end{aligned} \tag{2.3}$$

However, this definition would immediately lead to a problem if the wedge product is carried out subsequently one behind the other. For example, in the first step, following Eq. (2.1), we obtain

$$(\mathbf{v}_1 \wedge \mathbf{v}_2) \wedge \mathbf{v}_3 \;=\; (\mathbf{v}_1 \otimes \mathbf{v}_2 - \mathbf{v}_2 \otimes \mathbf{v}_1) \wedge \mathbf{v}_3 \;=\; (\mathbf{v}_1 \otimes \mathbf{v}_2) \wedge \mathbf{v}_3 - (\mathbf{v}_2 \otimes \mathbf{v}_1) \wedge \mathbf{v}_3.$$

If we now naively apply Eq. (2.1) to it the second time, we would get the (false) expression

$$(\mathbf{v}_1 \wedge \mathbf{v}_2) \wedge \mathbf{v}_3 \;=\; \mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \mathbf{v}_3 - \mathbf{v}_2 \otimes \mathbf{v}_1 \otimes \mathbf{v}_3 - \mathbf{v}_3 \otimes \mathbf{v}_1 \otimes \mathbf{v}_2 + \mathbf{v}_3 \otimes \mathbf{v}_2 \otimes \mathbf{v}_1$$

Compared to Eq. (2.3) there are two terms missing in this expression, and there are also wrong signs. Similarly, we get another wrong results when bracketing the second and the third factor:

$$\mathbf{v}_1 \wedge (\mathbf{v}_2 \wedge \mathbf{v}_3) \;=\; \mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \mathbf{v}_3 - \mathbf{v}_1 \otimes \mathbf{v}_3 \otimes \mathbf{v}_2 - \mathbf{v}_2 \otimes \mathbf{v}_3 \otimes \mathbf{v}_1 + \mathbf{v}_3 \otimes \mathbf{v}_2 \otimes \mathbf{v}_1$$

Thus we can conclude that the wedge product, extended to more than two factors, would not be associative.

This apparent contradiction arises from the naive iteration of the wedge product, whose mode of action was defined on only two vectors in Eq. (2.1), to a rank-2 tensor and a vector. This is obviously not correct. We therefore need to define the wedge product *depending on the rank of the tensors to join*, where Eq. (2.1) is only the special case for the rank 1-1.

Antisymmetrization operator

In order to construct an associative wedge product, let us first define an *antisymmetrization operator* $\mathcal{A}$ which completely antisymmetrizes a factorizing tensor of rank $(q, 0)$ by

$$\mathcal{A}\big[\mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \ldots \otimes \mathbf{v}_q\big] \;:=\; \frac{1}{q!} \sum_{\sigma \in P_q} \operatorname{sgn}(\sigma) \bigotimes_{k=1}^{q} \mathbf{v}_{\sigma_k}\,. \tag{2.4}$$

As can be seen, this operator antisymmetrizes all tensor components by summing over all possible permutations with the corresponding sign. The prefactor $1/q!$ ensures that

$$\mathcal{A}^2 = \mathcal{A}$$

so that a renewed antisymmetrization leaves the already antisymmetrized tensor unchanged. Since a general non-factorizing tensor of rank $(q,0)$ can always be represented as a linear combination of factorizing tensors, the definition (2.4) can also be applied to general tensors of rank q.

Equivalently, it is possible to permute the arguments of the tensor. For example, if a factorizing tensor $\mathbf{T}$ of rank $(q,0)$ acts on a product tensor of rank $(0,q)$, we can antisymmetrize $\mathbf{T}$ by permuting the factors in the argument:

$$\Big(\mathcal{A}\mathbf{T}\Big)(\alpha_1 \otimes \ldots \otimes \alpha_q) \;=\; \frac{1}{q!}\sum_{\sigma\in P_q} \operatorname{sgn}(\sigma)\, \mathbf{T}(\alpha_{\sigma_1} \otimes \ldots \otimes \alpha_{\sigma_q})\,. \tag{2.5}$$

Having introduced the antisymmetrization operator $\mathcal{A}$, the wedge product of two tensors $\mathbf{T}_1$ and $\mathbf{T}_2$ with the ranks q_1 and q_2 is defined as

$$\boxed{\mathbf{T}_1 \wedge \mathbf{T}_2 = \frac{(q_1+q_2)!}{q_1!q_2!}\mathcal{A}\Big[\mathbf{T}_1 \otimes \mathbf{T}_2\Big]} \tag{2.6}$$

As can be verified easily, the wedge product defined in this way is associative and compatible with the special case in Eq. (2.3). The wedge product thus not only antisymmetrizes the left and right argument *en bloc*, but it also antisymmetrizes all the internal tensor components contained in these blocks.

2.1.2 q-multivectors

The antisymmetric tensors constructed by the $\wedge$ product are called *factorizable* or *separable*. Although linear combinations of such tensors are again antisymmetric, they are not necessarily factorizable. However, as in the case of the ordinary tensor product (see Sect. 1.5.8 on page 23) it turns out that each antisymmetric tensor can be described as a finite linear combination of factorizable antisymmetric tensors. Such an antisymmetric tensor of rank $(q,0)$ is called q-*multivector*. Such q-multivectors obey the following properties:

Two vectors (rank $q_\mathbf{A} = q_\mathbf{B} = 1$):	$\mathbf{A}\wedge\mathbf{B} = \mathbf{A}\otimes\mathbf{B} - \mathbf{B}\otimes\mathbf{A}$
Multivectors with rank q_1 and q_2:	$\boxed{\mathbf{A}\wedge\mathbf{B} = \frac{(q_\mathbf{A}+q_\mathbf{B})!}{q_\mathbf{A}!q_\mathbf{B}!}\mathcal{A}\Big[\mathbf{A}\otimes\mathbf{B}\Big]}$
Associativity:	$(\mathbf{A}\wedge\mathbf{B})\wedge\mathbf{C} = \mathbf{A}\wedge(\mathbf{B}\wedge\mathbf{C})$
Linearity left:	$(\lambda\mathbf{A}+\mu\mathbf{B})\wedge\mathbf{C} = \lambda\mathbf{A}\wedge\mathbf{C} + \mu\mathbf{B}\wedge\mathbf{C}$
Linearity right:	$\mathbf{A}\wedge(\lambda\mathbf{B}+\mu\mathbf{C}) = \lambda\mathbf{A}\wedge\mathbf{B} + \mu\mathbf{A}\wedge\mathbf{C}$
Commutation relation:	$\boxed{\mathbf{A}\wedge\mathbf{B} = (-1)^{q_\mathbf{A}q_\mathbf{B}}\,\mathbf{B}\wedge\mathbf{A}}$

Proof: To prove the commutation relation, we first think of $\mathbf{A}$ and $\mathbf{B}$ as factorizable tensors. To get from$\mathbf{A}\wedge\mathbf{B}$ to $\mathbf{B}\wedge\mathbf{A}$, each tensor component of B must be commuted through each tensor component of A, which each time causes a minus sign. Overall, there are $q_A q_B$ of such swap processes, giving a total of a pre-factor $(-1)^{q_A q_B}$.

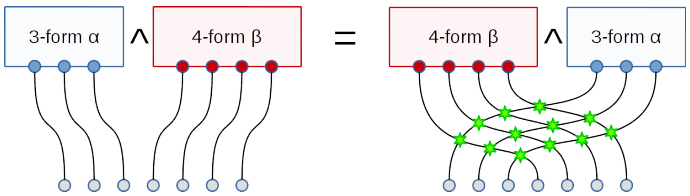


Figure 2.1: Illustration of the commutation rules for antisymmetric forms in the framework of the exterior algebra. Exchanging a 3-form α and a 4-form β amounts to 12 pairwise exchanges, marked by the green stars. Each exchange contributes with a minus sign. Since there are $3 \cdot 4 = 12$ exchanges involved, the total sign is $(-1)^{3\cdot4} = (-1)^{12} = +1$, hence $\alpha \wedge \beta = \beta \wedge \alpha$ in this case.

2.1.3 p-forms

In contrast to the ordinary tensor product, which can combine completely different vector spaces, the wedge product can only combine objects generated from the same vector space. Thus we can conclude that antisymmetric tensors must either be completely contravariant or completely covariant. Covariant antisymmetric tensors of rank p are referred to as *p-forms* and will play a central role in the theory of relativity. The wedge product on p-forms is defined analogously:

Two 1-forms:	$\alpha \wedge \beta = \alpha \otimes \beta - \beta \otimes \alpha$ (only for 1-forms)
Two general forms:	$\boxed{\alpha \wedge \beta = \frac{(p_1+p_2)!}{p_1!p_2!}\mathcal{A}\left[\alpha \otimes \beta\right]}$
Associativity:	$(\alpha \wedge \beta) \wedge \gamma = \alpha \wedge (\beta \wedge \gamma)$
Linearity left:	$(\lambda\alpha + \mu\beta) \wedge \gamma = \lambda\alpha \wedge \gamma + \mu\beta \wedge \gamma$
Linearity right:	$\alpha \wedge (\lambda\beta + \mu\gamma) = \lambda\alpha \wedge \beta + \mu\alpha \wedge \gamma$
Commutation relations:	$\boxed{\alpha \wedge \beta = (-1)^{p_\alpha p_\beta}\beta \wedge \alpha}$

The general definition can easily be extended from 2 to n factors: If $\alpha^{(1)}, \ldots, \alpha^{(n)}$ are forms of rank $p_1, \ldots, p_n$, then their wedge product is given by

$$\boxed{\alpha^{(1)} \wedge \ldots \wedge \alpha^{(n)} = \frac{(\sum_{i=1}^n p_i)!}{p_1!p_2!\cdots p_n!}\mathcal{A}\left[\alpha^{(1)} \otimes \ldots \otimes \alpha^{(n)}\right]} \tag{2.7}$$

Although the wedge product acts on multivectors and forms in a perfectly symmetric manner, it is quite common in differential geometry to regard the wedge product as preferentially acting on p forms, so there is some bias towards V^*. Therefore, if you look at a wedge product in the literature and it is not obvious whether it acts on forms or vectors, you can usually assume that it acts on forms.

2.1.4 Exterior algebra

The set of factorizable p-forms together with of all linear combinations constitutes a vector space, which is called *exterior power* of V^*. This space of general antisymmetric p-forms is denoted by $\bigwedge^p V^*$, with the special cases $\bigwedge^0 V^* = \mathbb{R}$ and $\bigwedge^1 V^* = V^*$. This means that the wedge product maps

$$\wedge : \bigwedge^m V^* \times \bigwedge^n V^* \to \bigwedge^{n+m} V^*$$

and thus establishes a relation between the vector spaces. Analogous to the tensor algebra (1.77), the direct sum of all these vector spaces

$$\bigwedge V^* := \bigoplus_p \bigwedge^p V^* \tag{2.8}$$

equipped with the wedge product is known as the *exterior algebra* or *Grassmann algebra* over V^*. The exterior algebra provides a set of vector spaces together with a closed and consistent set of computational rules for dealing with antisymmetric covariant tensors.

The exterior algebra closes

The antisymmetry has a strong limiting effect on the dimension of these vector spaces. To see this, let us assume that the underlying vector space V^* is finite-dimensional, having the basis $\{\mathbf{e}^1, \ldots, \mathbf{e}^n\}$. Then it is clear that each antisymmetric tensor of rank p can be represented as a linear combination of basis tensors

$$\mathbf{e}^{i_1} \wedge \ldots \wedge \mathbf{e}^{i_p},$$

where $i_1, \ldots, i_p \in \{1, 2, \ldots, n\}$. However, these basis tensors are nonzero if and only if all the indices are pairwise different. In addition, the transposition (exchange) of two indices gives only a minus sign, so that we can assume the set of indices to be ordered by $1 \leq i_1 < i_2 < \ldots < i_p \leq n$. The dimension of the $\bigwedge^p V^*$ is just the number of possibilities to select p values out of $\{1, 2, \ldots, n\}$, i.e.,

$$\dim\left(\bigwedge^p V^*\right) = \frac{n!}{p!(n-p)!} = \binom{n}{p}, \tag{2.9}$$

which sum up to

$$\sum_{p=0}^{n} \dim\left(\bigwedge^p V^*\right) = \sum_{p=0}^{n} \binom{n}{p} = 2^n. \tag{2.10}$$

In contrast to the usual tensor algebra $\otimes$, which allows us to create tensors of arbitrarily high rank (i.e., with any number of indices), the outer algebra *closes*, i.e., there are only finitely many (namely 2^n) linearly independent antisymmetric tensors whose rank is less than or equal to the dimension of the underlying vector space.

Example: In the vector space $\mathbb{R}^3$ there are three independent basis 1-forms $\mathbf{e}^1, \mathbf{e}^2, \mathbf{e}^3$, giving rise to three 2-forms $\mathbf{e}^1 \wedge \mathbf{e}^2$, $\mathbf{e}^1 \wedge \mathbf{e}^3$, $\mathbf{e}^2 \wedge \mathbf{e}^3$, as well as to a single 3-form $\mathbf{e}^1 \wedge \mathbf{e}^2 \wedge \mathbf{e}^3$, which is proportional to the volume form (see below). The total number of tensors is in fact equal to $1 + 3 + 3 + 1 = 8 = 2^3$.

2.1.5 Representation of p-forms

In Sect. 1.5.5 on page 21 we have seen that a tensor $\mathbf{T}$ of rank (q, p) can be represented in a given basis by

$$\mathbf{T} = T^{j_1 \dots j_q}{}_{i_1 \dots i_p} \mathbf{e}_{j_1} \otimes \dots \otimes \mathbf{e}_{j_q} \otimes \mathbf{e}^{i_1} \otimes \dots \otimes \mathbf{e}^{i_p} \tag{2.11}$$

with the components

$$T^{j_1 \dots j_q}{}_{i_1 \dots i_p} = \mathbf{T}(\mathbf{e}^{j_1}, \dots, \mathbf{e}^{j_q}; \mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_p}). \tag{2.12}$$

Analogously it is possible to represent an *antisymmetric* tensor of rank p by

$$\alpha = \alpha_{i_1, \dots, i_p} \mathbf{e}^{i_1} \otimes \dots \otimes \mathbf{e}^{i_p}. \tag{2.13}$$

with the components

$$\alpha_{i_1 \dots i_p} = \alpha(\mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_p}) \tag{2.14}$$

which are antisymmetric under the transposition of the indices. Of course we want to keep Eqs. (2.13)-(2.14) valid, but we would like to stay within the exterior algebra, avoiding the use of the ordinary tensor product $\otimes$. To do that, we have to rewrite Eq. (2.13) in terms of the wedge product. To this end we first use Eq. (2.7) to compute

$$\mathbf{e}^{i_1} \wedge \dots \wedge \mathbf{e}^{i_p} = \underbrace{\frac{(1 + \dots + 1)!}{1! \cdots 1!}}_{= p!} \mathcal{A}\left[\mathbf{e}^{i_1} \otimes \dots \otimes \mathbf{e}^{i_p}\right]. \tag{2.15}$$

On the other hand we know that α is antisymmetric, hence $\mathcal{A}[\alpha] = \alpha$. Since the antisymmetrization operator $\mathcal{A}$ is linear, we can apply it to Eq. (2.13), obtaining

$$\alpha = \alpha_{i_1, \dots, i_p} \mathcal{A}\left[\mathbf{e}^{i_1} \otimes \dots \otimes \mathbf{e}^{i_p}\right]. \tag{2.16}$$

Combining this relation with Eq. (2.15) we arrive at

$$\boxed{\alpha = \frac{1}{p!} \alpha_{i_1 \dots i_p} \mathbf{e}^{i_1} \wedge \dots \wedge \mathbf{e}^{i_p}.} \tag{2.17}$$

Compared with Eq. (2.13) the only formal difference is the combinatorial prefactor $\frac{1}{p!}$ which compensates for the double counting caused by antisymmetrization. Forgetting this combinatorial factor is a common mistake in the exterior calculus.

Remark: Note that although α is a fully antisymmetric form, there is no need for the components to be fully antisymmetric as well, that is, the corresponding matrix does not need to be fully antisymmetric. In fact, when adding non-antisymmetric contributions to $\alpha_{i_1 \dots i_p}$, this change will not appear in (2.17) since the antisymmetry of $\mathbf{e}^{i_1} \wedge \dots \wedge \mathbf{e}^{i_p}$ cancels these contributions automatically. In this sense the componetents are not unique.

2.1.6 Representation of the wedge product

Let α and β be forms of the rank p_α and p_β and let $\gamma = \alpha \wedge \beta$. According to the definition of the wedge product, γ is a $(p_\alpha + p_\beta)$-form given by

$$\alpha \wedge \beta = \frac{(p_\alpha + p_\beta)!}{p_\alpha! p_\beta!} \mathcal{A}\Big[\alpha \otimes \beta\Big].$$

How are the corresponding components related?

To answer this question, we use the representation of antisymmetric tensors given in Eq. (2.17):

$$\begin{aligned} \gamma = \alpha \wedge \beta &= \frac{1}{p_\alpha! p_\beta!} \alpha_{i_1 \ldots i_{p_\alpha}} \beta_{j_1 \ldots j_{p_\beta}} \left(\mathbf{e}^{i_1} \wedge \ldots \wedge \mathbf{e}^{i_{p_\alpha}}\right) \wedge \left(\mathbf{e}^{j_1} \wedge \ldots \wedge \mathbf{e}^{j_{p_\beta}}\right) \\ &= \frac{1}{p_\alpha! p_\beta!} \alpha_{i_1 \ldots i_{p_\alpha}} \beta_{j_1 \ldots j_{p_\beta}} \left(\mathbf{e}^{i_1} \wedge \ldots \wedge \mathbf{e}^{i_{p_\alpha}} \wedge \mathbf{e}^{j_1} \wedge \ldots \wedge \mathbf{e}^{j_{p_\beta}}\right) \end{aligned} \tag{2.18}$$

On the other hand we know that γ is represented by

$$\gamma = \frac{1}{(p_\alpha + p_\beta)!} \gamma_{i_1 \ldots i_{p_\alpha} j_1 \ldots j_{p_\beta}} \left(\mathbf{e}^{i_1} \wedge \ldots \wedge \mathbf{e}^{i_{p_\alpha}} \wedge \mathbf{e}^{j_1} \wedge \ldots \wedge \mathbf{e}^{j_{p_\beta}}\right). \tag{2.19}$$

Comparing both expressions allows us to express the wedge product in terms of the components by

$$\gamma_{i_1 \ldots i_{p_\alpha} j_1 \ldots j_{p_\beta}} = \frac{(p_\alpha + p_\beta)!}{p_\alpha! p_\beta!} \alpha_{i_1 \ldots i_{p_\alpha}} \beta_{j_1 \ldots j_{p_\beta}}. \tag{2.20}$$

Example:

- $\alpha \wedge \beta$:
  For two 1-forms $\alpha = \alpha_i \mathbf{e}^i$ and $\beta = \beta_i \mathbf{e}^i$ we have $\alpha \wedge \beta = \alpha_i \beta_j \mathbf{e}^i \wedge \mathbf{e}^j$. The actual result depends significantly on the dimension of the basic vector space. In one dimension we have $\alpha \wedge \beta = 0$, in two dimensions we have $\alpha \wedge \beta = (\alpha_1 \beta_2 - \alpha_2 \beta_1)(\mathbf{e}^1 \wedge \mathbf{e}^2)$, while in three dimensions we get three terms with a cross-product-like structure:
  $$\alpha \wedge \beta = (\alpha_1 \beta_2 - \alpha_2 \beta_1)(\mathbf{e}^1 \wedge \mathbf{e}^2) + (\alpha_1 \beta_3 - \alpha_3 \beta_1)(\mathbf{e}^1 \wedge \mathbf{e}^3) + (\alpha_2 \beta_3 - \alpha_3 \beta_2)(\mathbf{e}^2 \wedge \mathbf{e}^3)$$
- $\alpha \wedge \beta \wedge \gamma$:
  The triple wedge product vanishes in one and two dimensions. In three dimensions we get
  $$\alpha \wedge \beta \wedge \gamma \;=\; \alpha_i \beta_j \gamma_k \, \mathbf{e}^i \wedge \mathbf{e}^j \wedge \mathbf{e}^k \;=\; \alpha_i \beta_j \gamma_k \, \epsilon^{ijk} \, \mathbf{e}^1 \wedge \mathbf{e}^2 \wedge \mathbf{e}^3,$$
  hence the result is proportional to the volume form multiplied by the triple product (determinant) of the vector components:
  $$\alpha \wedge \beta \wedge \gamma \;=\; (\alpha_1 \beta_2 \gamma_3 - \alpha_1 \beta_3 \gamma_2 + \alpha_3 \beta_1 \gamma_2 - \alpha_3 \beta_2 \gamma_1 + \alpha_2 \beta_3 \gamma_1 - \alpha_3 \beta_2 \gamma_1) \, \mathbf{e}^1 \wedge \mathbf{e}^2 \wedge \mathbf{e}^3$$
- $(\alpha \wedge \beta)(\mathbf{u} \wedge \mathbf{v})$:
  If one applies the (antisymmetric) 2-form $\alpha \wedge \beta$ to the (antisymmetric) 2-vector $\mathbf{u} \wedge \mathbf{v}$,

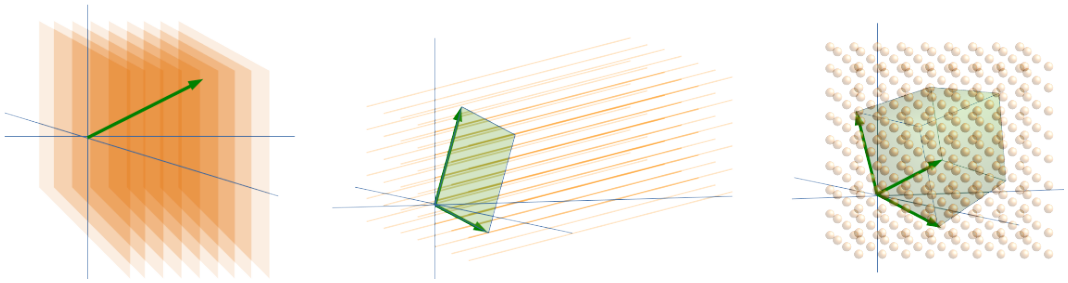


Figure 2.2: Visual interpretation of p-forms in the vector space $\mathbb{R}^3$. The figure illustrates the action of a 1-form (left), a 2-form (center), and a volume form (right), see text.

one obtains

$$\begin{aligned}(\alpha\wedge\beta)(\mathbf{u}\wedge\mathbf{v}) &= \alpha_i\beta_j u^k v^l\ (\mathbf{e}^i\wedge\mathbf{e}^j)(\mathbf{e}_k\wedge\mathbf{e}_l)\\ &= \alpha_i\beta_j u^k v^l\ (\mathbf{e}^i\otimes\mathbf{e}^j-\mathbf{e}^j\otimes\mathbf{e}^i)(\mathbf{e}_k\otimes\mathbf{e}_l-\mathbf{e}_l\otimes\mathbf{e}_k)\\ &= \alpha_i\beta_j u^k v^l\ (\delta^{ij}_{kl}-\delta^{ij}_{lk}-\delta^{ji}_{kl}+\delta^{ji}_{lk}) \;=\; (\alpha_i\beta_j-\beta_i\alpha_j)(u^iv^j-v^iu^j)\end{aligned}$$

- None-separable 2-form:
  The 2-form $\gamma := \mathbf{e}^1\wedge\mathbf{e}^2+\mathbf{e}^3\wedge\mathbf{e}^4$ cannot be factorized, i.e., it cannot be written in the form $\gamma=\alpha\wedge\beta$. Assuming the contrary, we would have $\alpha\wedge\gamma=\alpha\wedge\alpha\wedge\beta=0$ for all 1-forms $\alpha\neq 0$. However, an explicit calculation yields

  $$\alpha\wedge\gamma \;=\; \alpha_1\,\mathbf{e}^1\wedge\mathbf{e}^3\wedge\mathbf{e}^4+\alpha_2\,\mathbf{e}^2\wedge\mathbf{e}^3\wedge\mathbf{e}^4+\alpha_3\,\mathbf{e}^1\wedge\mathbf{e}^2\wedge\mathbf{e}^3+\alpha_4\,\mathbf{e}^1\wedge\mathbf{e}^2\wedge\mathbf{e}^4$$

  which vanishes if and only if $\alpha=0$.

2.1.7 Visual interpretation of the exterior algebra

Visual interpretation of p-forms

To gain more intuition it is very helpful to interpret p-forms in $\mathbb{R}^n$ visually as oriented $(n-p)$-dimensional hypersurfaces. Generally this interpretation works as follows: The p-vectors, which are put into the slots of a p-form as arguments, span a p-dimensional parallelepiped that pervades or encloses these hypersurfaces. The 'number' of the enclosed or punctured hypersurfaces is just the number that the p-form produces as its output. In Fig. 2.2 we illustrate how this interpretation looks like in the vector space $\mathbb{R}^3$:

- A 0-form is a scaler, i.e., it is simply a constant on the entire vector space.
- A 1-form in $\mathbb{R}^3$ can be thought of as a space-filling staggering of parallel planes. A vector to which the 1-form is applied (green arrow) penetrates a certain number of these planes. This number is the resulting value returned by the 1-form.
- A 2-form in $\mathbb{R}^3$ can be interpreted as as a space-filling staggering of parallel bars. The two vectors to which the 2-form is applied form a area element (the green parallelogram in the figure). The number of bars enclosed by this area element is the resulting value of the 2-form.
- A 3-form in $\mathbb{R}^3$ is the so-called volume form, which we will discuss in more detail in the following section. It can be interpreted as a regular arrangement of small

cells or spheres. The three vectors that serve as arguments span a certain spatula volume. The number of balls included in this volume element is the resulting value of the 3-form.

Visual interpretation of the wedge products

The wedge product combines two antisymmetric forms into a new antisymmetric form. As a first step towards a visual interpretation of the wedge product, let us consider the product $\gamma = \alpha \wedge \beta$ of two 1-forms. As pointed out above, each 1-form can be thought of as a staggering of parallel planes. Since both 1-forms have to be different (because otherwise $\gamma = 0$), their planes intersect to form square tubes or rods. In electrodynamics, these tubes are nothing but field lines, and therefore 2-forms are the appropriate mathematical object for describing field lines.

Similarly, the wedge product of a 2-form (staggered bars) with a 1-form (staggered planes) result into a regular patter of intersection points, representing the resulting 3-form.

Using this visual interpretation, it should be noted that the discretization in planes, rods, and spheres is for intuition only and that in reality all these objects are continuous. It is also important that one must imagine all geometric elements as being oriented. For example, the rods have a clearly defined sense of rotation, which determines whether they are counted positively or negatively when penetrating the surface element. A very detailed discussion of the geometric interpretation of arbitrary p-forms can be found in the classical book by Misner, Thorne and Wheeler [**?**].

2.1.8 The volume form ω

In an n-dimensional vector space the antisymmetric covariant n-form

$$\Omega \;:=\; \mathbf{e}^1 \wedge \mathbf{e}^2 \wedge \ldots \wedge \mathbf{e}^n \tag{2.21}$$

plays a special role, because on the one hand it has the highest possible rank, on the other hand there is (up to rescaling) only a single n-form of this kind, since its vector space $\bigwedge^n V^*$ (see Eq. (2.9)) is one-dimensional. In other words, all n-forms are proportional to Ω.

Since the above definition refers to a particular basis, Ω is representation-dependent and therefore cannot be used as an abstract sense. However, since there is only a single n-form except for rescaling, Ω cannot change by more than a scalar factor under a basis transformation. To calculate this scale factor we first examine how Ω transforms under a change of the basis. To this end let us define the antisymmetric *Levi-Civitá symbols* with lower indices

$$\boxed{\varepsilon_{i_1 \ldots i_n} = \begin{cases} 1 & \text{if } \{i_1, \ldots, i_n\} \text{ is an even permutation of } 1, \ldots, n, \\ -1 & \text{if } \{i_1, \ldots, i_n\} \text{ is an odd permutation of } 1, \ldots, n, \\ 0 & \text{otherwise (if at least one of the indices occurs twice).} \end{cases}} \tag{2.22}$$

In linear algebra it is well-known that the Levi-Civitá-symbols allow us to compute the determinant of a square matrix $A^i{}_j$ by

$$\det(A) \;=\; \sum_{i_1\ldots i_n} \varepsilon_{i_1\ldots i_n}\, A^1{}_{i_1} \ldots A^n{}_{i_n}\,. \tag{2.23}$$

The Levi-Civitá symbols with upper indices are *defined* (!) as

$$\boxed{\varepsilon^{i_1\ldots i_n} \;=\; s\,\varepsilon_{i_1\ldots i_n}\,,} \tag{2.24}$$

where

$$\boxed{s \;=\; \operatorname{sgn}\big[\det(\mathbf{g})\big] \;=\; \pm 1} \tag{2.25}$$

is the sign of the determinant of $\mathbf{g}$, i.e., the product of all signs occurring in the signature of the metric. It can be shown that s is invariant under basis changes, i.e., it is a representation-independent variable. In the Euclidean $\mathbb{R}^3$ we have $s = 1$, while in the four-dimensional theory of relativity we always have $s = -1$. The reason for this definition will become clear shortly.

In practice, the ϵ-symbols are very useful to isolate the antisymmetry in a wedge product as a prefactor and thus to bring the wedge product into a standard form:

$$\begin{aligned}
\mathbf{e}_{i_1} \wedge \mathbf{e}_{i_2} \wedge \cdots \wedge \mathbf{e}_{i_n} \;&=\; \epsilon_{i_1\ldots i_n}\, \mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \ldots \wedge \mathbf{e}_n \\
\mathbf{e}^{i_1} \wedge \mathbf{e}^{i_2} \wedge \cdots \wedge \mathbf{e}^{i_n} \;&=\; s\,\epsilon^{i_1\ldots i_n}\, \mathbf{e}^1 \wedge \mathbf{e}^2 \wedge \ldots \wedge \mathbf{e}^n\,.
\end{aligned}$$

Conversely one has

$$\begin{aligned}
\mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \ldots \wedge \mathbf{e}_n \;&=\; \frac{s}{n!}\, \epsilon^{i_1\ldots i_n}\, \mathbf{e}_{i_1} \wedge \mathbf{e}_{i_2} \wedge \cdots \wedge \mathbf{e}_{i_n} \\
\mathbf{e}^1 \wedge \mathbf{e}^2 \wedge \ldots \wedge \mathbf{e}^n \;&=\; \frac{1}{n!}\, \epsilon_{i_1\ldots i_n}\, \mathbf{e}^{i_1} \wedge \mathbf{e}^{i_2} \wedge \cdots \wedge \mathbf{e}^{i_n}\,.
\end{aligned}$$

> **Note: IMPORTANT:** $\varepsilon_{i_1\ldots i_n}$ and $\varepsilon^{i_1\ldots i_n}$ are *symbols*, not tensors! In fact, a tensor changes its components under basis transformations and therefore cannot have constant components. This is the reason why the $\varepsilon_{i_1\ldots i_n}$ are denoted as **symbols**. We will get back to a similar case further below when introducing the so-called *Christoffel symbols*, which also look like a tensor components although they do not transform like a tensor.

We now calculate how Ω transforms under a basis transformation. In the dashed basis of the dual space $\mathbf{e}'^j = M^i{}_j \mathbf{e}^j$ (see Sect. 1.54 on page 19) one obtains

$$\begin{aligned}
\Omega' \;&=\; \mathbf{e}'^1 \wedge \mathbf{e}'^2 \wedge \ldots \wedge \mathbf{e}'^n \\
&=\; M^1{}_{i_1} M^2{}_{i_2} \cdots M^n{}_{i_n}\;\; \mathbf{e}^{i_1} \wedge \mathbf{e}^{i_2} \wedge \ldots \wedge \mathbf{e}^{i_n} \\
&=\; M^1{}_{i_1} M^2{}_{i_2} \cdots M^n{}_{i_n}\;\; s\,\varepsilon^{i_1\ldots i_n}\;\; \mathbf{e}^1 \wedge \mathbf{e}^2 \wedge \ldots \wedge \mathbf{e}^n \;=\; \det(M)\,\Omega\,,
\end{aligned} \tag{2.26}$$

i.e. the determinant of M is the factor in question by which Ω changes under a basis transformations. This proves that Ω has no invariant meaning.

To eliminate the basis dependency in the above definition (2.21), it is necessary to multiply Ω by a reciprocal correction factor in such a way that $\det(M)$ drops out. A

perfect candidate is the determinant of the metric (see Sect. 1.105 on page 33) which transforms according to

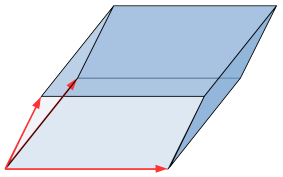
$$g \;\to\; g' = \frac{g}{\det(M)^2}. \tag{2.27}$$

Here the determinant of M appears as a square in the denominator. We can therefore cleverly combine the basis-dependent form $\Omega = \mathbf{e}^1 \wedge \ldots \wedge \mathbf{e}^n$ and the basis-dependent determinant g in such that the basis dependency drops out. This leads us to define the n-form

$$\boxed{\omega \;:=\; \sqrt{|g|}\;\mathbf{e}^1 \wedge \ldots \wedge \mathbf{e}^n\,.} \tag{2.28}$$

This definition is invariant under basis transformations, so it is valid in any basis and thus has an abstract representation-independent meaning. Note that the modulus under the square root is necessary because g can be negative for non-Euclidean geometries.[1] In the theory of relativity, where $g < 0$, one often writes $\omega = \sqrt{-g}\;\mathbf{e}^1 \wedge \ldots \wedge \mathbf{e}^n$.

The linear form $\omega = \sqrt{|g|}\mathbf{e}^1 \wedge \ldots \wedge \mathbf{e}^n$ is called *volume-form*. Indeed, applying ω to n vectors just yields the oriented (i.e., signed) volume of the n -dimensional parallelepiped spanned by these vectors.



Parallelepiped spanned by three vectors

Proof: To make this plausible, consider a parallelepiped spanned by three vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$ in $\mathbb{R}^3$ (see figure). Because of the invariance of ω, the choice of the basis does not matter, so we may use the standard orthonormal basis without restriction of generality. In this basis

$$\omega(\mathbf{a},\mathbf{b},\mathbf{c}) \;=\; a^i b^j c^k\, \omega(\mathbf{e}_i, \mathbf{e}_j.\mathbf{e}_k) \;=\; a^i b^j c^k\, \varepsilon_{ijk} \;=\; \begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix}$$

the spat product formed by the three vectors, which is known to be equal to the oriented volume of the parallelepiped. The corresponding sign is given by the well known 'right hand rule '.

Sometimes one also needs contravariant tensor $\omega^\sharp$, the dual which is the totally antisymmetric n multiplier you get by lifting all the tensor components. An analogous procedure as above leads to

$$\omega^\sharp \;=\; \frac{s}{\sqrt{|g|}}\;\mathbf{e}_1 \wedge \ldots \wedge \mathbf{e}_n\,, \tag{2.29}$$

where s again denotes the sign of the determinant of the metric.

2.1.9 Representation of the volume form

The volume form can be written as $\omega \;=\; \frac{\sqrt{|g|}}{n!}\,\epsilon_{i_1\ldots i_n}\,\mathbf{e}^{i_1} \wedge \ldots \wedge \mathbf{e}^{i_n}$, meaning that the tensor ω has the following components:

$$\boxed{\omega_{i_1\ldots i_n} \;=\; \sqrt{|g|}\,\epsilon_{i_1\ldots i_n}\,.} \tag{2.30}$$

[1] The volume form ω is often referred to in the literature as ϵ, but this can easily get into conflict with the Levi-Civitá symbols. Another common notation, as we shall see below, is $\star 1$, see Sect. 2.2.5 on page 54.

The corresponding dual tensor can be written as $\omega^\sharp = \frac{1}{\sqrt{|g|}n!}\,\epsilon^{i_1\ldots i_n}\,\mathbf{e}_{i_1}\wedge\ldots\wedge\mathbf{e}_{i_n}$, where $\epsilon^{i_1\ldots i_n} = s\epsilon_{i_1\ldots i_n}$. Consequently we have

$$\boxed{\omega^{i_1\ldots i_n} \;=\; \frac{1}{\sqrt{|g|}}\,\epsilon^{i_1\ldots i_n}\,.} \tag{2.31}$$

2.1.10 Contraction ι

The *contraction* of antisymmetric tensors is basically carried out in the same way as for general tensors. However, there are three special features of antisymmetric tensors that have to be taken into account:

- A contraction can be viewed as tracing out a contravariant with a covariant tensor slot. As there are no mixed antisymmetric tensors in the exterior algebra, this means that we can only contract q-multivectors with p-forms.
- Because of the antisymmetric structure it does not play a role (up to minus sign) which of the tensor components are contracted, since the sum over all permutation ensures that every tensor component is involved equally in the contraction process. This simplifies the notation considerably.
- The ordinary contraction, as it was introduced in Sect. 1.5.10 on page 24, renders combinatorically factors, e.g. we have $\mathcal{C}^{1\ldots n}_{1\ldots n}(\omega^\sharp\otimes\omega) = sn!$.

For these reasons, it is useful to introduce a separate contraction operator in the framework of the exterior algebra, which is better adapted to the notation and automatically compensates for the occurring combinatorial factors. This operator is denoted by the Greek letter *iota* (ι) and contracts all slots of a q-multivector $\mathbf{A}$ with a p-form α, assuming that $q\leq p$:

$$\boxed{\iota_{\mathbf{A}}\alpha \;:=\; \frac{1}{q!}\mathcal{C}^{1\ldots q}_{1\ldots q}(\mathbf{A}\otimes\alpha)} \tag{2.32}$$

Remember: The iota operator contracts q-multivectors with p-forms, where $q\leq p$. It differs from the ordinary contraction by a combinatorial factor $1/q!$, where q is the number of indices which are contracted.

Examples:

- The volume form contracted with its own dual:
$$\iota_{\omega^\sharp}\omega \;=\; \frac{1}{n!}\mathcal{C}^{1\ldots n}_{1\ldots n}(\omega^\sharp\otimes\omega) \;=\; \frac{1}{n!}\frac{\sqrt{|g|}}{\sqrt{|g|}}\epsilon^{i_1\ldots i_n}\epsilon_{i_1\ldots i_n} \;=\; s \;=\; \pm 1$$
- A vector $\mathbf{X}$ contracted with a factorizable 2-form $\alpha\wedge\beta$:
$$\iota_{\mathbf{X}}(\alpha\wedge\beta) \;=\; \iota_{\mathbf{X}}(\alpha\otimes\beta-\beta\otimes\alpha) \;=\; (\iota_{\mathbf{X}}\alpha)\beta-\alpha(\iota_{\mathbf{X}}\beta)$$
- A vector $\mathbf{X}$ contracted with a factorizable 3-form $\alpha\wedge\beta\wedge\gamma$:
$$\iota_{\mathbf{X}}(\alpha\wedge\beta\wedge\gamma) \;=\; (\iota_{\mathbf{X}}\alpha)(\beta\wedge\gamma)-(\iota_{\mathbf{X}}\beta)(\alpha\wedge\gamma)+(\iota_{\mathbf{X}}\gamma)(\alpha\wedge\beta)$$

It can be shown that the contraction of a vector $\mathbf{X}\in V$ with the wedge product of two

p-forms α, β with ranks p_α, p_β is given by

$$\iota_{\mathbf{X}}(\alpha \wedge \beta) \;=\; (\iota_{\mathbf{X}}\alpha) \wedge \beta \;+\; (-1)^{p_\alpha}\alpha \wedge (\iota_{\mathbf{X}}\beta)\,. \tag{2.33}$$

The contraction thus behaves formally like a signed product rule. It is particularly interesting to study the successive execution of various contractions. Here one finds

$$\iota_{\mathbf{X}} \circ \iota_{\mathbf{Y}} \;=\; -\iota_{\mathbf{Y}} \circ \iota_{\mathbf{X}}\,. \tag{2.34}$$

In particular, $\iota_{\mathbf{X}} \circ \iota_{\mathbf{X}} = 0$. Therefore, consecutive contractions *applied to antisymmetric tensors* behave formally in a similar way as the wedge product.

Proof: $\iota_{\mathbf{X}} \circ \iota_{\mathbf{X}}$ can be interpreted as $\iota_{\mathbf{X}\otimes\mathbf{X}}$. However, since $\mathbf{X}\otimes\mathbf{X}$ is a symmetric tensor, contraction with an antisymmetric tensor always gives zero. Similarly, the tensors $\mathbf{X}\otimes\mathbf{Y}$ und $\mathbf{Y}\otimes\mathbf{X}$, each contracted with the same antisymmetric tensor, can differ only by one sign.

2.1.11 Representation of the contraction ι in the exterior algebra

In a given basis $\{\mathbf{e}_i\}$ and the corresponding dual basis $\{\mathbf{e}^j\}$ we have

$$\iota_{\mathbf{e}_1\wedge\ldots\wedge\mathbf{e}_q}\; \mathbf{e}^1 \wedge \ldots \wedge \mathbf{e}^q = 1\,. \tag{2.35}$$

To prove this relation we compute

$$\begin{aligned}
&\mathcal{C}\big((\mathbf{e}_1 \wedge \ldots \wedge \mathbf{e}_q) \otimes (\mathbf{e}^1 \wedge \ldots \wedge \mathbf{e}^q)\big) \\
&\quad= \sum_{\sigma,\tau\in P_q} \epsilon^{\sigma_1\ldots\sigma_q}\epsilon_{\tau_1\ldots\tau_q}\mathcal{C}(\mathbf{e}_{\sigma_1}\otimes\ldots\otimes\mathbf{e}_{\sigma_q}\otimes\mathbf{e}^{\tau_1}\otimes\ldots\otimes\mathbf{e}^{\tau_q}) \\
&\quad= \sum_{\sigma,\tau\in P_q} \delta^{\tau_1}_{\sigma_1}\delta^{\tau_2}_{\sigma_2}\cdots\delta^{\tau_q}_{\sigma_q} \;=\; \sum_{\sigma\in P_q} 1 \;=\; q!
\end{aligned}$$

Similarly one can show that for $q \le p$ the following relation holds:

$$\iota_{\mathbf{e}_1\wedge\ldots\wedge\mathbf{e}_q}\; \mathbf{e}^1 \wedge \ldots \wedge \mathbf{e}^p = \mathbf{e}^{q+1} \wedge \ldots \wedge \mathbf{e}^p \tag{2.36}$$

Now let $\mathbf{A}$ be a q-multivector and α a p-form with $q \le p$. As can be shown (exercise), Eq. (2.32) can be represented in components by

$$\iota_{\mathbf{A}}\alpha \;=\; \frac{1}{q!(p-q)!}\,\alpha_{i_1\ldots i_p}\,\mathbf{A}^{i_1\ldots i_q}\;\mathbf{e}^{i_{q+1}}\wedge\ldots\wedge\mathbf{e}^{i_p}\,, \tag{2.37}$$

hence the components of the contraction result are given by:

$$[\iota_{\mathbf{A}}\alpha]_{i_{q+1}\ldots i_p} \;=\; \frac{1}{q!}\,\alpha_{i_1\ldots i_p}\,\mathbf{A}^{i_1\ldots i_q}\,. \tag{2.38}$$

Here one can see that $\iota_{\mathbf{A}}\alpha$ differs from the usual contraction by the factor $\frac{1}{q!}$. This compensates for the multiple occurrence of possible combinations.

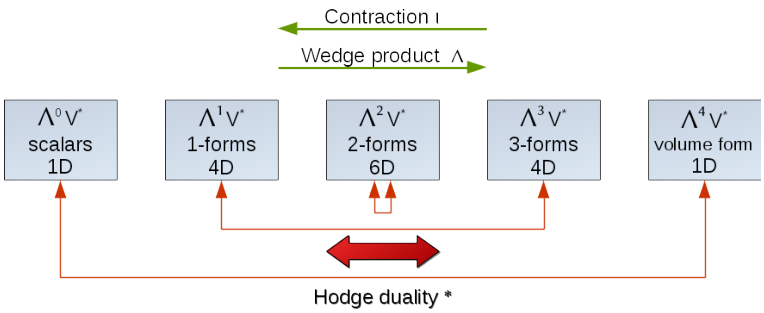
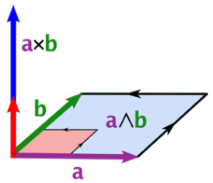


Figure 2.3: How the Hodge duality connects the spaces of antisymmetric forms in a four-dimensional vector space.

2.2 Hodge duality

2.2.1 Illustrative description of the Hodge duality

The *hodge-duality* came across to each of us indirectly in the form of the so-called *cross-product* in three-dimensional vector analysis. Instead of characterizing a surface element by two spanning vectors **a** and **b**, it can much more elegantly be represented by a normal vector **c** which is given by the cross product $\mathbf{c} = \mathbf{a} \times \mathbf{b}$. Remarkably, in this representation we only need three instead of six components to characterize the orientation and the size of the surface element.



Vector product (cross product)

The Hodge duality is based on the fact that the vector spaces $\bigwedge^p(V^*)$ and $\bigwedge^{n-p}(V^*)$, i.e., the linear spaces of the p-forms and the $(n-p)$-forms, have exactly the same dimension (cf. Eq. (2.9)):

$$\dim\left(\bigwedge^p V^*\right) = \binom{n}{p} = \binom{n}{n-p} = \dim\left(\bigwedge^{n-p} V^*\right) \tag{2.39}$$

The Hodge duality is a linear invertible transformation between these spaces for which one uses the symbol $\star$.

To compute the hodge-dual of a p-form, its inputs are simply contracted with the volume form ω, as sketched in the figure. Since the volume form itself has n inputs, there are still $n-p$ channels available, which are considered as the new inputs of $\star\alpha$. However, in this illustration is misleading in so far that as the covariant slots of the form α are contracted with the covariant inputs of the form ω, but as we have seen above, this is not allowed (only contravariant slots can be contracted with covariant ones). However, as we will see in the following, it is possible to contract two covariant slots indirectly by using the metric. To do this we first need the so-called generalized or *induced scalar product*.

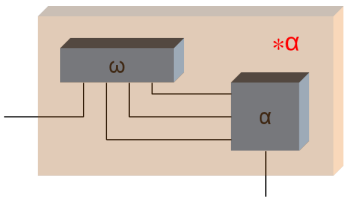


Illustration of the Hodge-dual of a 3-form in $\mathbb{R}^4$, creating a 1-form $\star\alpha$.

2.2.2 Induced scalar product on p-forms

On the space of the (antisymmetric) p -forms $\bigwedge^p V^*$ one can construct a scalar product

$$\mathbf{g}_p^* : \bigwedge^p V^* \times \bigwedge^p V^* \to \mathbb{R}$$

as follows: Let $\alpha = \alpha^{(1)} \wedge \ldots \wedge \alpha^{(p)}$ and $\beta = \beta^{(1)} \wedge \ldots \wedge \beta^{(p)}$ be two factorizable p forms. Then we define $\mathbf{g}^*$ as

$$\mathbf{g}_p^*(\alpha, \beta) \;:=\; \epsilon_{i_1 i_2 \ldots i_p} \prod_{j=1}^{p} \mathbf{g}^*(\alpha^{(j)}, \beta^{(i_j)}) \;=\; \det\{\mathbf{g}^*(\alpha^{(i)}, \beta^{(j)})\}. \tag{2.40}$$

On the right hand side we have the determinant of a $p \times p$ matrix formed by the results of the ordinary scalar product $\mathbf{g}^* = \mathbf{g}_1^*$ between all possible pairs of 1-forms $\alpha^{(i)}$ and $\beta^{(j)}$. It it is possible to prove that the product defined above indeed fulfills all the definition properties of a scalar product.

As usual, this scalar product is also defined on *non-factorizable* p-forms because such forms can always be written as finite linear combinations of factorizable forms, which allows us to compute the scalar product by using its bilinearity.

Many textbooks use a simplified notation for $\mathbf{g}_p^*$. For example, in many cases the subscript p and the asterisk are omitted because the type and the rank of the arguments clearly indicates which scalar product we have to use. Another widespread notation uses angle brackets $\langle \cdot, \cdot \rangle$, reminding of the Dirac notation in quantum mechanics.

Having defined a scalar products for forms, we can do geometry with forms. For example, it is possible to define the norm of a p -form by $||\alpha|| = \sqrt{|g^*(\alpha, \alpha)|}$ or to get a criterion telling us whether two p-forms are "perpendicular" to each other or not. An illustrative interpretation, however, is not always that easy.

Examples:

- As an example, let us consider the $\mathbb{R}^3$ with the ordinary Cartesian scalar product. As pointed out above, the 2-forms $\mathbf{e}^1 \wedge \mathbf{e}^2$ and $\mathbf{e}^1 \wedge \mathbf{e}^3$ can be interpreted as rods in the z or y direction. The scalar product of these two 2-forms is

  $$\mathbf{g}^*(\mathbf{e}^1 \wedge \mathbf{e}^2, \mathbf{e}^1 \wedge \mathbf{e}^3) \;=\; \begin{vmatrix} \mathbf{g}^*(\mathbf{e}^1, \mathbf{e}^1) & \mathbf{g}^*(\mathbf{e}^1, \mathbf{e}^3) \\ \mathbf{g}^*(\mathbf{e}^2, \mathbf{e}^1) & \mathbf{g}^*(\mathbf{e}^2, \mathbf{e}^3) \end{vmatrix} \;=\; \begin{vmatrix} g^{11} & g^{13} \\ g^{21} & g^{23} \end{vmatrix} \;=\; \begin{vmatrix} 1 & 0 \\ 0 & 0 \end{vmatrix} \;=\; 0\,,$$

  that is, the two 2-forms are in fact "perpendicular" on each other.

- The norm of the volume form on $\mathbb{R}^3$ induced by the scalar product defined above is given by

  $$||\omega|| \;=\; \sqrt{\mathbf{g}^*(\omega, \omega)} \;=\; \sqrt{|\det g|}\, \sqrt{\mathbf{g}^*(\mathbf{e}^1 \wedge \mathbf{e}^2 \wedge \mathbf{e}^3,\, \mathbf{e}^1 \wedge \mathbf{e}^2 \wedge \mathbf{e}^3)}$$
  $$= \sqrt{|\det g|}\, \begin{vmatrix} g^{11} & g^{12} & g^{13} \\ g^{21} & g^{22} & g^{23} \\ g^{31} & g^{32} & g^{33} \end{vmatrix}^{1/2} \;=\; 1$$

  which also makes a lot of sense.

In the last section, we have seen that the consecutive execution of contractions (applied to antisymmetric tensors) behaves formally like a wedge product. Having introduced

the generalized scalar product, it can be shown that both operations are indeed dual to one another. To this end let α be a 1-form, β a p-form, and γ a $p+1$-form. Then

$$\boxed{\mathbf{g}^*(\alpha \wedge \beta, \gamma) \;=\; \mathbf{g}^*(\beta, \iota_{\alpha^\sharp}\gamma)\,.} \tag{2.41}$$

Note that on the left side of the equation we have a scalar product of two $p+1$-forms, while on the right side there is a scalar product of two p-forms. This relation shows us that we can convert a wedge product in one of the arguments of the generalized scalar product into a contraction in the other argument.

Proof: To convince ourselves, consider the case $p = 1$ and assume that $\gamma = \eta \wedge \rho$ is factorizable. Then the left side reads

$$\mathbf{g}^*(\alpha \wedge \beta, \eta \wedge \rho) = \mathbf{g}^*(\alpha,\eta)\mathbf{g}^*(\beta,\rho) - \mathbf{g}^*(\alpha,\rho)\mathbf{g}^*(\beta,\eta) = \alpha^i\eta_i\beta^j\rho_j - \alpha^i\rho_i\beta^j\eta_j$$

Because of $\iota_{\alpha^\sharp}(\eta \wedge \rho) = [\alpha^\sharp(\eta)]\rho - [\alpha^\sharp(\rho)]\eta$ the right side

$$\mathbf{g}^*(\beta, \iota_{\alpha^\sharp}(\eta \wedge \rho)) = [\alpha^\sharp(\eta)]\mathbf{g}^*(\beta,\rho) - [\alpha^\sharp(\rho)]\mathbf{g}^*(\beta,\eta) = \alpha^i\eta_i\beta^j\rho_j - \alpha^i\rho_i\beta^j\eta_j$$

gives the same result.

2.2.3 Representation of the generalized scalar product

Let

$$\alpha = \frac{1}{p!}\alpha_{i_1\ldots i_p}\mathbf{e}^{i_1} \wedge \ldots \wedge \mathbf{e}^{i_p}\,, \qquad \beta = \frac{1}{p!}\beta_{j_1\ldots j_p}\mathbf{e}^{j_1} \wedge \ldots \wedge \mathbf{e}^{j_p}\,, \tag{2.42}$$

be two p-forms. Then, according to Eq. (2.40) the generalized scalar product is given by

$$\begin{aligned}
\mathbf{g}^*(\alpha,\beta) &= \frac{1}{(p!)^2}\,\alpha_{i_1\ldots i_p}\,\beta_{j_1\ldots j_p}\,\underbrace{\mathbf{g}^*(\mathbf{e}^{i_1} \wedge \ldots \wedge \mathbf{e}^{i_p},\, \mathbf{e}^{j_1} \wedge \ldots \wedge \mathbf{e}^{j_p})}_{=\det_{kl}(\mathbf{g}^*(\mathbf{e}^{i_k},\mathbf{e}^{j_l}))} \\
&= \frac{1}{(p!)^2}\,\alpha_{i_1\ldots i_p}\,\beta_{j_1\ldots j_p}\,\epsilon_{k_1\ldots k_p}\prod_{r=1}^{p} g^{i_r j_{k_r}} \\
&= \frac{1}{(p!)^2}\,\alpha^{j_{k_1}\cdots j_{k_r}}\,\beta_{j_1\ldots j_p}\,\epsilon_{k_1\ldots k_p} \;=\; \frac{1}{p!}\,\alpha^{j_1\ldots j_p}\,\beta_{j_1\ldots j_p} \;=\; \iota_{\alpha^\sharp}\beta
\end{aligned} \tag{2.43}$$

Therefore, calculating the scalar product of two forms of the same rank amounts to contracting all pairs of indices, which is quite intuitive.

2.2.4 Hodge duality on the basis of the generalized scalar product

Behind the fact that $\iota_{\mathbf{X}} \circ \iota_{\mathbf{Y}} = -\iota_{\mathbf{Y}} \circ \iota_{\mathbf{X}}$, telling us that contractions executed one after the other work formally like a wedge product, there is a special symmetry of the exterior algebra, which is known as the *Hodge duality*. To understand the Hodge duality, which should not be confused with the 'usual' duality $V \leftrightarrow V^*$, we first need a lemma:

Lemma: Let V be a vector space equipped with a scalar product $\mathbf{g}$ and let $f : V \to \mathbb{R}$ be a given linear function. Then there is a unique vector $\mathbf{u} \in V$

such that

$$f(\mathbf{v}) \;=\; \mathbf{g}(\mathbf{v},\mathbf{u}) \qquad \forall \mathbf{v} \in V \tag{2.44}$$

Proof: In an orthonormal basis we have $f(\mathbf{e}_i) = \mathbf{g}(\mathbf{e}_i,\mathbf{u}) = g_{ij}u^j$. If we multiply with the inverse matrix g^{ki} from the left we obtain $u^k = g^{ki}f(\mathbf{e}_i)$, hence $\mathbf{u} = g^{ki}f(\mathbf{e}_i)\mathbf{e}_k$.

Remark: You may already know this lemma from quantum mechanics or mathematics as the ***theorem by Riesz***: For every linear functional $F[\psi]$ acting on a wavefunction ψ there exists another wavefunction $|\phi\rangle$ in such a way that $F[\psi] = \langle\phi|\psi\rangle$.

This lemma holds in every vector space. We now apply it to the vector space $\wedge^p V^*$ equipped with the generalized scalar product introduced above. Let α be a fixed p -form and β an arbitrary $n-p$-form. This implies that $\alpha\wedge\beta$ is an n-form. Since the space $\wedge^n V^*$ is one-dimensional, $\alpha\wedge\beta$ must be proportional to the volume form (see Sect. 2.1.8 on page 45), i.e.,

$$\alpha\wedge\beta \;=\; f_\alpha(\beta)\omega\,, \tag{2.45}$$

where f_α is a linear function. Then, according to the lemma, there exists a unique $n-p$-form γ_α such that

$$f_\alpha(\beta) \;=\; \mathbf{g}^*(\beta,\gamma_\alpha)$$

which, inserted into (2.45) allows us to express the wedge product by

$$\alpha\wedge\beta \;=\; \mathbf{g}^*(\beta,\gamma_\alpha)\omega \qquad \forall\beta\in\wedge^{n-p}V^*\,. \tag{2.46}$$

Thus we can assign to each p-form α a unique $n-p$ -form γ_α. In fact, the dimensions of the corresponding vector spaces

$$\dim\Big(\bigwedge\nolimits^p(V^*)\Big) = \binom{n}{p}\,, \qquad \dim\Big(\bigwedge\nolimits^{n-p}(V^*)\Big) = \binom{n}{n-p}\,, \tag{2.47}$$

are identical since the corresponding binomial coefficients are invariant under the replacement $p\leftrightarrow n-p$.

Example: In order to understand this construction, let us consider the vector space $\mathbb{R}^3$ with Cartesian coordinates. Furthermore let α be a 1-form and β a 2-form with the representations

$$\alpha \;=\; \alpha_1\mathbf{e}^1+\alpha_2\mathbf{e}^2+\alpha_3\mathbf{e}^3 \tag{2.48}$$

$$\beta \;=\; \beta_{12}(\mathbf{e}^1\wedge\mathbf{e}^2)+\beta_{13}(\mathbf{e}^1\wedge\mathbf{e}^3)+\beta_{23}(\mathbf{e}^2\wedge\mathbf{e}^3)\,. \tag{2.49}$$

Then we have

$$\begin{aligned}\alpha\wedge\beta \;&=\; \alpha_1\beta_{23}\,(\mathbf{e}^1\wedge\mathbf{e}^2\wedge\mathbf{e}^3)+\alpha_2\beta_{13}\,(\mathbf{e}^2\wedge\mathbf{e}^1\wedge\mathbf{e}^3)+\alpha_3\beta_{12}\,(\mathbf{e}^3\wedge\mathbf{e}^1\wedge\mathbf{e}^2) \\ &=\; \underbrace{(\alpha_1\beta_{23}-\alpha_2\beta_{13}+\alpha_3\beta_{12})}_{=f_\alpha(\beta)}\,\mathbf{e}^1\wedge\mathbf{e}^2\wedge\mathbf{e}^3\,.\end{aligned} \tag{2.50}$$

According to the lemma there exists the 2-form

$$\gamma = \gamma_{12}(\mathbf{e}^1\wedge\mathbf{e}^2)+\gamma_{13}(\mathbf{e}^1\wedge\mathbf{e}^3)+\gamma_{23}(\mathbf{e}^2\wedge\mathbf{e}^3)$$

such that $\mathbf{g}_2^*(\gamma,\beta) = f_\alpha(\beta)$. For example, the generalized scalar product acting on 2-forms is given by

$$\mathbf{g}_2^*(\mathbf{e}^1\wedge\mathbf{e}^2,\mathbf{e}^1\wedge\mathbf{e}^2) \;=\; \begin{vmatrix}\mathbf{g}^*(\mathbf{e}^1,\mathbf{e}^1) & \mathbf{g}^*(\mathbf{e}^1,\mathbf{e}^2)\\ \mathbf{g}^*(\mathbf{e}^2,\mathbf{e}^1) & \mathbf{g}^*(\mathbf{e}^2,\mathbf{e}^2)\end{vmatrix} = \begin{vmatrix}1 & 0\\ 0 & 1\end{vmatrix} = 1$$

and analogously by $\mathbf{g}_2^*(\mathbf{e}^i\wedge\mathbf{e}^j,\mathbf{e}^k\wedge\mathbf{e}^l) \;=\; \delta^{ik}\delta^{jl}$ für $i\neq j$. For this reason we get

$$\mathbf{g}_2^*(\gamma,\beta) = f_\alpha(\beta) = \gamma_{12}\beta_{12}+\gamma_{13}\beta_{13}+\gamma_{23}\beta_{23}\,.$$

By comparing the coefficients one obtains $\gamma_{12} = \alpha_1$, $\gamma_{13} = -\alpha_2$ and $\gamma_{23} = \alpha_3$, determining the asymmetric 2-form γ completely.

Remark: From electrodynamics you know that in many cases it is convenient to express an oriented surface element ($\sim$ 2-form) by a vector ($\sim$ 1-form) that is perpendicular to the surface element and whose length corresponds to the surface area. The Hodge duality mediates between these two equivalent descriptions, generalized to arbitrary p-forms in n dimensions.

2.2.5 Hodge-star operator$\star$

To formalize the Hodge duality, we introduce a special operator $\star$, which maps a p-form to a $n-p$-form. This so-called *Hodge-star operator* is defined as the map $\star : \bigwedge^p(V^*) \to \bigwedge^{n-p}(V^*) : \alpha \mapsto \star\alpha$ with

$$\boxed{\star\alpha \;:=\; \iota_{\alpha^\sharp}\omega} \tag{2.51}$$

where ω is the volume form introduced above. The Hodge-$\star$ operator is a linear operation. Since both the musical isomorphism $\sharp$ and the volume form depend on the metric tensor, the Hodge-Star operator always refers by definition to a certain metric.

Proof: In the nomenclature of the previous section we have $\star\alpha = s\gamma$. Using the relation Eq. (2.41) one obtains

$$\mathbf{g}^*(\beta,\gamma)\omega \;=\; \frac{\mathbf{g}^*(\beta,\iota_{\alpha^\sharp}\omega)\omega}{s} \;=\; \frac{\mathbf{g}^*(\alpha\wedge\beta,\omega)\omega}{s} \;=\; \alpha\wedge\beta\frac{\mathbf{g}^*(\omega,\omega)}{s} \;=\; \alpha\wedge\beta\,,$$

that is, the definition is compatible with Eq. (2.46).

2.2.6 Representation of the Hodge-star operator$\star$

In a given basis $\{\mathbf{e}_i\}$ the p-form α is represented by

$$\alpha = \frac{1}{p!}\,\alpha_{i_1\ldots i_p}\,\mathbf{e}^{i_1}\wedge\ldots\wedge\mathbf{e}^{i_p}\,. \tag{2.52}$$

The corresponding dual vector

$$\alpha^\sharp = \frac{1}{p!}\,\alpha^{i_1\ldots i_p}\,\mathbf{e}_{i_1}\wedge\ldots\wedge\mathbf{e}_{i_p} \tag{2.53}$$

can be obtained by raising the indices of α. Therefore we have

$$\star\,\alpha \;=\; \iota_{\alpha^\sharp}\omega \;=\; \frac{1}{p!(n-p)!}\,\sqrt{|g|}\,\epsilon_{i_1\ldots i_n}\,\alpha^{i_1\ldots i_p}\,\mathbf{e}^{i_{p+1}}\wedge\ldots\wedge\mathbf{e}^{i_n}\,. \tag{2.54}$$

Thus the $n-p$-form $\star\alpha$is given in terms of the components

$$\boxed{[\star\alpha]_{i_{p+1}\ldots i_n} \;=\; \frac{1}{p!}\,\sqrt{|g|}\,\epsilon_{i_1\ldots i_n}\alpha^{i_1\ldots i_p}\,.} \tag{2.55}$$

For example, this allows us to compute the components of $\star\star\alpha$

$$[\star\star\alpha]_{i_{n-p+1}\ldots i_n} \;=\; \frac{|g|}{p!(n-p)!}\,\epsilon_{i_1\ldots i_n}\,g^{i_1 j_{p+1}}\cdots g^{i_{n-p} j_n}\,\epsilon_{j_1\ldots j_n}\,g^{j_1 k_1}\cdots g^{j_p k_p}\,\alpha_{k_1\ldots k_p} \tag{2.56}$$

which, after some algebra, implies that $\star\star\alpha = \pm\alpha$ (see exercise).

2.2.7 Properties of the Hodge-star operator $\star$

The Hodge dual of a scalar is the volume form

$$\star 1 \;=\; \omega\,. \tag{2.57}$$

In some books, therefore, the volume form is not denoted by ω but simply by $\star 1$. Conversely, the Hodge dual of the volume form reads

$$\star\omega \;=\; \iota_{\omega^\sharp}\omega \;=\; \frac{s}{n!}(\mathbf{e}^1\wedge\ldots\wedge\mathbf{e}^n)(\mathbf{e}_1\wedge\cdots\wedge\mathbf{e}_n) \;=\; s \;=\; \pm 1\,. \tag{2.58}$$

Evaluating Eq. (2.56) one can show that

$$\boxed{\star\star\alpha \;=\; s\,(-1)^{p(n-p)}\,\alpha} \qquad\Rightarrow\qquad \star^2 = \star\star = \pm 1\,, \tag{2.59}$$

that is, the $\star$-operator is (up to a minus sign) an involution.

As shown above, the Hodge duality is closely related to the generalized scalar product. For example, one can write Eq. (2.46) in the form

$$\alpha\wedge\star\beta \;=\; \mathbf{g}^*(\alpha,\beta)\omega \tag{2.60}$$

where α and β are two arbitrary p-forms. Conversely it is possible to express the generalized scalar product using $\wedge$ and $\star$ by

$$\mathbf{g}^*(\alpha,\beta) \;=\; s\star(\alpha\wedge\star\beta) \tag{2.61}$$

2.2.8 Hodge-$\star$ operator represented in an orthonormal basis

As pointed out earlier, the Hodge duality always refers to a given metric $\mathbf{g}$, which is the reason why in the literature we sometimes find the notation $\star_g$. For this reason, the rules for calculating the $\star$ product become particularly simple when working in an orthonormal basis. In this case, the Hodge dual of a p-form consists simply of the complementary basis forms. For example,

$$\star\,(\mathbf{e}^1\wedge\ldots\wedge\mathbf{e}^p) \;=\; \mathbf{e}^{p+1}\wedge\ldots\wedge\mathbf{e}^n\,. \qquad \text{(if } \{\mathbf{e}^i\} \text{ orthonormal)} \tag{2.62}$$

If the left side does not consist of an ordered sequence of the first p basis vectors, but of an arbitrary sequence of basis vectors, we have instead

$$\star\,(\mathbf{e}^{i_1}\wedge\ldots\wedge\mathbf{e}^{i_p}) \;=\; \pm\mathbf{e}^{i_{p+1}}\wedge\ldots\wedge\mathbf{e}^{i_n}\,, \qquad \text{(if } \{\mathbf{e}^i\} \text{ orthonormal)} \tag{2.63}$$

where the sign depends on whether $(i_1, \ldots, i_n)$ is an even or odd permutation of $1, \ldots, n$.

Example:

- In the $\mathbb{R}^2$ equipped with the ordinary scalar product $\mathbf{g} = \mathrm{diag}(1,1)$ and the orthonormal standard basis $\{\mathbf{e}^i\}$ we have
$$\star 1 = \mathbf{e}^1 \wedge \mathbf{e}^2\,; \qquad \star \mathbf{e}^1 = \mathbf{e}^2\,; \qquad \star \mathbf{e}^2 = -\mathbf{e}^1\,; \qquad \star(\mathbf{e}^1 \wedge \mathbf{e}^2) = 1$$
- In the $\mathbb{R}^3$ with the ordinary scalar product $\mathbf{g} = \mathrm{diag}(1,1,1)$ and the standard basis we have
$$\star 1 = \mathbf{e}^1 \wedge \mathbf{e}^2 \wedge \mathbf{e}^3\,; \qquad \star \mathbf{e}^1 = \mathbf{e}^2 \wedge \mathbf{e}^3\,; \qquad \star \mathbf{e}^2 = -\mathbf{e}^1 \wedge \mathbf{e}^3\,; \qquad \star \mathbf{e}^3 = \mathbf{e}^1 \wedge \mathbf{e}^2$$
$$\star(\mathbf{e}^1 \wedge \mathbf{e}^2) = \mathbf{e}^3\,; \qquad \star(\mathbf{e}^1 \wedge \mathbf{e}^3) = -\mathbf{e}^2\,; \qquad \star(\mathbf{e}^2 \wedge \mathbf{e}^3) = \mathbf{e}^1\,; \qquad \star(\mathbf{e}^1 \wedge \mathbf{e}^2 \wedge \mathbf{e}^3) = 1$$

2.2.9 Self-duality *

A form α is called *self-dual* with respect to the Hodge star operation if

$$\star\,\alpha \propto \alpha\,, \tag{2.64}$$

i.e., if the form α is some kind of "eigenvector" of $\star$. Self-dual forms have a high degree of symmetry. Although here we want to introduce self-duality only for the sake of completeness, it should be noted that there is a self-dual formulation of general relativity, which is increasingly influential in current research, especially in the context of quantum loop gravity.

Since the Hodge operator maps p-forms to $n-p$-forms, self-dual forms must inevitably have the degree $p = n/2$, where n is the dimension of the underlying vector space. Therefore, self-dual forms can only exist vector spaces with even dimension. This reduces Eq. (2.59) to:

$$\star\star\alpha = s\alpha\,, \tag{2.65}$$

where $s = \mathrm{sgn}(\det\{g_{ij}\})$. Thus one distinguishes the following two cases:

	Riemann ($s=1$)	Lorentz ($s=-1$)
self-dual	$\star\alpha = \alpha$	$\star\alpha = i\alpha$
anti-self-dual	$\star\alpha = -\alpha$	$\star\alpha = -i\alpha$

The lowest dimension in which self-duality might occur is $n = 2$, but it is easy to show (exercise) that self-dual 1-forms do not exist in $\mathbb{R}^2$, since the corresponding equations do not have such solutions. Thus vector spaces with self-dual forms have to be at least four-dimensional. As an example, consider the Euclidean $\mathbb{R}^4$. In this case the hodge-dual of a 2-form α is given in terms of the components

$$[\star\alpha]_{ij} = \frac{1}{2}\epsilon_{ijkl}\alpha^{kl} = \frac{1}{2}\epsilon_{ijkl}\eta^{km}\eta^{ln}\alpha_{nm}\,. \tag{2.66}$$

Regarding the components of the antisymmetric 2-form α as a 6-component vector, the $\star$-operation can be interpreted as a linear representation in the form of a 6×6 matrix.

The eigenvalues are $(-1,-1,-1,1,1,1)$. So in $\mathbb{R}^4$ there are three self-dual and three anti-self-dual forms.

Calculation: Check this with *Mathematica*®

```
map={{1,2},{1,3},{1,4},{2,3},{2,4},{3,4}}
M=Table[Signature[Join[map[[i]],map[[j]]]],{i,1,6},{j,1,6}]
M//Eigenvalues
```

The six 2-forms read

$$e^1 \wedge e^2 \pm e^3 \wedge e^4\,, \qquad e^1 \wedge e^3 \pm e^2 \wedge e^4\,, \qquad e^1 \wedge e^4 \pm e^2 \wedge e^3\,. \tag{2.67}$$

Since the second exterior power $\bigwedge^2(\mathbb{R}^{4^*})$, i.e., the space of all 2-forms acting on $\mathbb{R}^4$, is also six-dimensional, you can see immediately that these six 2-forms span the entire space of all 2-forms. Thus, it is possible to express each 2-form as the sum of a self-dual and an anti-self-dual form, in the same way as any matrix can be expressed as the sum of a symmetric and an antisymmetric matrix. There are accordingly two projectors

$$P^{\pm} \;=\; \frac{1}{2}(\mathbb{1} \pm \star) \tag{2.68}$$

with $P^+ + P^- = \mathbb{1}$ represented in terms of components by

$$P^{\pm}{}_{ij}{}^{kl} \;=\; \frac{1}{2}\left(I_{ij}{}^{kl} \pm \epsilon_{ij}{}^{kl}\right), \tag{2.69}$$

where

$$I_{ij}{}^{kl} = \frac{1}{2}(\delta_i^k\delta_j^l - \delta_j^k\delta_i^l) \tag{2.70}$$

is the antisymmetric identical map in the space $\bigwedge^2(\mathbb{R}^{4^*})$.

2.3 Functions, coordinate systems and differential forms

So far we have dealt with the structure of vector spaces and the linear forms defined on them. In non-relativistic physics as well as in special relativity, space and time can be considered as vector spaces *per se* , i.e., the points of space-time are in one-to-one correspondence with certain position vectors. However, in the theory of general relativity, in which the spacetime becomes a curved dynamic object, this concept fails. In general relativity the spacetime is represented by a so-called *manifold* - a set of points which locally still looks like a vector space but which does not have a vector space structure globally, meaning that the points on the manifold can no longer be described by vectors. However, on short distances the space-time still looks like a planar vector space, just as the surface of the ocean locally looks like a $\mathbb{R}^2$. Vectors can therefore be used to specify *directions* in space-time.

In this chapter, we will not yet be dealing with curved manifolds, but for the time being only with flat spaces, more specifically with open connected subsets $U \subset \mathbb{R}^n$. But to facilitate the generalization to curved manifolds, let us attempts to make no use of the vector space structure in U. Concretely, this means that we do not want to consider

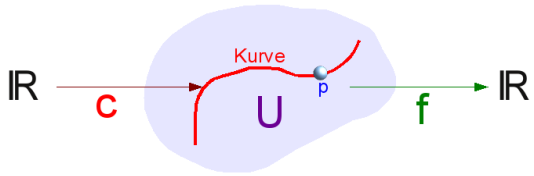


Figure 2.4: A smooth curve $c : \mathbb{R} \to U$ running through the point p can be used to define a directional derivative $\nabla_c f_p$ of the function $f : U \to \mathbb{R}$ at the point $p$along this curve (see text).

the points $p \in U$ as vectors, in particular we do not want to add, subtract, or multiply them by a scalar like vectors.

2.3.1 Scalar functions, curves and directional derivatives

Scalar functions:
A scalar function is a map $f : U \to \mathbb{R}$ which associates with every point $p \in U$ a certain number f_p. In the following we will always assume f to be continuously differentiable in a sense to be made more precise below. Note that the function f may be nonlinear.

Ordinary directional derivative:
Since U is generally multidimensional, the way in which a function changes depends on the chosen direction. Normally, in $\mathbb{R}^n$, we would define the *directional derivative* of the function f at the point $\mathbf{u} \in U$ in the direction $\mathbf{v}$ by the limit

$$\nabla_{\mathbf{v}} f(\mathbf{u}) \;=\; \lim_{\mu \to 0} \frac{f(\mathbf{u} + \mu \mathbf{v}) - f(\mathbf{u})}{\mu} \,. \tag{2.71}$$

However, as one can see, this definition makes explicit use of the vector space structure of U by adding to the point $\mathbf{u}$ the displacement $\mu\mathbf{v}$. However, according to the rules of the game formulated above, we do not want to make use of the vector space structure. We are therefore looking for an alternative way to define a directional derivative.

Curves:
As a possible way out let us consider *parameterized curves*, i.e., functions $c : \mathbb{R} \to U$, which map a parameter $\lambda \in (a, b) \subset \mathbb{R}$ onto a point $c(\lambda) \in U$. These maps are assumed to be continuous and continuously differentiable, which basically means that the curve consists of a single smooth piece.

To define a directional derivative at the point $p \in U$, we now consider a curve c that runs through p (see Fig. 2.4), with the parameterization chosen in that $c(0) = p$. The concatenated map $f \circ c$ is then a map $\mathbb{R} \to \mathbb{R}$, which can be differentiated in the usual way without the use of vectors. So we use a curve instead of a vector to specify a direction and to define the corresponding directional derivative:

$$\nabla_c f_p \;:=\; \frac{\mathrm{d}}{\mathrm{d}\lambda} f(c(\lambda)) \bigg|_{\lambda=0} \tag{2.72}$$

Directional derivative:
The directional derivative ∇_c defined above seems to depend only on the direction and velocity with respect to its parameter at which the curve c traverses the point p, but

not the on the shape and the velocity of the curve outside of p. Thus two curves are declared to be *equivalent* in the point p, if they give the same directional derivation for all possible functions:

$$c \underset{p}{\sim} c' \qquad \Leftrightarrow \qquad \nabla_c f_p = \nabla_{c'} f_p \quad \forall f\,. \tag{2.73}$$

A directional derivative in p is thus an *equivalence class* of curves $[c]_p$, which we want to denoted by capital letters X_p in anticipation of the usual notations in differential geometry:

$$X_p := [c]_p\,. \tag{2.74}$$

If we apply this directional derivative to a function, we just get the corresponding directional derivative

$$X_p f \;=\; \nabla_c f_p\big|_{c\in X_p}\,, \tag{2.75}$$

where the curve c is just an arbitrary representative of the equivalence class X_p.

Tangent space:
It can be shown that the directional derivatives X_p form a linear vector space, referred to as the *tangent space* T_pU of U in the point p.

For the case $U \subset \mathbb{R}^n$ considered here, this is obvious because every curve crossing p is associated with a velocity vector $\mathbf{v} = \frac{\mathrm{d}}{\mathrm{d}\lambda}c(\lambda)|_{\lambda=0}$, and therefore the addition or scalar multiplication of directional derivatives thus corresponds to the addition and scalar multiplication of such velocity vectors. As we shall see below, the concept of the tangent space will also work for curved manifolds, where the tangent space is also a linear vector space.

First, it is important that we get familiar with a completely new perspective, which is common in differential geometry:

The vectors of the tangent space T_pU are directional derivatives.

Note: Vectors are derivatives - this is not easy to accept for newcomers. However, we must get used to the fact that vectors no longer specify *distances* between points. A vector represents merely a certain direction combined with a number (magnitude of the vector). The only thing that can be done with such a vector is to build a directional derivative, that is to ask, how much e.g. a coordinate or a function varies if we go along the manifold in a given direction. This allows us to identify tangent vectors with the corresponding directional derivative.

Vector fields:
A directional derivative $X_p \in T_pU$ is defined at a certain point p. In general, however, the directional derivative is declared not only in a single point, but as a bundle of many directional derivatives on the whole space U. Such a bundle is said to be a *vector field* denoted X. Applying this vector field to a function f, what we get is a new function Xf with $Xf|_p = X_p f$.

According to the new view, a vector field X acts on functions by carrying out the corresponding directional derivative at each point. The vector field X is therefore a linear operator

$$X(\lambda f + \mu g) = \lambda X f + \mu X g \qquad (f, g \text{ functions};\ \lambda, \mu \in \mathbb{R}) \tag{2.76}$$

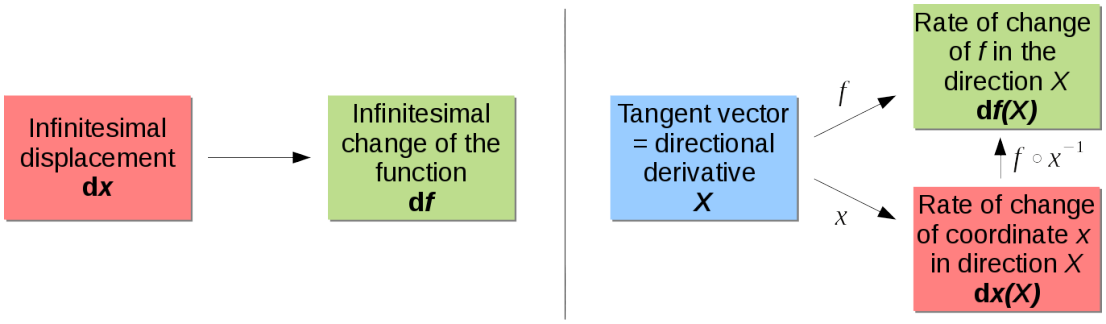


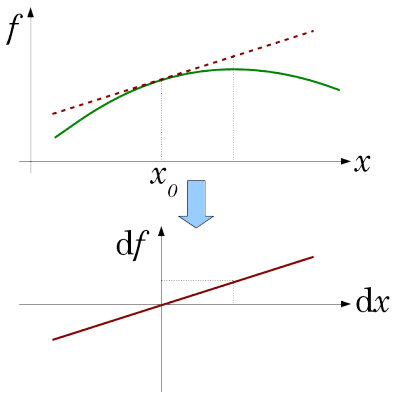
Figure 2.5: The traditional interpretation of differentials as tiny displacements and and tiny changes compared with the interpretation of differentials as 1-forms in differential geometry.

which obeys the *Leibniz Rule* (product rule) on products of functions:

$$\boxed{X(fg) \;=\; fXg + gXf\,.} \tag{2.77}$$

2.3.2 Differentials

The derivative of a scalar function f at the point x_0 is known to be a linear approximation of the function around x_0: If you move a little bit in the direction $\mathrm{d}x$, the function changes in a linear approximation by the amount $\mathrm{d}f = f'(x_0)\,\mathrm{d}x$, where $\mathrm{d}f$ and $\mathrm{d}x$ *infinitesimal differentials*. In higher-dimensional spaces we get $\mathrm{d}f = \nabla f \cdot \mathrm{d}\mathbf{x}$, where ∇f is the gradient of f.



Differentials interpreted as linear approximations of a function in a given point.

Infinitely small vectors - we will now have to say goodbye to this strange but over the years cherished idea of infinitesimally small pieces. It is replaced by an interpretation compatible with the abstract notion of directional vectors $X \in T_pU$ introduced above.

The different way of thinking is sketched schematically in Fig. 2.5:

- The left side shows the traditional way of thinking: The starting point here is a function $f(x)$, i.e., a direct mapping of a coordinate x to the value of the function. One then asks how f changes when x changes. For large changes, this relationship may be nonlinear, but for infinitesimally small changes $\mathrm{d}x$, the infinitesimally small change $\mathrm{d}f$ will depend linearly on $\mathrm{d}x$.
- The right side shows the new perspective of differential geometry: The starting point here is the abstract 'physical' space U. Note that *two* functions are defined in this space: A function $f : U \to \mathbb{R}$ as well as a coordinate function $x : U \to \mathbb{R}$. Thus each point p is assigned a function value f_p and a coordinate x_p. In order to investigate how these functions change locally, we have to form the directional

derivative of both.

Now let the directional derivative X_p act on these two functions and let us define

$$\mathrm{d}f(X_p) \;:=\; X_p f \qquad\qquad \mathrm{d}x(X_p) \;:=\; X_p x \tag{2.78}$$

These two quantities describe how the function f as well as the coordinate x change in a linear approximation when moving away from the point p in the direction X_p. As can be seen, the differentials $\mathrm{d}f$ and $\mathrm{d}x$ are no longer infinitesimal quantities here, rather they are linear functions on tangent vectors, i.e., linear mappings $T_pU \to \mathbb{R}$. They are therefore elements of the vector space dual to T_pU, the so-called *cotangential space* or *cotangent space* T_p^*U:

Differentials are 1-forms and therefore they are elements of the cotangent space T_p^*U.

As 1-forms, *differentials* differentials map abstract tangent vectors $X_p \in T_pU$ linearly to numbers. The differential $\mathrm{d}f$ of a function f is defined pointwise by $\mathrm{d}f_p(X_p) = X_p f$. If we intend to perform this operation in each point simultaneously, we briefly write

$$\boxed{\mathrm{d}f(X) \;=\; Xf} \tag{2.79}$$

where X is a vector field and $\mathrm{d}f$ is a field of 1-forms.

Note: From the new perspective, the good old textbook mathematics is a crude simplification, which omits the painful detour via U and T_pU and instead considers only the concatenated mapping $f \circ x^{-1} : \mathbb{R} \to \mathbb{R} : x \mapsto f(x) := f(p(x))$ (the vertical arrow in the figure). However, without introducing a tangent space there is in principle no way to specify a direction. This problem is circumvented by representing directions through the components of tiny vectors, interpreted as infinitesimal differentials $\mathrm{d}x$. This is the price to pay if we want to avoid the definition of a tangent space.

Remember: Summary of notations:

f	Function $U- > \mathbb{R}$
f_p	Value of the function at the point p
X_p	Tangent vector in the point p
$X_p f$	Derivative of f in the direction X at point p
X	Tangent vector field
Xf	Derivative of f along the tangent vector field
$\mathrm{d}f_p$	Differential form of f in the point p, mapping tangent vectors onto numbers
$\mathrm{d}f_p(X_p)$	Change of f in linear approximation in the direction X
$\mathrm{d}f$	Field of differential forms of the function f on U
$\mathrm{d}x_p$	Differential forms of x in the point p, mapping tangent vectors onto numbers
$\mathrm{d}x(X)$	Change of the coordinate x in linear approximation moving into the direction of X
$\mathrm{d}x_p(X_p)$	Change of the coordinate x in linear approximation in the direction X at point p

2.3.3 Coordinate systems

An *coordinate system* S is a set of n continuously differentiable functions $x^i : U \to \mathbb{R}$, which uniquely identifies each point $p \in U$ with n *coordinates* $x^i(p) \in \mathbb{R}$, where n is, as usual, the dimension of the space. If we arrange the coordinates $x^i(p)$ as the components of a vector $\mathbf{x}(p) \in \mathbb{R}^n$, we can also interpret a coordinate system as a

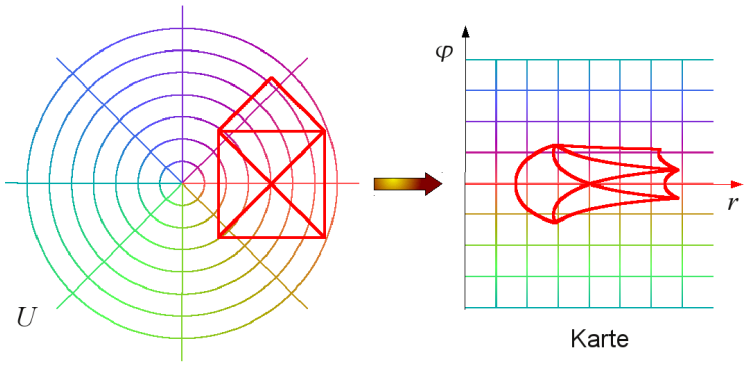


Figure 2.6: A coordinate system S bijectively maps the 'reality' (left) onto a planar map (right). To this end, every point $p \in U$ is assigned a point $\mathbf{x} \in \mathbb{R}^2$ on the map. This example shows the well-known polar coordinates and illustrates how a 'house of Santa Claus' is represented on the map. Each point $p \in U$ is uniquely mapped to $x^1 = r$ and $x^2 = \varphi$. The grit lines are obtained by varying one coordinate and keeping the others constant.

continuously differentiable bijection $S : U \to \mathbb{R}^n$, so to speak, as an invertible mapping from the physical space to a planar map. The inverted map S^{-1} is denoted by $p(\mathbf{x}) = p(x^1, \ldots, x^n)$.

As an example let us consider the well-known *polar coordinates*, which is shown in Fig. 2.6. In this coordinate system, every point $p \in U$ is assigned a radius $x^1 = r$ and an angle $x^2 = \varphi$. On the corresponding map, the 'reality ', e.g. the 'house of Santa Claus' shown on the left, appears to be heavily distorted. In particular, the edges of the house appear to be curved.

A coordinate system is represented by a grid of lines as shown in Fig. 2.6. These lines are obtained by varying one coordinate while keeping the others constant. If $p \in U$ is a point and $\mathbf{x}(p)$ is the corresponding point on the map, then these lines are given by

$$c_j(\lambda) = p(x^1, \ldots, x^{j-1},\ x^j + \lambda\ , x^{j+1}, \ldots, x^n)\,. \tag{2.80}$$

On the 'physical' space U these mesh lines may be curved, as the example of the polar coordinates demonstrates, while on the map they appear as parallel lines.

2.3.4 Coordinate basis

Each of these mesh lines c_j running through the point p is a certain curve defined on U, representing a directional derivative X_j at the point p. Since the coordinate mapping is bijective, the directional derivatives $X_1, \ldots, X_n$ are linearly independent and thus form a (not necessarily orthonormal) basis of the tangent space T_pU. A coordinate system thus provides a natural basis in which vectors (= directional derivatives) and 1-forms (= differentials) can be represented. This basis is called *coordinate basis*.

> **Note:** A coordinate basis is non necessarily orthogonal or normalized. Note that each point p has its own tangent space T_pU and therefore its own basis. How to get from one tagential space to an adjacent one will be one of the essential cornerstones of differential geometry.

We could now proceed and define a coordinate basis of T_pU denoted as $\mathbf{e}_1, \ldots, \mathbf{e}_n$, and

then construct as usual a corresponding dual basis $\mathbf{e}^1, \ldots, \mathbf{e}^n$ such that $\mathbf{e}^i(\mathbf{e}_j) = \delta^i_j$. In differential geometry, however, another notation prevails, which is intended to be a formal reminder of the calculation rules used in vector analysis, but for beginners it takes initially some time to get used to it.

Basis vectors of the tangent space:
Starting point is the observation that the basis vectors $\mathbf{e}_j$ have to be interpreted as a directional derivatives applied to an arbitrary function f which, when re-presented in coordinates, is just the usual partial derivative:

$$\mathbf{e}_j f \;=\; \frac{\partial}{\partial x^j} f\big(p(x_1, \ldots, x_n)\big)\,. \tag{2.81}$$

Therefore one introduces the formal notation

$$\boxed{\mathbf{e}_j \;=\; \partial_j \;=\; \frac{\partial}{\partial x^j}}\,. \tag{2.82}$$

This gives the symbol $\frac{\partial}{\partial x^j}$ a kind of double life: Strictly speaking, it is a base vector of the tangent space along the j-th coordinate line, but on the corresponding map the symbol $\frac{\partial}{\partial x^j}$ acts like the traditional partial derivative.

Basis-1-forms of the cotangent space:
Similarly one proceeds at the cotangential space. Its basis vectors $\mathbf{e}^1, \ldots, \mathbf{e}^n$ turn out to be the differentials $\mathbf{e}^i = \mathrm{d}x^i$ of coordinate functions x^i. To see this we have to check that they satisfy the definition property $\mathbf{e}^i(\mathbf{e}_j) = \delta^i_j$, and in fact, we have $\mathrm{d}x^i(\mathbf{e}_j) = \partial_j x^i = \delta^i_j$. Therefore, in differential geometry one uses the notations

$$\boxed{\mathbf{e}^i \;=\; \mathrm{d}x^i\,.} \tag{2.83}$$

This definition is formally compatible with the algebraic rule $\mathrm{d}x^i(\frac{\partial}{\partial x^j}) = \frac{\partial}{\partial x^j} x^i = \delta^i_j$.

Coordinate basis of T_pU : Partial derivatives $\mathbf{e}_j = \partial_j = \frac{\partial}{\partial x^j}$

Dual coordinate basis of T^*_pU: Differentials $\mathbf{e}^i \;=\; \mathrm{d}x^i$

2.3.5 Representation of fields in coordinate systems

Vectors and tensors can be represented in the usual way in the coordinate basis, whereby instead of $\mathbf{e}_i$ and $\mathbf{e}^j$ we simply write $\partial/\partial x^i$ and $\mathrm{d}x^j$. The only new thing is that they are vector fields or 1-form fields on U without explicitly referring to the vector space structure of U. If one refers to a certain point $p \in U$ and thus to a certain tangent space T_pU or cotangential space T^*_pU, this is expressed by a subscript p (not to be confused with the rank a p-form). If, on the other hand, one wants to look at all tangent spaces simultaneously, one simply omits the index p. For example, f denotes a function, while f_p denotes the value of this function at the point p. Similarly, TU denotes the totality of all tangent spaces while T_pU stands for the individual tangent space attached to p.

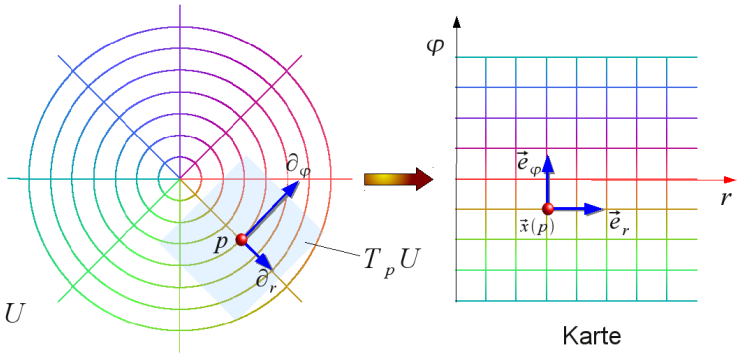


Figure 2.7: Coordinate basis: In every point $p \in U$ one can imagine a tangent space $T_pU \cong \mathbb{R}^2$ as being attached to this point. The coordinate system gives rise to a special basis in this space in the direction of the coordinate lines running through p.

Vector fields and fields of differential forms:
In a coordinate basis the vector field $\mathbf{X} : U \to TU$ is represented by

$$\mathbf{X} = X^i \frac{\partial}{\partial x^i} = X^i \partial_i \,. \tag{2.84}$$

Similarly, a field of q-multivectors $\mathbf{T} : U \to \wedge^q TU$ is represented by

$$\mathbf{T} = \frac{1}{q!}\, T^{i_1 \ldots i_q}\, \partial_{i_1} \wedge \ldots \wedge \partial_{i_q}. \tag{2.85}$$

Analogously the field of p-forms $\alpha : U \to \wedge T_p^* U$ can be represented in a coordinate system by

$$\alpha = \frac{1}{p!}\, \alpha_{i_1 \ldots i_p}\, \mathrm{d}x^{i_1} \wedge \ldots \wedge \mathrm{d}x^{i_p} \,. \tag{2.86}$$

Differentials of functions
The differential of a scalar function f is a 1-form and therefore it can be represented by $\mathrm{d}f = \alpha_i \, \mathrm{d}x^i$ with certain coefficients α_i. To determine these coefficients, let $\mathrm{d}f$ act on a basis vector ∂_j. On the one hand we get the usual partial derivatives $\mathrm{d}f(\partial_j) = \partial_j f$ because of Eq. (2.79), on the other hand $\mathrm{d}f(\partial_j) = \alpha_i \, \mathrm{d}\underbrace{x^i(\partial_j)}_{=\delta^i_j}$. Thus, $\alpha_j = \partial_j f$, i.e.,

$$\boxed{\mathrm{d}f = (\partial_j f)\, \mathrm{d}x^j \,.} \tag{2.87}$$

In the coordinate representation we get the well-known *total differential*. This formal correspondence is one reason why we no longer denote the base vectors by $\mathbf{e}_i$ and $\mathbf{e}^j$, but instead use the notations ∂_i and $\mathrm{d}x^j$.

old notation	new notation
$\mathbf{e}_i$	$\partial_i := \partial/\partial x^i$
$\mathbf{e}^j$	$\mathrm{d}x^j$
$\mathbf{X} = X^i \mathbf{e}_i$	$\mathbf{X} = X^i \partial_i$
$\mathbf{A} = \frac{1}{q!} A^{i_1 \ldots i_q} \mathbf{e}_{i_1} \wedge \ldots \wedge \mathbf{e}_{i_q}$	$\mathbf{A} = \frac{1}{q!} A^{i_1 \ldots i_q} \partial_{i_1} \wedge \ldots \wedge \partial_{i_q}$
$\alpha = \frac{1}{p!} \alpha_{j_1 \ldots j_p} \mathbf{e}^{j_1} \wedge \ldots \wedge \mathbf{e}^{j_p}$	$\alpha = \frac{1}{p!} \alpha_{j_1 \ldots j_p} \mathrm{d}x^{j_1} \wedge \ldots \wedge \mathrm{d}x^{j_p}$

Table 2.1: Comparison of the previous notation with the new notation used in differential geometry

2.3.6 Changing between different coordinate systems

For the representation of a space U there are many possible coordinate systems. For example, in the theory of relativity, each frame of reference is a valid coordinate system. Therefore, one often faces the task of transforming representations in one coordinate system into representations of another coordinate system.

Let us consider two given coordinate systems $S : p \to x^i(p)$ und $S' : p \to x^{i'}(p)$. Since the two coordinate mappings by objective, there exist concatenated maps

$$\begin{aligned} S' \circ S^{-1} : &\quad x^i \to x^{i'}(x^1, \ldots, x^n) \\ S \circ S'^{-1} : &\quad x^{i'} \to x^i(x^{1'}, \ldots, x^{n'}) \,. \end{aligned} \tag{2.88}$$

For a scalar function f defined on U we have

$$\frac{\partial f}{\partial x^i} = \frac{\partial x^{j'}}{\partial x^i} \frac{\partial f}{\partial x^{j'}} \tag{2.89}$$

or, in short

$$\partial_i = (\partial_i x^{j'}) \partial'_j \qquad \text{and} \qquad \partial'_i = (\partial'_i x^j) \partial_j \tag{2.90}$$

where we used the notation $\partial_i = \frac{\partial}{\partial x^i}$ and $\partial'_j = \frac{\partial}{\partial x^{j'}}$. Because the partial derivatives play the role of basis vectors in tangent space, this transformation law expresses nothing more than a usual basis transformation. A comparison with the old notation $\mathbf{e}'_i = \mathbf{e}_k \tilde{M}^k{}_i$ in Eq. (1.17) yields

$$M^i{}_j = \frac{\partial x^{i'}}{\partial x^j} \quad \text{and} \quad \tilde{M}^i{}_j = \frac{\partial x^i}{\partial x^{j'}}. \tag{2.91}$$

The transformation matrix is nothing but the Jacobian of the concatenated map. Consequently, the components of a tangent vector $\mathbf{v} = v^i \partial_i$ transform according to

$$v^i \to v^{i'} = \frac{\partial x^{i'}}{\partial x^j} v^j = (\partial_j x^{i'}) v^j \,. \tag{2.92}$$

Remark: Please note that only components of tangent vectors v^i transform in this way, but not the coordinates x^i themselves, i.e., $x^{i'} \neq \frac{\partial x^{i'}}{\partial x^j} x^j$. The reason is that the coordinates x^i are

	T_pU	T_p^*U
object	$\mathbf{X} = X^i\partial_i$	$\alpha = \alpha_i\,\mathrm{d}x^i$
basis	$\partial'_i = (\partial'_i x^j)\partial_j$	$\mathrm{d}x^{i'} = (\partial_j x^{i'})\,\mathrm{d}x^j$
components	$X^{i'} = (\partial_j x^{i'})X^j$	$\alpha'_i = (\partial'_i x^j)\alpha_j$

Table 2.2: Summary of the transformation laws for vectors $\mathbf{X}$ and 1-forms αbetween different coordinate systems. Tensors of higher ranking transform analogously in each of their indices.

functions on the manifold and not components of a tangent vector field.

Similar relationships can be found for the cotangent space. A comparison with Eq. (1.54) on page 19 provides the transformation law for the basis differentials

$$\mathrm{d}x^i = \frac{\partial x^i}{\partial x^{j'}}\,\mathrm{d}x^{j'} \qquad \text{bzw.} \qquad \mathrm{d}x^{i'} = \frac{\partial x^{i'}}{\partial x^j}\,\mathrm{d}x^j \tag{2.93}$$

Remark: You already encountered this transformation law as a chain rule for infinitesimal differentials. The new notation introduced above has indeed been chosen to reproduce this familiar appearance. It should be remembered, however, that the differentials $\mathrm{d}x^i$ are linearly independent 1-forms of the cotangent space T_p^*U and not just components of infinitesimally tiny vectors.

Correspondingly, the components of a 1-form $\alpha = \alpha_i\,\mathrm{d}x^i$ transform as

$$\alpha_i \to \alpha'_i = \frac{\partial x^j}{\partial x^{i'}}\alpha_j = (\partial'_i x^j)\alpha_j \tag{2.94}$$

Example:

Cartesian coordinates and polar coordinates:
Consider the space $U \subset \mathbb{R}^2$ equipped with the usual scalar product. The simplest representation is the usual Cartesian coordinate system $S : p \to (x^1, x^2) = (x, y)$. But in many cases one also uses polar coordinates $S' : p \to (x^{1'}, x^{2'}) = (r, \phi)$. The transformation equations between the two coordinate systems read

$$S \circ S'^{-1} : (r, \phi) \to (x, y) = (r\cos\phi, r\sin\phi)$$

or in short $x = r\cos\phi$ und $y = r\sin\phi$. In the Cartesian basis with the basis vectors $\partial_x, \partial_y$the field of the metric tensor is constant on $U$and is given by

$$(g_{ij}) = \begin{pmatrix} \mathbf{g}(\partial_x, \partial_x) & \mathbf{g}(\partial_x, \partial_y) \\ \mathbf{g}(\partial_y, \partial_x) & \mathbf{g}(\partial_y, \partial_y) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The basis vectors in polar coordinates read

$$\begin{aligned} \partial_r &= (\partial_r x)\partial_x + (\partial_r y)\partial_y &&= \cos\phi\,\partial_x + \sin\phi\,\partial_y \\ \partial_\phi &= (\partial_\phi x)\partial_x + (\partial_\phi y)\partial_y &&= -r\sin\phi\,\partial_x + r\cos\phi\,\partial_y \end{aligned}$$

Therefore, the field of the metric tensor represented in polar coordinates is now given by

$$(g'_{ij}) = \begin{pmatrix} \mathbf{g}(\partial_r, \partial_r) & \mathbf{g}(\partial_r, \partial_\phi) \\ \mathbf{g}(\partial_\phi, \partial_r) & \mathbf{g}(\partial_\phi, \partial_\phi) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix},$$

i.e., it depends also on the location (more specifically, on the radius). In both cases, the metric tensors are diagonal, i.e., they describe orthogonal basis systems. Polar coordinates are shown graphically in Fig. 2.7 on page 64.

2.3.7 Degenerate differential forms and zero vector fields

A p-form α is called *degenerate*, if there exists a non-vanishing vector $\mathbf{X} \neq 0$ such that

$$\alpha(\mathbf{X}, \mathbf{Y}_{(1)}, \ldots, \mathbf{Y}_{(p-1)}) = 0 \qquad \forall, \mathbf{Y}_{(1)}, \ldots, \mathbf{Y}_{(p-1)}, \tag{2.95}$$

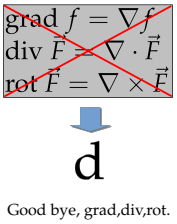
that is, if there is a vector that forces the form to render zero, regardless of what is going on at the other input slots. Such a vector $\mathbf{X}$ is also called *zero vector* (not to be confused with the neutral element of the vector space). If α is a field of p-forms, then this field is called *degenerate* if there is an entire vector field $\mathbf{X}$ that has the above property at every point $p \in U$. Such a vector field is called *zero vector field*.

As an example, let us consider a 2-form α with the representation $\alpha(\mathbf{X}, \mathbf{Y}) = \frac{1}{2}\alpha_{ij}X^iY^i$. Apparently, a 2-form is degenerate if and only if the 'matrix' α_{ij} has an eigenvector of eigenvalue zero. Since the matrix is real and antisymmetric, it has a purely imaginary spectrum of conjugate complex eigenvalues. It follows immediately that a 2-form in a vector space with an odd dimension always has a zero vector field.

Sketch of a proof: A real matrix has either individual real or pairs of conjugate complex eigenvalues, as one can easily show by complex conjugation of the eigenvalue equation. A real antisymmetric matrix multiplied by i is a Hermitian matrix that we know from quantum mechanics to have only real eigenvalues, so it is clear that a real antisymmetric matrix has a purely imaginary spectrum. The eigenvalue spectrum of a real antisymmetric matrix must therefore consist of either zeroes or $\pm$-pairs of purely imaginary eigenvalues. It follows that in odd dimensions, at least one of the eigenvalues is equal to zero.

2.4 Derivatives

In the theory of differential forms, the good old differential operators such as gradient, divergence and rotation, are replaced by the *exterior derivative*. By introducing this concept, the process of taking the derivative is founded on a more general basis and incorporated naturally into the formalism of exterior algebra.



Good bye, grad,div,rot.

2.4.1 Generalized differential

How the differential operates:

In page 60 we have already seen how to compute the derivative of (scalar) functions $f : U \to \mathbb{R}$. The result is a field of 1-forms, the so-called *differential* $\mathrm{d}f$. The differential $\mathrm{d}f_p$

at the point $p \in U$ is an element of the cotangent space $T_p^* U$, which can be thought of as a flat vector space attached to the point p. Applying the 1-form $\mathrm{d}f_p$ to a vector pointing into a certain direction yields as a result the change of f in a linear approximation along this direction (see Eq. (2.87)):

$$\mathrm{d}f(\partial_j) = \partial_j f \qquad \Rightarrow \qquad \mathrm{d}f = (\partial_j f)\,\mathrm{d}x^j.$$

In the following we would like to introduce a more general derivative which can act not only on scalar functions but also on fields of any p-forms α. The result is a *generalized differential* $\tilde{\mathrm{d}}\alpha$ (the tilde is a preliminary notation, which will be dismissed again in the following section). By analogy with $\mathrm{d}f$, this differential is intended to provide information about how the field of the p-form α changes along a given direction. Since α is a linear machine that maps p vectors to a number, $\tilde{\mathrm{d}}\alpha$ must be a machine that maps p-vectors as well as another vector pointing into certain direction (that is, a total of $p+1$ vectors) to a number. Because differentiating is a linear operation, we expect $\tilde{\mathrm{d}}\alpha$ to be a covariant tensor of rank $p+1$.

Remember: If we compute the derivative of a tensor, its rank increases by 1.

Embedding into the exterior algebra:
If we introduce the derivative in the sense described above and let it act on a 1-form α, then, for example, the second-order tensor $\tilde{\mathrm{d}}\alpha$ represented in a given basis would be given by

$$(\tilde{\mathrm{d}}\alpha)_{ij} \;=\; \partial_i \alpha_j\,.$$

As can be readily seen from this example, this tensor is not necessarily antisymmetric, i.e., it would leave the scope the exterior algebra, which is constrained to antisymmetric tensors. So, at first glance, the concept of a derivative seems to be incompatible with exterior algebra.

On closer inspection, however, this problem can be tackled. To start with, we note that each tensor can be written as the sum of a symmetric and an antisymmetric component, e.g.

$$(\tilde{\mathrm{d}}\alpha)_{ij} \;=\; \frac{1}{2}(\partial_i \alpha_j + \partial_j \alpha_i) + \frac{1}{2}(\partial_i \alpha_j - \partial_j \alpha_i)\,.$$

If you want to do physics exclusively on the basis of antisymmetric tensors, you will eventually have to generate measurable quantities, i.e., scalars, by contraction. If, however, the above tensor is contracted with another antisymmetric tensor, the symmetric component drops out, as a symmetric tensor contracted with an antisymmetric tensor always yields zero. Without any claim to mathematical rigor this hand-waving argument turns out to be a consistent principle, namely, that physical theories have to be constructed in such a way that the symmetrical part of a derivative has no influence on scalars and is therefore not measurable.

One can therefore simply omit the symmetrical contribution without changing the physics. This can be achieved by always antisymmetrizing the result after differentiating. This *antisymmetrized* derivative

$$\mathrm{d} = \mathcal{A}(\tilde{\mathrm{d}}) \tag{2.96}$$

is referred to as the *exterior derivative* because it is defined inside the exterior algebra (for the definition of $\mathcal{A}$ please refer to Sect. 2.4 on page 38). With reference to the above example the representation of the exterior derivative applied to a 1-form α reads

$$\boxed{(\mathrm{d}\alpha)_{ij} \;=\; \partial_i \alpha_j - \partial_j \alpha_i} \tag{2.97}$$

However, the antisymmetrization has an unfamiliar consequence: Namely, if one wants to differentiate twice, one obtains

$$(\mathrm{d}^2\alpha)_{ijk} \;=\; (\mathrm{d}\,\mathrm{d}\alpha)_{ijk} \;=\; \partial_i\partial_j\alpha_k - \partial_j\partial_i\alpha_k + \partial_j\partial_k\alpha_i - \partial_k\partial_j\alpha_i + \partial_k\partial_i\alpha_j - \partial_i\partial_k\alpha_j \;=\; 0.$$

or, more generally:

$$\boxed{\mathrm{d}^2 = 0.} \tag{2.98}$$

Thus, in the exterior algebra there are no higher derivatives, only a first derivative. The formal reason is easy to understand: Since derivative operators commute, d^2 would be a symmetric construction, i.e., the two additional inputs of $\mathrm{d}^2\alpha$ attached are the two directions intention space would be symmetric under exchange and therefore would vanish on contraction with an antisymmetric tensor.

If exterior algebra is the correct basis for describing nature, it has a direct, highly restrictive consequence, namely:

Nature is based on first derivatives only.

The Hamilton equations of motion are differential equations of first order. But what about the Maxwell equations and the Schrödinger equation? Here the impression prevails that physics is based on second derivatives. And in fact, as we will see below, besides the d differential there is another *codifferential* $\mathrm{d}^\dagger$, which allows you to combine a derivative and a co-derivative and get something that in the conventional setting looks like a second derivative.

During your studies you certainly noticed that the fundamental laws of nature are built in such a way that only first and second derivatives occur, while no physical law involves the 27^{th} derivative. Often this is paraphrased by the so-called "heuristic principle of simplicity" - which basically means that physical laws are as simple as possible to be compliant with the required symmetry properties. The exterior algebra makes a lot of progress here because it provides a formal framework restricted to first and second derivatives, suggesting a possible explanation why we see only first and second derivatives in Nature. Of course, it still remains an open question why the exterior algebra plays such a fundamental role in the structure of space and time.

2.4.2 Exterior derivative

The *exterior derivative* is defined as an operator $\mathbf{d}$ that maps a field of p-forms on a field of $p+1$-forms. This operator has the following formal properties:

(i) The operator d applied to function f (i.e., a field of 0-forms) just reproduces the ordinary differential $\mathrm{d}f$ defined in Eq. (2.79) on page 61.

(ii) Applying d twice always returns zero, i.e., $\mathrm{d}^2 = 0$.

(iii) If d is applied to a wedge product, we can use the following product rule

$$\boxed{\mathrm{d}(\alpha \wedge \beta) = \mathrm{d}\alpha \wedge \beta + (-1)^{p_\alpha} \alpha \wedge \mathrm{d}\beta\,,} \tag{2.99}$$

where p_α is the rank of the field of forms α.

Since the exterior derivative increases the rank of the tensor, the exterior derivative of the volume form has to vanish:

$$\boxed{\mathrm{d}\omega = 0.} \tag{2.100}$$

2.4.3 Representation of the exterior algebra

Let $x_1, \ldots, x_n$ be a coordinate system and let ∂_i and $\mathrm{d}x^j$ be the corresponding vector fields of the tangent space and the cotangent space. By applying the above product rule and using $\mathrm{d}^2 = 0$ one can immediately show that the exterior derivative of the basis vectors of $\mathrm{d}x^{i_1} \wedge \ldots \wedge \mathrm{d}x^{i_p} \in \bigwedge^p T_p^* U$ vanish:

$$\mathrm{d}(\,\mathrm{d}x^{i_1} \wedge \ldots \wedge \mathrm{d}x^{i_p}) = 0. \tag{2.101}$$

Hence we can calculate the exterior derivative of a p-form

$$\alpha = \frac{1}{p!} \alpha_{i_1 \ldots i_p}\, \mathrm{d}x^{i_1} \wedge \ldots \wedge \mathrm{d}x^{i_p} \tag{2.102}$$

by interpreting the coefficients as functions (0-forms) and using the product rule (2.99), where $\alpha_{i_1 \ldots i_p}$ is the first factor and the basis vector $\mathrm{d}x^{i_1} \wedge \ldots \wedge \mathrm{d}x^{i_n}$ is the second factor. Since the second term in the product rule vansihes, the result reads:

$$\mathrm{d}\alpha = \frac{1}{p!} \left(\mathrm{d}\alpha_{i_1 \ldots i_p}\right) \wedge \mathrm{d}x^{i_1} \wedge \ldots \wedge \mathrm{d}x^{i_p}\,. \tag{2.103}$$

Because of $\mathrm{d}f = (\partial_j f)\,\mathrm{d}x^j$ one obtains

$$\boxed{\mathrm{d}\alpha = \frac{1}{p!} \frac{\partial \alpha_{i_1 \ldots i_p}}{\partial x^j}\, \mathrm{d}x^j \wedge \mathrm{d}x^{i_1} \wedge \ldots \wedge \mathrm{d}x^{i_p}.} \tag{2.104}$$

Because of the antisymmetry only those terms in the sum will contribute for which the index j differs from *all* other indices $i_1, \ldots, i_p$ of the tensor.

Examples:
1) Consider a 1-form $\alpha = \alpha_i\,\mathrm{d}x^i$ on $\mathbb{R}^3$. Then we have

$$\begin{aligned} \mathrm{d}\alpha &= (\partial_j \alpha_i)\,\mathrm{d}x^j \wedge \mathrm{d}x^i = \frac{1}{2!}\left(\partial_j \alpha_i - \partial_i \alpha_j\right) \mathrm{d}x^j \wedge \mathrm{d}x^i \\ &= (\partial_1 \alpha_2 - \partial_2 \alpha_1)\,\mathrm{d}x^1 \wedge \mathrm{d}x^2 + (\partial_1 \alpha_3 - \partial_3 \alpha_1)\,\mathrm{d}x^1 \wedge \mathrm{d}x^3 + (\partial_2 \alpha_3 - \partial_3 \alpha_2)\,\mathrm{d}x^2 \wedge \mathrm{d}x^3 \end{aligned}$$

2) For the 2-form $\gamma = \frac{1}{2!}\gamma_{ij}\,\mathrm{d}x^i \wedge \mathrm{d}x^j$ the exterior differential is given by

$$\mathrm{d}\gamma = \frac{1}{2}\partial_k \gamma_{ij}\,\mathrm{d}x^k \wedge \mathrm{d}x^i \wedge \mathrm{d}x^j = \frac{1}{2}\left(\partial_1 \gamma_{23} + \partial_2 \gamma_{31} + \partial_3 \gamma_{12}\right) \mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3$$

These coefficients can be made antisymmetric

$$(\mathrm{d}\gamma)_{ijk} = \frac{1}{2}(\partial_i\gamma_{jk} - \partial_i\gamma_{kj} + \partial_k\gamma_{ij} - \partial_k\gamma_{ji} + \partial_j\gamma_{ki} - \partial_j\gamma_{ik})$$

such that $\mathrm{d}\gamma = \frac{1}{3!}(\mathrm{d}\gamma)_{ijk}\,\mathrm{d}x^i \wedge \mathrm{d}x^j \wedge \mathrm{d}x^k$. Exercise: Show that $\mathrm{d}^2\gamma = 0$.

2.4.4 The Poincaré lemma

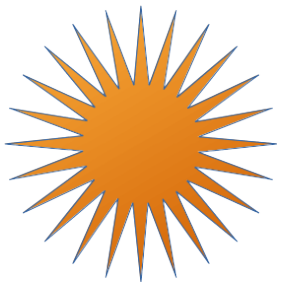
We start this section with two important definitions:

- A differential form α is said to be *closed* if the exterior derivative vanishes: $\mathrm{d}\alpha = 0$.
- A differential form $\alpha$is said to be *exact* if it can be expressed as the exterior derivative of another differential form β by $\alpha = \mathrm{d}\beta$. In this case the form β is called the *potential form* of α.

Because of $\mathrm{d}^2 = 0$ it is clear that every exact differential form is also closed. However, the opposite direction is not automatically fulfilled,. Whether it holds or not is the subject of the famous *Poincaré lemma*. Put simply, this lemma says:

> **Theorem:** In a star-shaped open set, each closed differential form is exact, i.e., for every closed p-form α one finds a $p-1$-form β, also called *potential form* or simply the *potential*, so that $\alpha = \mathrm{d}\beta$.

The star shape is required to ensure that the considered quantity does not have a hole, meaning that all closed integration contours can be topologically contracted to a single point. It suffices to prove the star shape with respect to a special coordinate system.



Star-shaped set.

> **Note:** This sounds familiar. A vortex-free vector field, i.e., a vector field on which the rotation operator returns zero, can be written as a gradient of a potential. The Poincaré lemma expresses this situation in a similar way for differential forms of arbitrary rank in arbitrary dimensions.

2.4.5 Relation to ordinary vector analysis

The differential operators of ordinary vector analysis can be expressed independent of coordinates in the differential calculus:

$$\begin{aligned} \operatorname{grad} f &= \nabla f &&= (\mathrm{d}f)^\sharp && (2.105)\\ \operatorname{div}\mathbf{X} &= \nabla\cdot\mathbf{X} &&= \star\,\mathrm{d}\star\mathbf{X}^\flat && (2.106)\\ \operatorname{rot}\mathbf{X} &= \nabla\times\mathbf{X} &&= (\star\,\mathrm{d}\mathbf{X}^\flat)^\sharp && (2.107) \end{aligned}$$

Here f is a scalar function and $\mathbf{X}$ denotes a vector field. With these expressions, known relations from standard vector analysis can be verified easily. For example we have

$$\operatorname{rot}\operatorname{grad} f = ((\star\,\mathrm{d}((\mathrm{d}f)^\sharp)^\flat)^\sharp = (\star\,\mathrm{d}^2 f)^\sharp = 0. \quad (2.108)$$

2.4.6 The co-differential operator

The *co-differential operator* $\mathrm{d}^\dagger$ is defined by

$$\boxed{\mathrm{d}^\dagger = s(-1)^{np+n+1} \star \mathrm{d} \star} \tag{2.109}$$

where $\star$ is the Hodge star operator, $s = \mathrm{sgn}(g)$and p is the rank of the differential form to which $\mathrm{d}^\dagger$it is applied.

The co-differential operator $\mathrm{d}^\dagger$ has in many ways the same properties as the normal differential operator. In particular we have

$$(\mathrm{d}^\dagger)^2 = 0 \tag{2.110}$$

One major difference, however, is that while d *increases* the rank of a p form to $p+1$, the co-differential operator *decreases* the rank:

$$p \underset{\star}{\longrightarrow} n-p \underset{\mathrm{d}}{\longrightarrow} n-p+1 \underset{\star}{\longrightarrow} p-1$$

i.e., the codifferential operator $\mathrm{d}^\dagger$ *decrements* the rank of a p-form by 1. An easy way to remember how the co-differential operator works is to first apply the ordinary derivative, which adds an additional input channel, and then to contract this additional channel with another input via Hodge.

Laplace-de Rham Operator

It should be noted at this point that in the exterior algebra the expression

$$\boxed{\Delta = (\mathrm{d} + \mathrm{d}^\dagger)^2 = \mathrm{d}\,\mathrm{d}^\dagger + \mathrm{d}^\dagger \mathrm{d}} \tag{2.111}$$

plays the role of the Laplace operator and is therefore referred to as the *Laplace-de Rham Operator* which is a generalization of the *Laplace-Beltrami-Operator*. In the Euclidean metric this operator coincides with the usual Laplacian up to a minus sign ($\Delta = -\nabla^2 = -\partial^\mu \partial_\mu$). The advantage of the Laplace-de Rham Operator is that it *does the right thing* in any metric, even, as we will see below, on curved manifolds. For example, a partial differential equation involving the Laplacian such as the wave equation can simply be put on a curved space by replacing the ordinary Laplacian with Δ defined above.

> **Remark:** Note that there is again a lot of confusion about the sign of the Lapalace-Beltrami operator. Most mathematicians prefer $\Delta = -\nabla^2$ because then Δ is positive definite. There are also two definitions of $\mathrm{d}^\dagger$ around which differ by a minus sign.

Hodge decomposition theorem

A differential form α is called *exact* if it can be written as the differential $\alpha = \mathrm{d}\beta$ of another form β of higher rank, the so-called *potential form*. Likewise, a differential form is called *co-exact* if it can be expressed as the codifferential $\alpha = \mathrm{d}^\dagger \gamma$ of another form γ of lower rank.

In addition, a differential form η is said to be *harmonic* if it satisfies the differential equation $\Delta\eta = 0$, where Δ is the Laplace-de-Rham operator introduced above.

The *Hodge decomposition theorem* states that any form can be written as a sum of an exact, a coexact, and a harmonic form.

$$\boxed{\text{any form } \alpha = \text{exact form } \mathrm{d}\beta + \text{coexact form } \mathrm{d}^\dagger\gamma + \text{harmonic form } \eta.}$$

2.4.7 Lie bracket

In Sect. 2.3.1 on page 58 we introduced *directional derivative* of functions and interpreted them as vectors or vector fields. The directional derivative of a function f, which is defined by

$$\mathbf{X}f = \mathrm{d}f(\mathbf{X}) = (\partial_j f)X^j \tag{2.112}$$

is obviously again a function. This allows us to apply several directional derivatives one after the other, for example we could apply $\mathbf{Y}\circ\mathbf{X}$ on f. Because of the usual product rule, the result expressed in components reads

$$\mathbf{Y}\circ\mathbf{X}f \;=\; \big(\partial_k[(\partial_j f)X^j]\big)Y^k \;=\; (\partial_k\partial_j f)X^jY^k + (\partial_j f)(\partial_k X^j)Y^k\,. \tag{2.113}$$

However, here the problem arises that in the first term there is a second derivative. This is bad news since a directional derivative applied to a scalar function should only involve first derivatives of the components, meaning that $\mathbf{Y}\circ\mathbf{X}$ is not a proper vector field. This means that the concatenation of vector fields leads us out of the exterior algebra. But as usual there is a way out, namely, if we define the antisymmetric commutator

$$[\mathbf{X},\mathbf{Y}] \;:=\; \mathbf{X}\circ\mathbf{Y} - \mathbf{Y}\circ\mathbf{X}\,, \tag{2.114}$$

the first term with the second derivative drops out and only products of first derivatives remain in place:

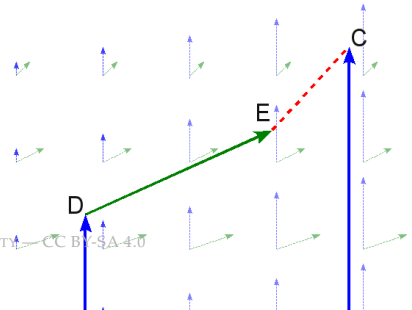
$$[\mathbf{X},\mathbf{Y}]f \;=\; (\partial_j f)(\partial_k Y^j)X^k - (\partial_j f)(\partial_k X^j)Y^k \;=\; (\partial_j f)\Big[\underbrace{(\partial_k Y^j)X^k - (\partial_k X^j)Y^k}_{[\mathbf{X},\mathbf{Y}]^j}\Big]\,. \tag{2.115}$$

Consequently, the *commutator* of two vector fields is again a healthy vector field. This means that the commutator of two vector fields is an operation which remains within the exterior algebra. This commutator is called *Lie-bracket* and plays a very fundamental role in mathematics and physics.

By construction, the Lie bracket is bilinear and antisymmetric. Moreover, under cyclic commutation it fulfills the so-called *Jacobi identity*

$$[\mathbf{X},[\mathbf{Y},\mathbf{Z}]] + [\mathbf{Z},[\mathbf{X},\mathbf{Y}]] + [\mathbf{Y},[\mathbf{Z},\mathbf{X}]] = 0\,. \tag{2.116}$$

The adjacent figure illustrates the geometric meaning of the Lie bracket. In the background you can see the vector fields $\mathbf{X}$ (green) and $\mathbf{Y}$



(blue). Starting at point A, you can move either with $\mathbf{Y} \circ \mathbf{X}$ along A-B-C, or you can move with $\mathbf{X} \circ \mathbf{Y}$ along A-D-E. However, if the length of the vector arrows happens to change in the other direction when moving in one direction, one does not arrive at the same point. Instead the final position C,E are separated by a gap, represented as a dashed red line in the figure. This difference is represented by the Lie bracket and scales bilinearly with the vector fields.

Recall that in a so-called coordinate basis (see Sect. 2.3.4 on page 62), the Lie-bracket of the basis vector fields is always equal to zero. Now we understand why: It is simply because coordinate systems are defined in such a way that each point is uniquely characterized by a certain set of coordinates. Hence, when moving along the grid lines, it is impossible to end up at different positions.

2.5 Integration of forms

When integrating forms, the rule is that the rank of the form over which to integrate is equal to the dimension of the geometric entity over which it is integrated. For example, curve integrals are integrated over 1-forms, surface integrals over 2-forms, etc.

Remember:

- 1-forms are integrated along curves
- 2-forms are integrated along surfaces
- p-forms are integrated over p-dimensional submanifolds
- The volume form is integrated over (sub)volumes

2.5.1 Special cases

1-form integrated along a curve

A 1-form α can be integrated along a curve c in U by choosing a convenient parametrization $c : (a, b) \to U$:

$$\int_c \alpha_c = \int_a^b \alpha_{c(\lambda)}\big[c'(\lambda)\big]\ \mathrm{d}\lambda\,. \tag{2.117}$$

Here $\lambda \in (a, b)$ is the curve parameter and $c'(\lambda) \in T_c U$ is the tangent vector along the curve with respect to the chosen parametrization (see Sect. 2.3.1 on page 58). In differential geometry, it is customary not to explicitly write down the dependency of the objects on the location, so the usual notation is

$$\int_c \alpha = \int_a^b \alpha\big(c'(\lambda)\big)\ \mathrm{d}\lambda\,. \tag{2.118}$$

If $\alpha = \mathrm{d}f$ is the differential of some function (called potential), the integral only depends on the endpoints and vanishes if the integration path is closed:

$$\int_c \mathrm{d}f = f_{c(b)} - f_{c(a)}\,, \qquad \oint_c \mathrm{d}f = 0\,. \tag{2.119}$$

Proof: To prove this statement, simply use the definition of directional derivative along a curve:

$$\int_c \mathrm{d}f_c \;=\; \int_a^b \mathrm{d}f_{c(\lambda)}(c'(\lambda))\,\mathrm{d}\lambda \;=\; \int_a^b \nabla_c f_{c(\lambda)}\,\mathrm{d}\lambda \;=\; \int_a^b \frac{\mathrm{d}}{\mathrm{d}\lambda} f_{c(\lambda)}\,\mathrm{d}\lambda \;=\; f_{c(b)} - f_{c(a)}\,. \tag{2.120}$$

Representation of a curve integral:
In a given coordinate system, the curve can be represented by coordinate functions $c^i(\lambda) = x^i(c(\lambda))$. The curve integral can then be represented as

$$\int_c \alpha = \int_a^b \alpha_i(\lambda)\frac{\mathrm{d}c^i(\lambda)}{\mathrm{d}\lambda}\,\mathrm{d}\lambda \tag{2.121}$$

Integrating n-form over the volume

To integrate over an n-dimensional volume in an n-dimensional space, the integrand Σ must be a n-form field. This n-form field Σ can only differ from the basis form $\Omega = \mathrm{d}x^1 \wedge \ldots \wedge \mathrm{d}x^n$ by a position-dependent factor σ:

$$\Sigma_p = \sigma_p\,\mathrm{d}x^1 \wedge \ldots \wedge \mathrm{d}x^n. \tag{2.122}$$

Here σ_p can be understood as a scalar function. Again, it is customary to suppress the location dependence by writing:

$$\Sigma = \sigma\,\mathrm{d}x^1 \wedge \ldots \wedge \mathrm{d}x^n. \tag{2.123}$$

In a given representation the volume integral can be expressed as an n-fold integral

$$\int_V \Sigma \;=\; \int \ldots \int \sigma(x^1, \ldots, x^n) dx^1 \cdots dx^n\,, \tag{2.124}$$

where the integration ranges have to be chosen according to the boundaries of the volume V. As you can imagine, the parametrization of a volume with a non-trivial shape can be a challenging task.

The factor σ is a function with different values at different points. In the volume form ω (see Sect. 2.1.8 on page 45), this function was just chosen in such a way that the volume integral returns the actual metric volume of the integration area.

Remark: The function σ depends on the choice of coordinates. In Cartesian coordinates in $\mathbb{R}^3$, the volume form is given by $\omega = \mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z$, corresponding to $\sigma = 1$, whereas in spherical coordinates we have $\omega = r^2 \sin\phi\; \mathrm{d}r \wedge \mathrm{d}\theta \wedge \mathrm{d}\phi$, corresponding to $\sigma = r\sin\phi$.

2.5.2 Generic integrals over p-forms

So far we have seen that 1-forms can be integrated along curves while n-forms can be integrated over volumes. Similarly, a p -form α can be integrated over connected p-dimensional domains G. The computation of such integrals requires a parameterization $\mathbf{x}(\lambda^1, \ldots, \lambda^p)$ of the domain in terms of p parameters $\lambda^1, \ldots, \lambda^p$. In a given presentation

$$\alpha = \frac{1}{p!}\alpha_{i_1 \ldots i_p}\, \mathrm{d}x^{i_1} \wedge \ldots \wedge \mathrm{d}x^{i_p} \tag{2.125}$$

the integral over the domain G can then be expressed by

$$\int_G \alpha \;=\; \int \cdots \int \alpha_{i_1 \ldots i_p}\big(\mathbf{x}(\lambda_1, \ldots, \lambda_p)\big) \left| \frac{\partial(x^{i_1}, \ldots, x^{i_p})}{\partial(\lambda^1, \ldots, \lambda^p)} \right| d\lambda^1\, d\lambda^2 \cdots d\lambda^p\,, \tag{2.126}$$

where $|\cdot|$ denotes the Jacobi matrix. Here the integration bounds should be have to be chosen in such a way that the entire area G is covered (which is sometimes not easy to realize).

2.5.3 Stokes theorem

From the introductory courses on vector analysis you know the two integral theorems of Gauss

$$\int_V \nabla \cdot \vec{A}\, dV \;=\; \oint_{\partial V} \vec{A} \cdot \vec{n}\; dS \tag{2.127}$$

and Stokes

$$\int_S (\nabla \times \vec{A}) \cdot d\vec{n} \;=\; \oint_{\partial S} \vec{A} \cdot d\vec{l}\,. \tag{2.128}$$

Forget it! We are now getting to something much simpler, namely, the *generalized Stokes Theorem*

$$\boxed{\int_G \mathrm{d}\alpha \;=\; \int_{\partial G} \alpha\,,} \tag{2.129}$$

where α is a p-form and G is a $p+1$-dimensional integration domain with the boundary ∂G. If the domain of integration does not have a boundary (as e.g. a closed curve on a sphere), the right hand side of the equation equals zero. With this very memorable theorem the handling of integrals is simplified considerably.

2.6 Tensor-valued forms *

So far, we have come to know two categories of tensors, namely, general tensors that are constructed with the ordinary tensor product '$\otimes$', and the subset of antisymmetric tensors (forms) constructed with the wedge product '$\wedge$' and for which we have defined a self-contained set of computational rules, the so-called exterior algebra.

In between there are also mixed kinds of tensors, which are antisymmetric in one part of their connections, but arbitrary in the others. In this case it is common to no longer

consider these tensors as mappings to $\mathbb{R}$, but to interpret the antisymmetric slots as inputs and the remaining ones as outputs.

As an example let us consider a *vector-valued p-Form*

$$\mathbf{T} \;=\; \frac{1}{p!}\, T^{i}{}_{j_1\ldots j_p}\, \mathbf{e}_i \otimes \mathbf{e}^{j_1} \wedge \ldots \wedge \mathbf{e}^{j_p}\,. \tag{2.130}$$

The inputs, i.e., the arguments, now refer exclusively to the antisymmetrized tensor components. This means that the form maps p vectors $\mathbf{X}_{(1)}, \ldots, \mathbf{X}_{(p)}$ to

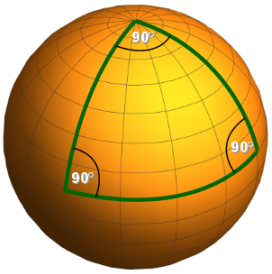
$$\mathbf{T}(\mathbf{X}_{(1)}, \ldots, \mathbf{X}_{(p)}) \;=\; T^{i}{}_{j_1\ldots j_p} X^{j_1}_{(1)} \ldots X^{j_p}_{(p)}\, \mathbf{e}_i\,. \tag{2.131}$$

Thus it is vector-valued, but apart from this it behaves algebraically as any other p-form.

3 Elementary concepts of differential geometry

3.1 Manifolds

Differential geometry deals with the geometry of curved spaces, so-called *manifolds*. A simple example is the surface of a sphere. A manifold $\mathcal{M}$ is characterized by a certain dimension n and it is built in such a way that it looks almost like a flat $\mathbb{R}^n$ on small distances, just like the surface of the sea appears to be locally flat.



The sum of the three angles of a triangle on a curved object differs from 180° [Wikimedia].

More precisely: A real (complex) n-dimensional manifold $\mathcal{M}$ is a Hausdorff space in which the local neighborhood of each point is *homeomorphic* to the vector space $\mathbb{R}^n$ ($\mathbb{C}^n$) ist. A *homeomorphisms* is a continuous invertible map whose inverse is also continuous.

Embedded and abstract manifolds

In many situations such a manifold is *embedded* in a higher-dimensional flat vector space[1] such as e.g. the two-dimensional spherical surface shown in the figure above is embedded in the $\mathbb{R}^3$. When differential geometry was developed, however, it turned out that there are also so-called *abstract manifolds* that can not be embedded into a higher dimensional vector space. As we shall see, the 4-dimensional curved space-time in the theory of general relativity is such an abstract manifold that can not be embedded in a superordinate 5-dimensional (flat) vector space.

In order to be able to describe such manifolds mathematically, differential geometry has to be formulated in such a way that that it does not require the concept of an embedding space. With reference to the example of a spherical surface shown above, this would mean that one describes its curved geometry without making use of the fact that we can move away from the spherical surface in the radial direction, exploring the space outside. In other words, we would like to characterize curvature without making use of the third dimension in normal direction. Modern differential geometry thus searches for an *intrinsic* description of the curved space, without resorting to a surrounding embedding space. For example, an *intrinsic curvature* can be detected by verifying that the sum of the three angles in a triangle is not equal to π (see figure).

[1] A spaces called *flat* if it does not have any curvature. Vector spaces are by definition flat.

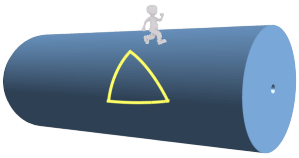


Figure 3.1: Triangle on a cylinder

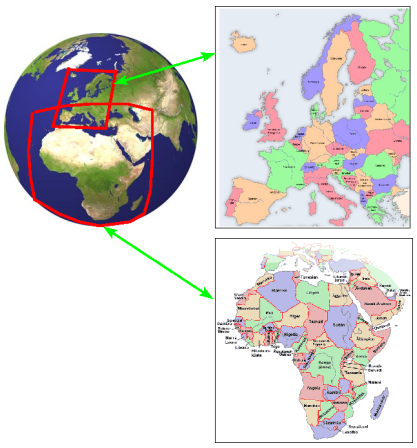
Remark: If a surface embedded into a flat vector space appears to be curved, this does not automatically imply that it is *intrinsically* curved. For example, a triangle on a cylinder always has the conrrect angle sum of 180°. A creature confined to the cylinder surface, to which the third dimension is not accessible, would therefore be unable to detect a local curvature. Thus, the intrinsic curvature of a cylinder is zero. In fact, we can make a cylinder out of a sheet of paper without wrinkling the paper, which would be impossible in the case of the sphere. Nevertheless the creature would be able to go once around returning to the same point, a property which is related to the *topology* of the cylinder.

3.1.1 Maps

Using the $\mathbb{R}^n$ as a mathematical description for the position space, we are accustomed to characterize the points in the space by the corresponding vector in a given coordinate system, for example by saying that a particle *is at location* $\mathbf{x} \in \mathbb{R}^n$. The same applies to the Minkowski spacetime in the special theory of relativity, where the so-called *events* (the points of the Minkowski space) are represented by four-vectors x^μ.

On a manifold, it is not so easy to describe the location of a point. If an embedding space was available, it would be of course possible to use vectors in the embedding space, e.g. we could describe the surface of a sphere by the set of all vectors $\{\mathbf{r}\}$ with the length $||\mathbf{r} - \mathbf{r}_0|| = R$ pointing from the center $\mathbf{r}_0$ to the surface. However, if one wishes to forget about the surrounding embedding space, this concept fails for several reasons. For example, the center of the sphere is not an element of its surface and therefore requires to make use of the third dimension perpendicular to the surface. If we restrict ourselves to the manifold, the position vectors should also live on the manifold, but have you ever seen a bent vector? Certainly such objects would not obey the usual vector space axioms.

To work around this problem, we can represent the manifold by one or several *maps*, much like real maps of the earth's surface. Since the manifold is nearly flat at short distances, we can assume that for every point $p \in \mathcal{M}$ there is a local environment $U(p) \subset \mathcal{M}$ equipped that can be represented on a map $\varphi : U \to \mathbb{R}^n$. Often, a single map is not enough to cover the entire manifold, so one needs a *collection* of several overlapping maps to cover the entire manifold. Referring to geography, such a collection is said to be an *atlas*.



Several maps forming an atlas.

In mathematical terms, a map is defined as a pair (U, φ) consisting of an open subset $U \subset \mathcal{M}$ and a *homeomorphism* $\varphi : U \to \mathbb{R}^n$. Recall that

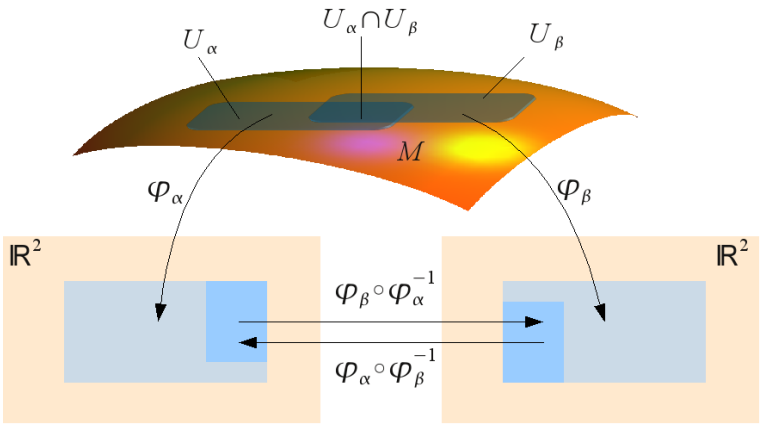


Figure 3.2: Going from one map to the other: The figure shows the manifold with two maps $(U_\alpha, \varphi_\alpha)$ and (U_β, φ_β), whose open original images overlap so that they have a common open intersection $U_\alpha \cap U_\beta$. As shown above, a map change is a bijective map between the images $\varphi_\alpha(U_\alpha \cap U_\beta)$ and $\varphi_\beta(U_\alpha \cap U_\beta)$, which can be constructed by a sequential application of φ_α^{-1} and φ_β.

a set is called *open* if every point $p \in U$ is surrounded by a neighborhood that lies completely in U, meaning that no point touches the boundary of the set. A collection $\{U_i\}$ of open subsets of $\mathcal{M}$ is called an *open cover* of $\mathcal{M}$ if $\bigcup_i U_i = \mathcal{M}$. The openness ensures that adjacent subsets overlap, i.e., they always have a non-empty intersection. A collection of maps $\{(U_i \varphi_i)\}$ whose subsets U_i cover the manifold $\mathcal{M}$ is referred to as an *atlas* of $\mathcal{M}$.

Atlases thus give us the possibility to map an n -dimensional manifold onto subsets of $\mathbb{R}^n$ and then to represent objects such as points in the usual way. Atlases are not unique because there are infinitely many possible projections and partitions. So, if you want to compute an abstract property of a manifold with the help of maps, the result must be independent of the chosen representation, that is, it has to coincide for all possible atlases.

Even the sphere $S^2 \subset \mathbb{R}^3$ cannot be represented by a single map, but requires at least two maps, e.g. one for the northern and another one for the southern hemisphere. In differential geometry, adjacent maps are chosen in such a way that they overlap. These overlap areas ensure that you can easily switch from one map to another.

3.1.2 Changes between different maps

Changes between different maps are much like changes between different coordinate systems. As an example let us consider two maps of an atlas, whose original images U_α und U_β on the manifold overlap (see Fig. 3.2), meaning that the common intersection $U_s := U_\alpha \cap U_\beta$ is represented in both maps. As shown in the figure, we can switch between the images $\varphi_\alpha(U_s) \in \mathbb{R}^n$ and $\varphi_\beta(U_s) \in \mathbb{R}^n$ by means of the sequential map

$$f : \mathbb{R}^n \to \mathbb{R}^n : \quad f := \varphi_\beta \circ \varphi_\alpha^{-1} \tag{3.1}$$

and its inverse $f^{-1} := \varphi_\alpha \circ \varphi_\beta^{-1}$. Such a mapping is continuous by construction and is called *coordinate transformation*.

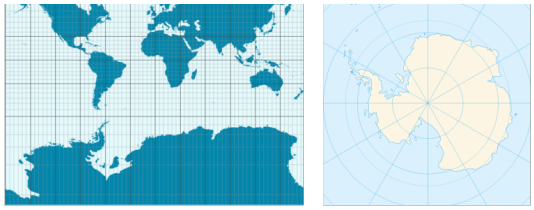


Figure 3.3: Antarctic represented in two different projections.

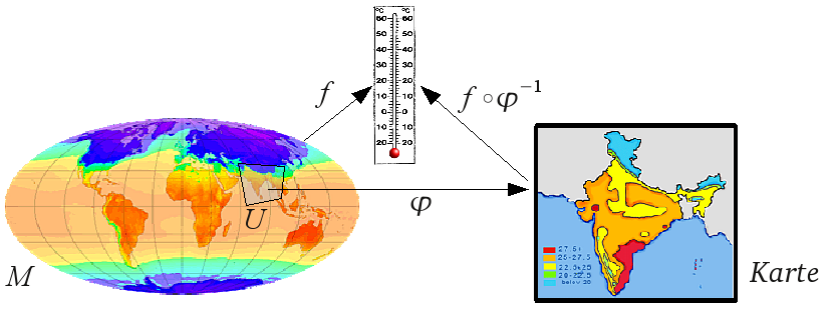


Figure 3.4: Function $f : M \to \mathbb{R}$ and its representation $f \circ \varphi^{-1}$ on a map (U, φ).

Remark: Although a coordinate transformation f maps subsets from $\mathbb{R}^n$ to $\mathbb{R}^n$, that is, an ordinary vector space on itself, this does not necessarily mean that f is linear. For example, the two maps on the right side overlap in the area of Antarctica. The Antarctic, however, appears to be extremely distorted in the top of the Mercator projection, while it appear to be very different from the bird's-eye view shown below. In this case it is clear that the map f mediating between these two representations is nonlinear.

Two maps are called C^k compatible if the coordinate transformation between them is k -fold-differentiable, that is, if all the k^{th} order partial derivatives exist. A manifold is called *differentiable* or *smooth*, if the maps of their atlases are C^∞-compatible, so coordinate transformations can be differentiated infinitely often. A manifold is called *analytic* if the coordinate transformations can be expanded.

3.1.3 Functions on manifolds

On a manifold $\mathcal{M}$ it is possible to declare *functions* f, which assign to every point $p \in \mathcal{M}$ a value $f(p)$. For example, the temperature on the Earth's surface can be understood as a map $f : \mathcal{M} \to \mathbb{R}$.

If we represent the manifold on a map, each point on the map is of course associated with a certain point on the manifold and thus with a certain value of f. This means that a map $f : \mathcal{M} \to \mathbb{R}$ automatically induces a corresponding function $F = f \circ \varphi^{-1} : \varphi(U) \to \mathbb{R}$, which maps a location $\mathbf{x}$ on the map to the corresponding value of the function $F(\mathbf{x}) := f(\varphi^{-1}(\mathbf{x}))$. This chaining of the maps is illustrated in Fig. 3.4.

A function $f : U_p \to \mathbb{R}$ on an open subset $U_p \subset \mathcal{M}$ is called *differentiable* at the point $p \in U$, if the associated function is on the map $F : \varphi(U_p) \to \mathbb{R}$ at the corresponding position $\varphi(p)$ is differentiable in the usual sense. It can be proved that this concept of differentiability is representation-independent, i.e., it is independent of the chosen

map. The set of all functions $f : U_p \to \mathbb{R}$ that can be differentiated in p will be denoted $\mathcal{F}_p(\mathcal{M})$.

A function $f : \mathcal{M} \to \mathbb{R}$ is called *(globally) differentiable* if it is differentiable in every point $p \in \mathcal{M}$. In the following the set of all differentiable functions on $\mathcal{M}$ will be denoted by $\mathcal{F}(\mathcal{M})$.

3.2 Tangent space and cotangent space

A sailor is instructed to sail at 20 knots speed in a north-westerly direction. This instruction can be interpreted as specifying a velocity vector $\mathbf{v}$ in a plane called *tangent space*.

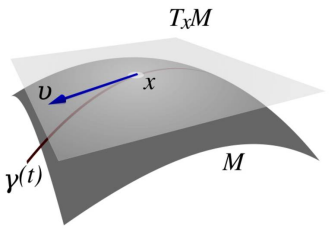


Illustration of the tangent space $T_p\mathcal{M}$ in a point $p \in \mathcal{M}$.

As shown in the figure on the right, the tangent space refers to a particular point p of the manifold and is therefore denoted $T_p\mathcal{M}$ (tangent space of $\mathcal{M}$ in p). One may view $T_p\mathcal{M}$ as a local space attached in p, which, in contrast to the manifold itself, is always flat, i.e., it is a vector space isomorphic to $\mathbb{R}^n$.

The graphic representation suggests to interpret the tangent space as a subset of a surrounding *embedding space*. But how do you define the tangent space if no embedding space is available? Vectors that 'stick out' of the manifold do not make sense here. So how can you characterize the speed of a ship on a curved manifold? The solution to this problem has already been addressed in Sect. 2.3.1 on page 58 and will be recalled in the following.

3.2.1 Directional derivatives and differentials:

The ship's path as a function of time is described by a parameterized smooth trajectory $c : \mathbb{R} \to \mathcal{M}$, where we want to assume without restriction that $c(0) = p$. Naively one would first try to define the speed of the ship as a derivative

$$c'(0) = \frac{\mathrm{d}}{\mathrm{d}t} c(t)\Big|_{t=0} = \lim_{\tau \to 0} \frac{c(\tau) - c(0)}{\tau}\,, \tag{3.2}$$

but, as you remember, this is impossible, since the manifold does not have a vector space structure, hence the difference of points $c(\tau) - c(0)$ has no meaning. To circumvent this difficulty, one asks oneself how *scalar functions* on the manifold change along the path, for example, we could ask how quickly the temperature changes when passing through the point p as a function of time. The point is that for a given function $f \in \mathcal{F}(\mathcal{M})$ that can be differentiated in p, the derivative of the concatenated map $f \circ c : \mathbb{R} \to \mathbb{R}$

$$\partial_c f := \frac{\mathrm{d}}{\mathrm{d}t} f(c(t))\Big|_{t=0} \tag{3.3}$$

is well-defined (see Eq. (2.72)). For a given point $p \in \mathcal{M}$, the value of this derivative seems to depend only on the *local* properties of the curve c and the function f in the point p, but not on properties of the curve in the function away from this point. For given p we can therefore define equivalence classes for both the trajectories and the functions:

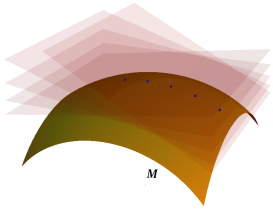
- **Equivalent trajectories:** $c_1 \sim c_2 \;\Leftrightarrow\; \partial_{c_1} f = \partial_{c_2} f \;\; \forall f \in \mathcal{F}_p(\mathcal{M})$
  Two trajectories are said to be equivalent in p if they pass the point p with the same direction and velocity. This set of the corresponding equivalence classes is denoted as the *tangent space* $T_p\mathcal{M}$, whose elements $X_p \in T_p\mathcal{M}$ can be interpreted as *directional derivatives*. One can easily show that the tangent space $T_p\mathcal{M}$ it is indeed a vector space.

- **Equivalent functions:** $f_1 \sim f_2 \;\Leftrightarrow\; X_p f_1 = X_p f_2 \;\; \forall X_p \in T_p M$.
  Two functions are said to be equivalent p, if they coincide in all their directional derivatives in the point p. The set of the corresponding equivalence classes is denoted as *cotangent space* $T_p^*\mathcal{M}$. The elements of this space $\mathrm{d}f_p \in T_p^*\mathcal{M}$ are interpreted as *differentials*, i.e., as 1-forms acting on the tangent space according to

$$\boxed{\mathrm{d}f_p(\mathbf{X}_p) = \mathbf{X}_p(f)\,.} \tag{3.4}$$

3.2.2 Tangent bundle and cotangent bundle

Each point $p \in \mathcal{M}$ of the manifold is assigned a *individual* tangent space $T_p\mathcal{M}$. Although all these spaces are isomorphic to $\mathbb{R}^n$, it is important to note that they are *different* spaces, that is, disjoint sets.



The tangent bundle is the collection of all tangent spaces

The disjoint union of all tangent spaces, so to speak the whole bunch of all tangent spaces for all points $p \in \mathcal{M}$, is called *tangent bundle* $T\mathcal{M}$. Similarly, the *cotangent bundle* $T^*\mathcal{M}$ is defined as a disjoint union of all cotangent spaces $T_p^*\mathcal{M}$.

Remark: One should realize that the tangent space...

- is always flat while the manifold can be curved.
- does not require the existence of an embedding space and therefore – contrary to what the above figure suggests – should not be interpreted as a subspace of an embedding space.
- is isomorphic to the $\mathbb{R}^n$. The same applies to maps of an atlas, but in no way maps should be confused with tangent spaces.

3.2.3 Excursus: fiber bundles *

Tangential and cotangent bundles are examples of so-called *fiber bundles*. To understand this term with a simple example, consider the xy plane shown in Fig. 3.5. This so-called *total space* $E = \mathbb{R}^2$ can be interpreted as a indbase space $B = \mathbb{R}$ (the x-axis), where a vertical *fiber* is attached at each point in the y direction, so that the total space is the

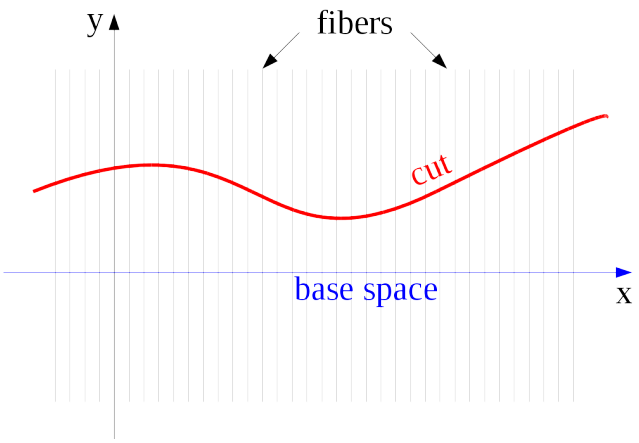


Figure 3.5: Simple example of a fiber bundle (see text)

disjoint union of all the fibers. For this reason such a construction is referred to as *fiber bundle*. In this space there exists a natural projection $\pi : E \to B$, the so-called *bundle projection*, which maps each fiber onto its base point. In the present example each fiber is mapped to the corresponding point on the x-axis.

Now imagine a function $f(x)$ in the xy-plane. This function cuts each fiber in half, thus defining a *cut* in the fiber bundle. If the function is continuously differentiable, this cut is said to be smooth.

In mathematics, the base space can be any topological space. In the context of general relativity it is usually a differentiable manifold, namely the curved space-time. Depending on how the fibers are made, i.e., what kind of mathematical objects are attached at the points of the base space, one distinguishes between different types of fiber bundles. Normally these are vector spaces, in this case one speaks of *vector bundles* or *vector space bundles*, i.e., families of vector spaces which are parameterized by the points of a manifold.

A *tangent bundle* $T\mathcal{M}$ is a special *vector space bundle* whose fibers are just the tangent spaces $T_p\mathcal{M}$. The fibers of the corresponding *cotangent bundle* $T^*\mathcal{M}$ are the dual cotangential spaces $T_p^*\mathcal{M}$. A *vector field* is a continuum of vectors in $T\mathcal{M}$ whose components are continuously differentiable functions of the coordinates on all possible maps. A vector field is therefore a *section* in the tangent bundle. Likewise, a field of 1-forms or differentials is accordingly a section in the cotangent bundle.

3.2.4 Coordinate basis

A differentiable manifold is represented by a collection of maps. Let $\varphi : U \to \mathbb{R}^n$ be such a map representing a subset $\varphi : U \to \mathbb{R}^n$. The vector components x^μ on the map can then be understood as n differentiable functions $x^\mu : U \to \mathbb{R}$, which are called *coordinates*. As already described in Sect. 2.3.4 on page 62, this singles out a particular basis:

- The curves, for which all coordinates except for x^μ are constant, represent in each point $p \in \mathcal{M}$ a set of directional derivatives $\mathbf{e}_\mu = \partial_\mu = \frac{\partial}{\partial x^\mu}$ which form the coordinate basis of $T_p\mathcal{M}$.
- The differentials $\mathrm{d}x^\nu$ of the coordinate functions, which by definition obey the relation $\mathrm{d}x^\nu(\partial_\mu) = \partial_\mu x^\nu = \delta^\nu_\mu$ form the corresponding dual coordinate basis of the cotangent space $T^*_p\mathcal{M}$.

This so-called *coordinate basis* defined above depends strongly on the choice of the map. With respect to a given metric $\mathbf{g}$, the basis vectors ∂_μ are generally neither normalized nor orthogonal. In the general theory of relativity, it is customary to mark indices referring to the coordinate basis by Greek indices.

Representation in the coordinate basis $\Leftrightarrow$ Greek indices

As we shall see in the following section, a coordinate basis is characterized by vanishing structural coefficients and therefore plays a special role.

3.2.5 Structural coefficients

The choice of a representation or a basis is to a large extent a matter of taste and has no influence on the physics, but it can have a significant influence on the computational effort. In some cases, the coordinate basis may prove inappropriate, and you would better work with a different particularly tailored basis $\{\mathbf{e}_i\}$. In that case, it is customary to write the components with Latin indices in order to distinguish them from the coordinate basis. Of course, such an alternative set of basis vector fields can again be represented in the coordinate basis:

$$\mathbf{e}_i = e^\mu_i \partial_\mu \tag{3.5}$$

The inverse map reads

$$\partial_\mu = e^k_\mu \mathbf{e}_k\,, \tag{3.6}$$

where e^k_μ is the inverse of the matrix e^μ_i. What distinguishes the coordinate basis from a general basis? What makes it so special? To understand this we consider the *Lie-bracket* of two basis vectors.

> **Remember:** Vectors, when applied to a function, yield the directional derivative, which in turn is again a function. This allows as to apply vectors several times, but the result of such a multiple application is generally no longer a vector since higher derivatives are created. In the Lie bracket (commutator), however, the second derivatives cancel out so that the Lie bracket maps two vectors to a new one. The Lie bracket therefore provides the proper operation within the exterior algebra. See Sect. 2.4.7 on page 73.

The Lie bracket reads

$$[\mathbf{e}_i, \mathbf{e}_j] = \Big((\partial_\nu e^\mu_j) e^\nu_i - (\partial_\nu e^\mu_i) e^\nu_j\Big)\partial_\mu \tag{3.7}$$

or

$$[\mathbf{e}_i, \mathbf{e}_j] = c^k{}_{ij}\mathbf{e}_k\,, \tag{3.8}$$

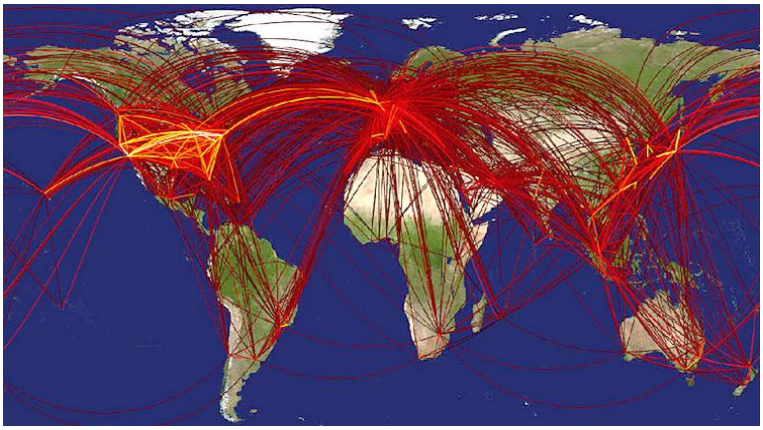


Figure 3.6: Although airplanes fly straight ahead on the shortest path, their routes appear curved on a map. [Image taken from: T. Geisel, Göttingen]

where the coefficients

$$c^k{}_{ij} = (\partial_\nu e_j^\mu) e_i^\nu e_\mu^k - (\partial_\nu e_i^\mu) e_j^\nu e_\mu^k \tag{3.9}$$

are the so-called *structural coefficients* of the basis. A simple calculation immediately shows that the structural coefficients in the coordinate basis are zero. This special property distinguishes a coordinate basis from a general basis.

This result can be interpreted in an illustrative way as follows. The Lie bracket $[\mathbf{X}, \mathbf{Y}]$ can be considered as a commutator of two displacements. This commutator tells us whether it makes a difference to lowest order if move first in $\mathbf{X}$ direction and then in $\mathbf{Y}$ direction or vice versa. It is easy to find examples for which this commutator is non-zero.[2] Coordinate systems are, however, by definition built in such a way that it does not matter in which order you follow the gridlines, that is, whether you move first in x-direction and then in y-direction or vice versa – you always get to the same point that is uniquely described by the coordinates.

3.3 Parallel transport

3.3.1 Transport of geometric objects

Although the surface of the earth is curved, captains or pilots still have a clear notion of what it means to move 'straight forward' by keeping the rudder in its neutral position. Such a trajectory describing a straight motion on a curved manifold is called a *geodesic line* or short *geodesic*. As we shall see, geodesic lines are precisely those lines which connect two points by the shortest route. For this reason, airplanes and ships prefer to move on geodesic lines. Although geodesic lines stand for 'moving straight', they usually do not appear as straight lines on a map (see Fig. 3.6).

A central problem in differential geometry is the transport of information from one place to another. What exactly does it mean to transport a mathematical object on a

[2] e.g. $\mathbf{X} = x\partial_x$ and $\mathbf{Y} = \partial_y$ in $\mathbb{R}^2$

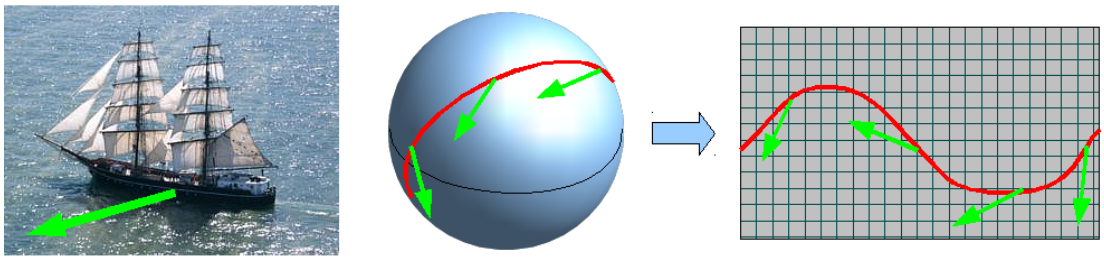


Figure 3.7: A ship transporting a tangential vector (see text).

ship? For scalars, this is very simple: the scalar, e.g. the number 27, is taken on board, transported and finally unloaded at the destination without changing its value. But how do you transport tangential vectors? Tangential vectors represent directions, they belong to a tangent space in a particular point, but what happens if we move from one tangential space the other?

For the ship's captain, the tangential vector to be transported is an arrow pointing in a certain direction. As long as the ship is traveling straight, it is reasonable to keep the arrow as it is, meaning that the relative angle between the ship and the arrow remains constant. However, If the ship changes its direction of travel by the angle ϕ, it is reasonable to turn the arrow relative to the ship by the opposite angle $-\phi$ in order to preserve the 'true' orientation of the vector. In fact, this simple protocol allows tangential vectors to be transported on any given orbit.

3.3.2 Parallel transport of tangent vectors

While transporting a tangential vector is a fairly straight-forward procedure from the captain's perspective, the situation on a map can look a lot more confusing. To see this, let us imagine a ship going straight on a great circle that is not in the equatorial plane (see Fig. 3.7). On a world map, this route has a shape that resembles a sine wave, and the vector being transported seems to be constantly changing direction. Of course, this change of direction is only an apparent one, which is due to the particular choice of the map. In fact, on a map, apparent (coordinate-related) and real (caused by changes in the course of the ship) direction changes are generally superimposed and both influences are not easy to disentangle at first glance.

We now want to formulate this phenomenon in more detail. Fig. 3.8 shows a part of the trajectory of a ship displayed on a map. This ship transports a tangent vector $\mathbf{Z}$ from the point $c(\lambda)$ to the point $c(\lambda+\delta\lambda)$ along a geodesic. The result of this displacement is the vector $\mathbf{Z}'$. Although *in reality* the vector keeps its direction, it will generally change its *apparent direction in the representation* on the map, that is, on the map $\mathbf{Z}'$ seems to point in a different direction than $\mathbf{Z}$. This illustrates that on the map it is not enough just to move the vector $\mathbf{Z}$ (which would give the dashed vector), but an additional correction is needed accounting for the nonlinearity of the representation, which is shown in the figure as a red difference vector.

This red difference vector will increase with magnitude of the displacement $\delta\lambda$ and of

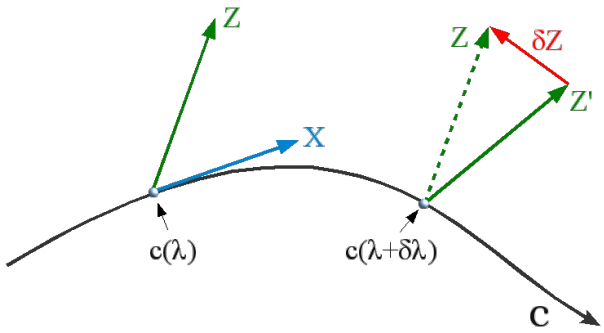


Figure 3.8: Parallel transport of a vector along a geodesic (see text).

course with the length of the vector to be transported. For small $\delta\lambda \ll 1$, it is reasonable to assume that both dependencies are linear. So we expect that

$$\delta\mathbf{Z} \;=\; \delta\lambda\, \Gamma(\mathbf{X}, \mathbf{Z})\,, \tag{3.10}$$

where $\mathbf{X} = \frac{d}{d\lambda} c(\lambda)$ is the tangent vector along the curve and where $\Gamma(\mathbf{X}, \mathbf{Z})$ is a yet to be specified bilinear mapping of two vectors $\mathbf{X}$ and $\mathbf{Y}$ onto some vector. So for a given $\mathbf{X}$, one has a linear map that generates the necessary directional correction of $\mathbf{Y}$ on the map in such a way that this vector retains its 'real' direction on $T\mathcal{M}$.

3.3.3 Covariant derivative of vector fields

Instead of a single vector, let us now consider a tangent vector field $\mathbf{Y}$ on the manifold. An example would be the wind speed on the earth's surface, which we imagine here to be location-dependent but time-independent. We would like to know how the vector field changes if we move in a given direction $\mathbf{X}$. Usually this requires to compute the *directional derivative*.

A directional derivative $\nabla_{\mathbf{X}}\mathbf{Y}$ of a vector field is expected to return the *rate* at which $\mathbf{Y}$ changes when moving in the direction of $\mathbf{X}$ straight ahead. This means that the directional derivative itself is vector-valued.

From the captain's perspective, this derivative is easy to calculate. First, he measures the wind speed and direction, thus determining the value of the vector field $\mathbf{Y}$ at current position of the ship. The captain then sails a short while in the direction given by $\mathbf{X}$, leaving the measured vector unchanged, that is, transporting it in parallel according to the protocol described above. Thereafter, the wind direction is measured again and compared with the previous one, the captain computes the difference between the current vector and the previous one which has been transported in parallel. Finally the difference divided by the distance traveled gives the directional derivative.

Because this procedure requires the parallel transport of a vector, the representation of such a directional derivative on a map must take into account the directional correction described above. This mechanism is outlined in Fig. 3.9. In comparison to the previous illustration, an additional vector field $\mathbf{Y}$ is shown here, indicated by orange field lines. At the starting point $c(\lambda)$, this vector field has the value $\mathbf{Y} = \mathbf{Y}(\lambda)$. The ship

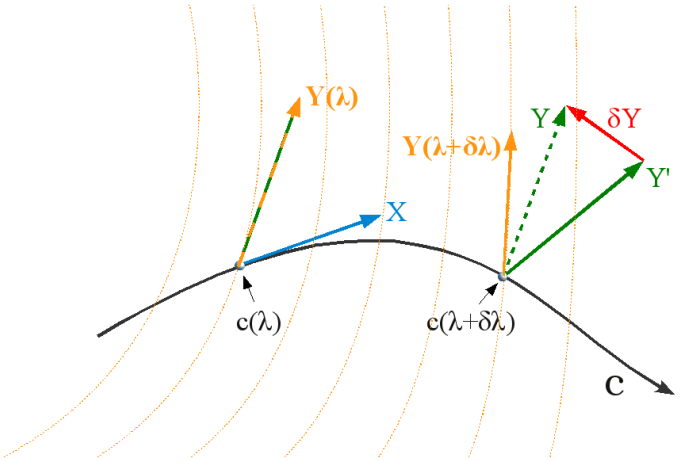


Figure 3.9: Heuristic motivation of the covariant derivative of a vector field (see text).

takes this vector on board and transports it in parallel to the destination $c(\lambda + \delta\lambda)$. In reality the direction of the transported vector $\mathbf{Y}'$ is unchanged at the destination, but on the map it appears to be twisted relative to the original vector $\mathbf{Y}$ with a correction $\delta\mathbf{Y} \approx \delta\lambda\,\Gamma(\mathbf{X},\mathbf{Y})$. The vector field at the destination $\mathbf{Y}(\lambda+\delta\lambda)$ is compared to the vector transported in parallel and divided by the distance, i.e., the directional derivative along the curve is given by

$$\nabla_{\mathbf{X}}\mathbf{Y} = \lim_{\delta\lambda\to 0} \frac{\mathbf{Y}(\lambda+\delta\lambda) - \mathbf{Y}'}{\delta\lambda}\,. \tag{3.11}$$

Therefore we get

$$\nabla_{\mathbf{X}}\mathbf{Y} = \lim_{\delta\lambda\to 0} \frac{(\mathbf{Y}(\lambda+\delta\lambda) - \mathbf{Y}) + (\mathbf{Y} - \mathbf{Y}')}{\delta\lambda} = \partial_{\mathbf{X}}\mathbf{Y} + \Gamma(\mathbf{X},\mathbf{Y}) \tag{3.12}$$

The 'true' directional derivative, which is known as the *covariant derivative* in differential geometry, thus differs from the usual directional derivative on the map by a correction in the form of a bilinear mapping $\Gamma(\mathbf{X},\mathbf{Y})$.

3.3.4 Connections

In the differential geometry, the process of correcting the apparent direction of an object on a map is realized by so-called *connections* (Zusammenhänge). A connection is a mathematical prescription for getting from one from one fiber of a fiber bundle to the adjacent fiber. In simple terms, this rule specifies the transformation by which adjacent fibers are 'glued together', i.e., how neighboring fibers of the fiber bundle are related. In the present case, where the fibers are the tangent spaces, the connection is just the covariant derivative. As we will see later, intrinsic curvature of the manifold can be thought of as some kind of distortion in gluing the tangent spaces.

Connections can be introduced and represented in different ways. Here we adopt a representation-free formulation introduced by *Koszul* in 1950. Accordingly, a connection is defined as a map ∇, which acts on two continuously differentiable vector fields

X and **Y**. The connection works as follows:

> **$\nabla_{\mathbf{X}}\mathbf{Y}$ is the rate at which the vector field Y changes relative to a parallel-transported vector when moving into the direction X.**

The connection ∇ fulfills the axioms

$$
\begin{aligned}
&(1) && \nabla_{\mathbf{X}_1+\mathbf{X}_2}\mathbf{Y} \;=\; \nabla_{\mathbf{X}_1}\mathbf{Y} + \nabla_{\mathbf{X}_2}\mathbf{Y} \\
&(2) && \nabla_{\mathbf{X}}(\mathbf{Y}_1+\mathbf{Y}_2) \;=\; \nabla_{\mathbf{X}}\mathbf{Y}_1 + \nabla_{\mathbf{X}}\mathbf{Y}_2 \\
&(3) && \nabla_{f\mathbf{X}}\mathbf{Y} \;=\; f\nabla_{\mathbf{X}}\mathbf{Y} \\
&(4) && \nabla_{\mathbf{X}}(g\mathbf{Y}) \;=\; g\nabla_{\mathbf{X}}\mathbf{Y} + \mathbf{X}(g)\mathbf{Y}\,,
\end{aligned}
$$

where f, g are continuous differentiable functions on the manifold. The first three axioms tell us that ∇ is a bilinear operator. Axiom (4) is also some kind of linearity law, but because of the possible dependence g on the position it looks like a product rule.

If one imagines the vector field **X** as a field directing the orbits of many ships, each carrying a tangential vector **Y** in parallel, then the parallel transport described in the previous section would give rise to a vector field **Y** whose connection in the direction of **X** vanishes:

$$\nabla_{\mathbf{X}}\mathbf{Y} = 0\,. \tag{3.13}$$

The vector field **X** describes *geodesic lines* or short *geodesics* if its own direction does not change when moving along these lines, that is:

$$\nabla_{\mathbf{X}}\mathbf{X} = 0\,. \tag{3.14}$$

3.3.5 Representation of the connection

Let $\{\mathbf{e}_i\}$ be an arbitrary basis vector field on $T\mathcal{M}$. Since the operator ∇ is bilinear, it is fully characterized by its operation on the basis vectors, i.e., the tangent vectors $\nabla_{\mathbf{e}_i}\mathbf{e}_j =: \nabla_i\mathbf{e}_j$ define the connection ∇ completely. As the result is a tangent vector indexed by i, j, it can be represented as a linear combination of the basis vectors, i.e.,

$$\boxed{\nabla_j\mathbf{e}_i = \Gamma^k{}_{ij}\mathbf{e}_k\,.} \tag{3.15}$$

Here the linear coefficients $\Gamma^k{}_{ij}$ are referred to as the *connection coefficients*. These coefficients tell us at which rate the k^{th} component of the basis vector field $\mathbf{e}_i$ changes when moving along the direction $\mathbf{e}_j$.

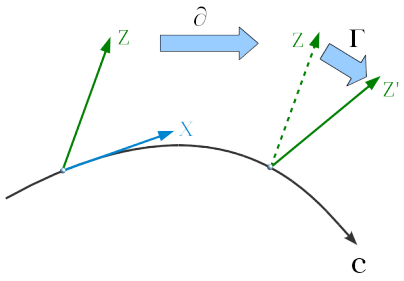
Representing the vector fields introduced in the last subsection in this basis by

$$\mathbf{X} = X^j\mathbf{e}_j\,, \qquad \mathbf{Y} = Y^i\mathbf{e}_i\,, \qquad \nabla_{\mathbf{X}}\mathbf{Y} = [\nabla_{\mathbf{X}}\mathbf{Y}]^k\mathbf{e}_k\,, \tag{3.16}$$

the axioms (3) and (4) lead us to the representation

$$\begin{aligned}[\nabla_{\mathbf{X}}\mathbf{Y}]^k &= \left[\nabla_{(X^i\mathbf{e}_j)}(Y^i\mathbf{e}_i)\right]^k = X^j\left[\nabla_j(Y^i\mathbf{e}_i)\right]^k \\ &= X^j\left[Y^i\nabla_j\mathbf{e}_i + \mathbf{e}_j(Y^i)\mathbf{e}_i\right]^k \\ &= X^jY^i\Gamma^k{}_{ij} + X^j\,\mathbf{e}_j(Y^i)\underbrace{[\mathbf{e}_i]^k}_{=\delta_i^k} = X^j\,\mathbf{e}_j(Y^k) + Y^i\Gamma^k{}_{ij}X^j\end{aligned} \tag{3.17}$$

The first term in the last line contains the basis vector $\mathbf{e}_i$, which can be interpreted as a directional derivative, hence this term describes the directional derivative of the vector component Y^k in the given representation. Since derivatives can be considered as generators of translations, the first term describes a shift of the vector $\mathbf{Y}$ in such a way that it does not change its direction *in the chosen representation*.



Covariant derivative

In the second term, the connection coefficients can be interpreted as a linear map depending on the direction of displacement, which corrects the shifted vector of the selected representation in such a way that 'in reality 'a parallel transport takes place. The entire expression is, as already mentioned, called the *covariant derivative* of the vector field.

3.3.6 Representation of the connection in the coordinate basis

The results of the last section apply to textitevery basis. We now consider the special case of a representation in the coordinate basis $\{\partial_\mu\}$ of a given coordinate system. As mentioned earlier, such bases are special in that the Lie bracket of the directional derivatives vanishes. To indicate that we are dealing here with a coordinate representation, we use Greek indices.

In the coordinate basis we have $\mathbf{e}_\mu = \partial_\mu$ so that Eq. (3.17) takes on the form

$$[\nabla_{\mathbf{X}}\mathbf{Y}]^\alpha = X^\nu\partial_\nu Y^\alpha + Y^\mu\Gamma^\alpha{}_{\mu\nu}X^\nu\,. \tag{3.18}$$

In the coordinate representation in the coefficients $\Gamma^\alpha{}_{\mu\nu}$ are denoted as *Christoffel symbols*. As can be shown by evaluating the Lie-brackets, the Christoffel symbols, in contrast to connection coefficients in a general basis, are symmetrical in the two lower indices:

$$\Gamma^\alpha{}_{\mu\nu} = \Gamma^\alpha{}_{\nu\mu}\,. \tag{3.19}$$

Such a connection is known as *Levi-Civitá connection*.

> **More precisely:** A Levi-Civitá connection is (a) conformal, i.e., the relative angle between two vectors with respect to the metric does not change under parallel transport, and it is (b) torsion-free, i.e., the transported vectors do not rotate helically around the transport direction. The space-time of the general theory of relativity fulfills this property, which manifests itself in a coordinate representation as a symmetry in the two lower indices of the Christoffel symbols.

In differential geometry, the following conventions have become common in the coordinate-based index notation:

$$\begin{array}{lll} Y^{\alpha}{}_{,\nu} & \text{Comma: Directional derivative} & Y^{\alpha}{}_{,\nu} = \partial_\nu(Y^\alpha) \\ Y^{\alpha}{}_{;\nu} & \text{Semicolon: Covariant derivative} & Y^{\alpha}{}_{;\nu} = [\nabla_{\mathbf{e}_\nu}\mathbf{Y}]^\alpha = \partial_\nu(Y^\alpha) + Y^\mu \Gamma^{\alpha}{}_{\mu\nu} \end{array}$$

With these abbreviation Eq. (3.17) can be written in the compact form

$$\boxed{Y^{\alpha}{}_{;\nu} \;=\; Y^{\alpha}{}_{,\nu} + Y^\mu \Gamma^{\alpha}{}_{\mu\nu}} \tag{3.20}$$

3.3.7 Covariant transformation behavior

Why is the covariant derivative called covariant? To understand this, we study how the covariant derivative changes moving from one map to another, i.e., how changes under coordinate changes $\{x^\mu\} \Leftrightarrow \{x^{\mu'}\}$. We first consider the ordinary partial derivative ∂_ν on the map. If this partial derivative acts on a function, we get

$$f_{,\nu} = \frac{\partial}{\partial x^\nu} f(\mathbf{x}) \quad \rightarrow \quad f_{,\nu}{}' = \frac{\partial}{\partial x^{\nu'}} f(\mathbf{x}') \;=\; \frac{\partial x^\rho}{\partial x^{\nu'}} \frac{\partial}{\partial x^\rho} f(\mathbf{x}) = \Lambda_\nu{}^\rho f_{,\rho} \; . \tag{3.21}$$

In other words, the partial derivative $\partial/\partial x^\nu$, *when acting on a function*, transforms like the components of a 1-form, that is, covariant. However, the situation looks quite different when the same partial derivative acts on a vector field $\mathbf{Y}(\mathbf{x})$:

$$\begin{aligned} Y^{\mu}{}_{,\nu} = \frac{\partial}{\partial x^\nu} Y^\mu(\mathbf{x}) \quad \rightarrow \quad \left[Y^{\mu}{}_{,\nu}\right]' &= \frac{\partial x^\rho}{\partial x^{\nu'}} \frac{\partial}{\partial x^\rho} \left(\frac{\partial x^{\mu'}}{\partial x^\tau} Y^\tau(\mathbf{x}') \right) \\ &= \frac{\partial x^\rho}{\partial x^{\nu'}} \frac{\partial x^{\mu'}}{\partial x^\tau} Y^{\tau}{}_{,\rho} + \frac{\partial x^\rho}{\partial x^{\nu'}} \frac{\partial^2 x^{\mu'}}{\partial x^\rho \partial x^\tau} Y^\tau \\ &= \Lambda_\nu{}^\rho \, \Lambda^\mu{}_\tau \, Y^{\tau}{}_{,\rho} + \Lambda_\nu{}^\rho \frac{\partial^2 x^{\mu'}}{\partial x^\rho \partial x^\tau} Y^\tau \end{aligned} \tag{3.22}$$

If only the first term were present, $Y^{\mu}{}_{,\nu}$ would transform like a tensor of rank (1,1). However, the second term violates this transformation behavior, so that the partial derivative of a vector field is *not* a tensor. At this point the mathematics to a certain extent indicates that the partial derivative on curved manifolds is not the correct directional derivative.

In contrast, the covariant derivative $Y^{\alpha}{}_{;\nu}$ transforms correctly as tensor of rank (1,1). There is no need for proof, since the connection ∇ is defined without using a representation and since $Y^{\alpha}{}_{;\nu}$ consists only of the components of $\nabla_{\mathbf{e}_\nu}\mathbf{Y}$. Because of Eq. (3.20) it immediately follows that the Christoffel symbols $\Gamma^{\alpha}{}_{\mu\nu}$ have no tensor property, which is the reason why they are called 'symbols'.

3.3.8 Geodesic lines

Let $c(\lambda)$ be a curve and $\mathbf{u}(\lambda) = \frac{d}{d\lambda} c(\lambda)$ the tangent vector field along this curve. As already discussed, a curve is called a *geodesic line* or simply a *geodesic* if its tangent

vector retains its direction, that is, if the curve describes a motion straight ahead:

$$\boxed{\nabla_{\mathbf{u}}\mathbf{u} = 0.} \tag{3.23}$$

Represented in a coordinate basis $\mathbf{u} = X^\mu \partial_\mu$ this condition reads

$$[\nabla_{u^\mu \partial_\mu}\mathbf{u}]^\alpha \;=\; u^\mu [\nabla_\mu \mathbf{u}]^\alpha \;=\; u^\mu u^\alpha{}_{;\mu} \;=\; 0\,, \tag{3.24}$$

meaning that

$$u^\mu \partial_\mu u^\alpha + u^\mu \Gamma^\alpha{}_{\mu\nu} u^\nu = 0. \tag{3.25}$$

Realizing that $\frac{\mathrm{d}}{\mathrm{d}\lambda} = \frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda}\frac{\partial}{\partial x^\mu} = u^\mu \partial_\mu$ we arrive at

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} u^\alpha + \Gamma^\alpha{}_{\mu\nu} u^\mu u^\nu = 0. \tag{3.26}$$

If the curve is given in coordinates, i.e. $c(\lambda) \leftrightarrow x^1(\lambda), \ldots, x^n(\lambda)$, then it is clear that $u^\alpha = \dot{x}^\alpha$, where the dot stand for the derivative with respect to λ. This leads us to the final result

$$\boxed{\ddot{x}^\alpha + \Gamma^\alpha{}_{\mu\nu} \dot{x}^\mu \dot{x}^\nu = 0\,.} \tag{3.27}$$

This is the so-called *geodesic equation* which describes the trajectory of a free particle in the theory of general relativity.

3.3.9 How the connection is calculated

The above formula allows one to calculate the particle trajectory, provided that the connection ∇ or – in a coordinate representation – the Christoffel symbols are known. In principle, these could be chosen arbitrarily and then describe a specific way how the tangent spaces of the manifold are glued together. In fact, from the mathematical point of view the manifold does not even have to have a metric. However, if there is a metric, then there exists a special connection for which the following principle applies:

A geodesic line between two given points is the curve of minimal length.

Here the notion of the *length* of the curve refers to the selected metric. This principle is valid in the curved space-time of general relativity.

Finding the line of minimal length

For a manifold where this extremal principle is valid, the Christoffel symbols can be calculated explicitly as a function of the metric. For this one uses the variational calculus known from Lagrange mechanics. Accordingly, the length $\int_c \mathrm{d}s$ of the curve c is extremal if it does not change the lowest order under infinitesimal variation of the curve with fixed endpoints, that is

$$\delta \int_c \mathrm{d}s \;=\; 0\,. \tag{3.28}$$

Here the line element is given by $\mathrm{d}s^2 = g_{\mu\nu}\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu$, i.e.,

$$\mathrm{d}s = \sqrt{\left|g_{\mu\nu}\frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda}\frac{\mathrm{d}x^\nu}{\mathrm{d}\lambda}\right|}\,\mathrm{d}\lambda^2 \;=\; \sqrt{|g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu|}\,\mathrm{d}\lambda\,. \tag{3.29}$$

Obviously the line element plays the role of the Lagrange function

$$L(\mathbf{x},\dot{\mathbf{x}}) = \sqrt{|g_{\mu\nu}(\mathbf{x})\dot{x}^\mu\dot{x}^\nu|}$$

with

$$\delta\int_{\lambda_1}^{\lambda_2} L(\mathbf{x},\dot{\mathbf{x}})\,\mathrm{d}\lambda \;=\; 0\,. \tag{3.30}$$

The trajectory is then known to be a solution of the Lagrange equations

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\frac{\partial L}{\partial\dot{x}^\mu} - \frac{\partial L}{\partial x^\mu} = 0\,. \tag{3.31}$$

It should be noted that in the theory of general relativity the metric $\mathbf{g}$ depends on the location of the manifold, and the specific location dependence $\mathbf{g}(\mathbf{x})$ encodes the gravitational field. With some patience (see below) one arrives at the differential equation

$$\ddot{x}^\alpha + \frac{1}{2}g^{\alpha\beta}(g_{\beta\mu,\nu} + g_{\beta\nu,\mu} - g_{\mu\nu,\beta})\dot{x}^\mu\dot{x}^\nu \;=\; 0\,. \tag{3.32}$$

Comparing this result with Eq. (3.27) we realize immediately that the Christoffel symbols are given by

$$\boxed{\Gamma^\alpha{}_{\mu\nu} \;=\; \frac{1}{2}g^{\alpha\beta}(g_{\beta\mu,\nu} + g_{\beta\nu,\mu} - g_{\mu\nu,\beta})}\,. \tag{3.33}$$

So we are now able to calculate the trajectories of particles for any given metric.

Proof: Let us restrict ourselves to time-like curves with $ds \geq 0$ so that we can omit the modules operation, i.e., $L = \sqrt{g_{\mu\nu}(\mathbf{x})\dot{x}^\mu\dot{x}^\nu}$. At first we compute

$$\frac{\partial L}{\partial x^\rho} = \frac{1}{2L}\frac{\partial g_{\mu\nu}}{\partial x^\rho}\dot{x}^\mu\dot{x}^\nu\,, \qquad \frac{\partial L}{\partial\dot{x}^\rho} = \frac{1}{2L}(g_{\rho\nu}\dot{x}^\nu + g_{\mu\rho}\dot{x}^\mu) = \frac{1}{L}g_{\rho\kappa}\dot{x}^\kappa$$

From the second term, we have to find the total derivative of λ, which we express by using the chain rule as

$$\frac{d}{d\lambda}\left[\frac{\partial L}{\partial\dot{x}^\rho}\right] = \frac{\partial}{\partial x^\tau}\left[\frac{\partial L}{\partial\dot{x}^\rho}\right]\frac{dx^\tau}{d\lambda} + \frac{\partial}{\partial\dot{x}^\tau}\left[\frac{\partial L}{\partial\dot{x}^\rho}\right]\frac{d\dot{x}^\tau}{d\lambda} = \frac{\partial^2 L}{\partial x^\tau\partial\dot{x}^\rho}\dot{x}^\tau + \frac{\partial^2 L}{\partial\dot{x}^\tau\partial\dot{x}^\rho}\ddot{x}^\tau\,.$$

The two summands contain the derivatives

$$\begin{aligned}\frac{\partial^2 L}{\partial x^\tau\partial\dot{x}^\rho} &= -\frac{1}{2L^3}\frac{\partial g_{\mu\nu}}{\partial x^\tau}\dot{x}^\mu\dot{x}^\nu g_{\rho\kappa}\dot{x}^\kappa + \frac{1}{L}\frac{\partial g_{\rho\kappa}}{\partial x^\tau}\dot{x}^\kappa\\ \frac{\partial^2 L}{\partial\dot{x}^\tau\partial\dot{x}^\rho} &= -\frac{1}{L^3}g_{\rho\nu}\dot{x}^\nu g_{\tau\mu}\dot{x}^\mu + \frac{1}{L}g_{\rho\tau}\end{aligned}$$

so that the Lagrange equations, multiplied on both sides by $2L$, are given by:

$$\frac{\partial g_{\mu\nu}}{\partial x^\rho}\dot{x}^\mu\dot{x}^\nu \;=\; -\frac{1}{L^2}\frac{\partial g_{\mu\nu}}{\partial x^\tau}\dot{x}^\mu\dot{x}^\nu g_{\rho\kappa}\dot{x}^\kappa\dot{x}^\tau + 2\frac{\partial g_{\rho\kappa}}{\partial x^\tau}\dot{x}^\kappa\dot{x}^\tau - \frac{2}{L^2}g_{\rho\nu}\dot{x}^\nu g_{\tau\mu}\dot{x}^\mu\ddot{x}^\tau + 2g_{\rho\alpha}\ddot{x}^\alpha$$

These rather complicated equations describe the shortest path for any parameterization of the curve. It is now possible to choose a special parameterization so that the equations become simple (similar to choosing a special gauge for the wave equation in electrodynamics).

We want to choose the parameterization in such a way that we move along the curve at a constant velocity, i.e., $ds/d\lambda = const$ which means that $L = const$. It should be noted that we are allowed to choose a particular gauge only *after* the variational calculation has been completed! This causes the first and the third term on the right side to vanish in the above equation. With this gauge the simplified equations read

$$g_{\rho\alpha}\ddot{x}^\alpha + \frac{1}{2}\left(2\frac{\partial g_{\rho\kappa}}{\partial x^\tau}\dot{x}^\kappa\dot{x}^\tau - \frac{\partial g_{\mu\nu}}{\partial x^\rho}\dot{x}^\mu\dot{x}^\nu\right) = 0$$

or

$$\ddot{x}^\alpha + \frac{1}{2}g^{\alpha\rho}\left(2g_{\rho\mu,\nu} - g_{\mu\nu,\rho}\right)\dot{x}^\mu\dot{x}^\nu = 0$$

what can be brought into the desired form. Please keep in mind that the geodesic differential equations generate geodesic lines with a special parameterization designed to move along the curve at constant velocity.

Example: Straight lines in $\mathbb{R}^2$ in polar coordinates

As we have seen before, a two-dimensional plane can be represented by polar coordinates $(x^1, x^2) = (r, \phi)$. In this representation the metric tensor is given by

$$g_{\mu\nu} = \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix}, \qquad g^{\mu\nu} = \begin{pmatrix} 1 & 0 \\ 0 & r^{-2} \end{pmatrix}. \tag{3.34}$$

In this case it is clear that the only non-vanishing partial derivative of the tensor components is $g_{22,1}$. Thus the only non-vanishing Christoffel symbols are

$$\Gamma^1{}_{22} = \frac{1}{2}g^{11}(g_{12,2} + g_{12,2} - g_{22,1}) = -\frac{1}{2}g^{11}g_{22,1} = -r \tag{3.35}$$

$$\Gamma^2{}_{12} = \Gamma^2{}_{21} = \frac{1}{2}g^2 2(g_{21,2} + g_{22,1} - g_{12,2}) = \frac{1}{2}g^{22}g_{22,1} = \frac{1}{r}. \tag{3.36}$$

In this case the equations for the geodesic lines read

$$\ddot{x}^t + \Gamma^1{}_{22}\dot{x}^2\dot{x}^2 = \ddot{r} - r\dot{\phi}^2 = 0 \tag{3.37}$$

$$\ddot{x}^2 + 2\Gamma^2{}_{12}\dot{x}^1\dot{x}^2 = \ddot{\phi} + \frac{2}{r}\dot{r}\dot{\phi} = 0 \tag{3.38}$$

Amusingly, we have obtained two rather complicated differential equations, just to describe a straight line in $\mathbb{R}^2$. This is probably one of the worst choices of coordinates to describe a straight line.

Example: Surface of a sphere S^2

The surface of a sphere $S^2 \in \mathbb{R}^3$ can be parameterized by spherical coordinates by two angles $(x^1, x^2) = (\theta, \phi)$, where the angle θ is measured starting from the z axis, i.e., from the north pole. The metric tensor reads

$$g_{\mu\nu} = \begin{pmatrix} 1 & 0 \\ 0 & \sin^2\theta \end{pmatrix}, \qquad g^{\mu\nu} = \begin{pmatrix} 1 & 0 \\ 0 & \sin^{-2}\theta \end{pmatrix}. \tag{3.39}$$

Again the only non-vanishing partial derivative of the tensor components is $g_{22,1}$. Consequently the non-vanishing Christoffel symbols are the following:

$$\Gamma^1{}_{22} = \frac{1}{2}g^{11}(g_{12,2} + g_{12,2} - g_{22,1}) = -\frac{1}{2}g^{11}g_{22,1} = -\sin\theta\cos\theta \quad (3.40)$$

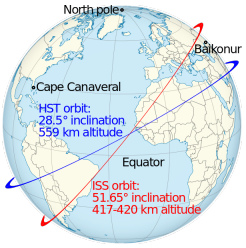
$$\Gamma^2{}_{12} = \Gamma^2{}_{21} = \frac{1}{2}g^2 2(g_{21,2} + g_{22,1} - g_{12,2}) = \frac{1}{2}g^{22}g_{22,1} = \cot\theta\,. \quad (3.41)$$

The equations for a geodesic line are in this case

$$\ddot{x}^t + \Gamma^1{}_{22}\dot{x}^2\dot{x}^2 = \ddot{\theta} - \dot{\phi}^2\sin\theta\cos\theta = 0 \quad (3.42)$$

$$\ddot{x}^2 + 2\Gamma^2{}_{12}\dot{x}^1\dot{x}^2 = \ddot{\phi} - 2\dot{\phi}\dot{\theta}\cot\theta = 0 \quad (3.43)$$

They describe large circles on the sphere, which, however, can be 'tilted' relative to the equatorial plane and are therefore represented with an oscillating θ component, just like the trajectory of the ISS.



ISS trajectory

3.3.10 Covariant derivative of arbitrary tensor fields

The covariant derivative introduced above acts on vector fields and generates the parallel transport of vectors. We now want to generalize the covariant derivative to tensors of any order.

Covariant derivative of functions

The parallel transport of a scalar does not affect the value of a scalar. Therefore the covariant derivative acting on a scalar function (0-form) is identical to the usual directional derivative of a scalar:

$$\nabla_{\mathbf{X}} f = \mathbf{X}(f) \quad (3.44)$$

In components we therefore have $\nabla_\mu f = \partial_\mu f$ bzw. $f_{;\mu} = f_{,\mu}$.

Covariant derivatives of 1-forms

We now consider a field of 1-forms $\alpha(\mathbf{x})$ acting on a vector field $\mathbf{Y}(\mathbf{x})$. At each point of the manifold, $\alpha(\mathbf{Y})$ returns a number, that is, a function on $\mathcal{M}$. Moving from one of these points in the direction $\mathbf{X}$, the function value will gradually change, with the rate of change being given by the usual directional derivative $\mathbf{X}(\alpha(\mathbf{Y})) = \nabla_{\mathbf{X}}(\alpha(\mathbf{Y}))$. This changing output value of the 1-form may have two causes, namely it may be caused

(a) by a change of the vector field $\mathbf{Y}$ or

(b) by a change of the 1-form α.

Tis gives two terms:

$$\nabla_{\mathbf{X}}(\alpha(\mathbf{Y})) = \alpha(\nabla_{\mathbf{X}}\mathbf{Y}) + (\nabla_{\mathbf{x}}\alpha)(\mathbf{Y})\,. \quad (3.45)$$

This identity defines the covariant derivative $\nabla_{\mathbf{X}}\alpha$ acting on a 1-form in a completely representation-free manner:

$$(\nabla_{\mathbf{X}}\alpha)(\mathbf{Y}) \;=\; \nabla_{\mathbf{X}}(\alpha(\mathbf{Y})) - \alpha(\nabla_{\mathbf{X}}\mathbf{Y})\,. \tag{3.46}$$

Resorting to the representation in terms of an arbitrary basis vector field $\{\mathbf{e}_i\}$ of $T\mathcal{M}$ and the corresponding dual basis vector field $\{\mathbf{e}^j\}$ of $T^*\mathcal{M}$, the equations given above imply that

$$(\nabla_j \mathbf{e}^k)(\mathbf{e}_i) \;=\; \nabla_j\Big(\underbrace{\mathbf{e}^k(\mathbf{e}_i)}_{=\delta^k_i}\Big) \;-\; \mathbf{e}^k(\nabla_j \mathbf{e}_i) \;=\; -\Gamma^k{}_{ij}\,, \tag{3.47}$$

where $\nabla_j = \nabla_{\mathbf{e}_j}$ and where the first term on the right hand side, which is the derivative of a constant, vanishes. This implies that $\nabla_j \mathbf{e}^k = -\Gamma^k{}_{ij}\mathbf{e}^i$ so that the covariant derivative acting on an arbitrary 1-form is given in this representation by

$$\nabla_j \alpha \;=\; \mathbf{e}_j(\alpha_k)\mathbf{e}^k - \alpha_k \Gamma^k{}_{ij}\mathbf{e}^i \tag{3.48}$$

or

$$\alpha_{i;j} \;=\; e_j(\alpha_i) - \alpha_k \Gamma^k{}_{ij} \tag{3.49}$$

or, in a coordinate basis

$$\boxed{\alpha_{\mu;\nu} \;=\; \alpha_{\mu,\nu} - \alpha_\rho \Gamma^\rho{}_{\mu\nu}}\,. \tag{3.50}$$

Covariant derivative of arbitrary tensor fields

For arbitrary tensor fields of higher rank, which can be written as a tensor product $\mathbf{T} = \mathbf{A} \otimes \mathbf{B}$, we can apply the following product rule for the covariant derivative:

$$\boxed{\nabla_{\mathbf{X}}(\mathbf{A}\otimes\mathbf{B}) \;=\; (\nabla_{\mathbf{X}}\mathbf{A})\otimes\mathbf{B} + \mathbf{A}\otimes(\nabla_{\mathbf{X}}\mathbf{B})\,.} \tag{3.51}$$

As an example, consider a covariant tensor field $\mathbf{T}$ of rank 2 represented in a given basis $\{\mathbf{e}_i\}$ by $\mathbf{T} = T_{ij}\mathbf{e}^i \otimes \mathbf{e}^j$. With the product rule given above, the covariant derivative can be calculated easily. It should be noted that the components T_{ij} of a tensor field depend on the location on the manifold, meaning that we have to take the derivative as if they were functions. All in all we get three terms:

$$\nabla_k \mathbf{T} \;=\; \nabla_k(T_{ij}\mathbf{e}^i\otimes\mathbf{e}^j) \;=\; (\nabla_k T_{ij})\mathbf{e}^i\otimes\mathbf{e}^j + T_{ij}(\nabla_k\mathbf{e}^i)\otimes\mathbf{e}^j + T_{ij}\mathbf{e}^i\otimes(\nabla_k\mathbf{e}^j) \tag{3.52}$$

or, in components

$$T_{ij;k} = \mathbf{e}_k(T_{ij}) - T_{mj}\Gamma^m{}_{ik} - T_{im}\Gamma^m{}_{jk}\,. \tag{3.53}$$

As can be seen, each index of the tensor is corrected by its own additive term, i.e., **each index is transformed separately by Christoffel symbols.** One can also compute the derivative of mixed tensors in this way:

$$T^{i_1\ldots i_q}_{j_1\ldots j_p;k} \;=\; \mathbf{e}_k(T^{i_1\ldots i_q}_{j_1\ldots j_p}) + \sum_{n=1}^{q} T^{i_1\ldots m\ldots i_q}_{j_1\ldots\ldots j_p}\Gamma^{i_n}{}_{mk} - \sum_{n=1}^{p} T^{i_1\ldots\ldots i_q}_{j_1\ldots m\ldots j_p}\Gamma^m{}_{j_n k} \tag{3.54}$$

Here, the rightmost index of the Christoffel symbols is always the index by which the derivative is taken, meaning that it is connected to the vector specifying the direction. Note that contravariant indices have positive corrections terms while covariant indices negative ones.

Covariant derivative of the metric

It is of course also possible to apply Eq. (3.53) to the metric tensor which is a symmetric tensor of rank 2. In a given coordinate basis one obtains

$$g_{\mu\nu;\tau} \;=\; g_{\mu\nu,\tau} - g_{\rho\nu}\Gamma^{\rho}{}_{\mu\tau} - g_{\mu\rho}\Gamma^{\rho}{}_{\nu\tau}\,. \tag{3.55}$$

In the theory of general relativity, the metric tensor has a special meaning. To understand that, let us again go back to the example of asailor. If the captain takes two vectors $\mathbf{X}, \mathbf{Y}$ on board and transports them along an arbitrary path, we expect that the angle between the two vectors will not change during the trip, meaning that $\mathbf{g}(\mathbf{X}, \mathbf{Y})$ is preserved while traveling. It can be shown that there is exactly one connection that satisfies this property, which is why it is called *metric* connection. Such a metric connection fulfills the property

$$\nabla_{\mathbf{X}}\mathbf{g} = 0 \;\; \forall \mathbf{X} \qquad \text{or} \qquad g_{\mu\nu;\tau} = 0. \tag{3.56}$$

The connection used in general relativity is of that kind. It can be shown that such a metric connection is always torsion-free.

3.3.11 Exterior derivative of the tensorial forms*

In Sect. 2.4.2 on page 69 we introduced the *exterior derivative* of differential forms. It differs from an ordinary derivative by a downstream antisymmetrization, ensuring that the operation is part of the exterior algebra.

The exterior derivative acting on pure p-forms works exactly as discussed in Sect. 2.4.2 on page 69. So nothing changes here. The situation is different with tensor-valued forms, which we briefly addressed in Sect. 2.6 on page 76. Such tensor-valued p-forms have p inputs that are antisymmetrized and satisfy the computational rules of outer algebra, as well as a number of vectorial outputs that need not be antisymmetrized. For example, a vector field is vector-valued 0-form. Usually the non-antisymmetrized components of the outputs are indexed, while the p antisymmetrized inputs are treated as representation-free differential forms without indices.

... Will be continued...

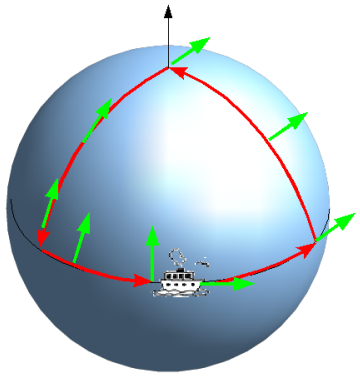


Figure 3.10: Curvature is expressed by the fact that the sum of the angles in a triangle is not equal to 180°. The captain of a ship can determine the deviation using parallel transport. If, as shown in the figure above, the ship sails along a three-quarter closed path (red) and transports a tangential vector (green), this vector points in a direction rotated by 90° at the destination. This rotation angle is exactly the same as the deviation of the angular sum from 90°. Note that such a measurement is also possible on abstract (not necessarily embedded) manifolds.

3.4 Curvature

3.4.1 Riemann curvature tensor

The procedure described above makes it possible to quantify the curvature of the manifold in the enclosed area. Mathematically, this process can be described as follows: Take two linearly independent vector fields $\mathbf{X}, \mathbf{Y}$ and move first along a geodesic line in $\mathbf{Y}$-direction and then along another geodesic line in $\mathbf{X}$-direction. Then repeat the process in reverse order (see Fig. 3.11). If the destinations happen to be different, then we have to connect the remaining gap in order to close the path. This difference is given to lowest order by the *Lie-bracket* $[\mathbf{X}, \mathbf{Y}]$, see Sect. 2.4.7 on page 73. Along these two contours a tangential vector $\mathbf{Z}$ has to be transported in parallel according to the usual protocol and finally we have to compare the results at the destination.

Mathematically, this procedure is expressed by alternately applying the corresponding covariant derivatives [3]. The curvature is thus described by the commutator-like map

$$\boxed{\mathbf{R}(\mathbf{X},\mathbf{Y})\mathbf{Z} = \nabla_{\mathbf{X}}\nabla_{\mathbf{Y}}\mathbf{Z} - \nabla_{\mathbf{Y}}\nabla_{\mathbf{X}}\mathbf{Z} - \nabla_{[\mathbf{X},\mathbf{Y}]}\mathbf{Z}\,,} \tag{3.57}$$

which is known as the *Riemann curvature tensor*. It is a tensor of rank $(1,3)$ that has three input vectors as arguments and renders a vector as output. The inputs $\mathbf{X}$ and $\mathbf{Y}$ define a locally closed contour while $\mathbf{Z}$ is the vector to be transported. The form returns the mismatch of $\mathbf{Z}$ when transported along the contour.

[3] In a coordinate basis, the last term is omitted, since in this case the Lie bracket vanishes.

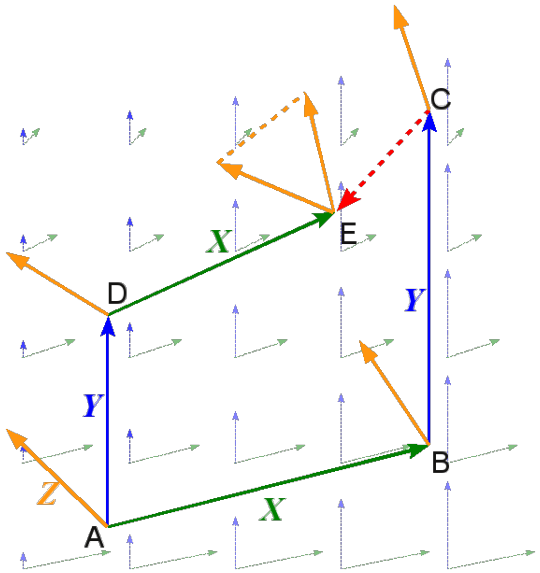


Figure 3.11: Construction of the Riemannian curvature tensor. A vector $\mathbf{Z}$ (orange) is transported in parallel in two different ways, namely by applying $\nabla_{\mathbf{X}} \circ \nabla_{\mathbf{Y}}$ along ADE and by using $\nabla_{\mathbf{Y}} \circ \nabla_{\mathbf{X}}$ along ABC. In addition, if the destinations do not match, the vector must be moved in parallel along the red dashed vector from C to E by using the Lie bracket $\nabla_{[\mathbf{X},\mathbf{Y}]}$. At the destination E, both vectors are compared. The difference gives information about how much the manifold is curved in the area enclosed by the path.

3.4.2 Representation of the Riemannian curvature tensor

In a given basis, the curvature tensor can be represented as a four-component quantity $R^{\mu}{}_{\nu\alpha\beta}$. These components are given by

$$\mathbf{e}_\mu R^{\mu}{}_{\nu\alpha\beta} \;=\; \big([\nabla_\alpha, \nabla_\beta] - \nabla_{[\mathbf{e}_\alpha, \mathbf{e}_\beta]}\big)\mathbf{e}_\nu \;=\; \big(\nabla_\alpha \nabla_\beta - \nabla_\beta \nabla_\alpha - c^{\rho}{}_{\alpha\beta}\nabla_\rho\big)\mathbf{e}_\nu \;, \tag{3.58}$$

where $c^{\rho}{}_{\alpha\beta}$ are the *structural coefficients*. By inserting the covariant derivative and comparing the coefficients one arrives at

$$\boxed{R^{\mu}{}_{\nu\alpha\beta} \;=\; \Gamma^{\mu}{}_{\nu\beta,\alpha} - \Gamma^{\mu}{}_{\nu\alpha,\beta} + \Gamma^{\rho}{}_{\nu\beta}\Gamma^{\mu}{}_{\rho\alpha} - \Gamma^{\rho}{}_{\nu\alpha}\Gamma^{\mu}{}_{\rho\beta} - c^{\rho}{}_{\alpha\beta}\Gamma^{\mu}{}_{\nu\rho}\,.} \tag{3.59}$$

In coordinate bases, the last term originating from the Lie derivative vanishes. Since the Christoffel symbols depend on the metric via Eq. (3.33), the Riemann curvature tensor can be calculated in a straight-forward way for a given metric. However, as we will see, not all of its $4^4 = 256$ components are independent.

3.4.3 Symmetries of the curvature tensor

The curvature tensor shown in components obeys the following symmetries. We use the compact notation with square brackets, which should symbolize a sum over the cyclic permutations of the indices contained therein.

- First Bianchi identity:

$$R^{\mu}{}_{[\nu\alpha\beta]} \;=\; R^{\mu}{}_{\nu\alpha\beta} + R^{\mu}{}_{\beta\nu\alpha} + R^{\mu}{}_{\alpha\beta\nu} \;=\; 0 \tag{3.60}$$

- Second Bianchi identity:

$$R^{\mu}{}_{\nu[\alpha\beta;\gamma]} \;=\; R^{\mu}{}_{\nu\alpha\beta;\gamma} + R^{\mu}{}_{\nu\gamma\alpha;\beta} + R^{\mu}{}_{\nu\beta\gamma;\alpha} \;=\; 0 \tag{3.61}$$

- Anti-symmetry in the first two indices:

$$R_{\mu\nu\alpha\beta} = -R_{\nu\mu\alpha\beta} \tag{3.62}$$

- Symmetry under the exchange of both pairs of indices:

$$R_{\mu\nu\alpha\beta} = R_{\alpha\beta\mu\nu} \tag{3.63}$$

Here $R_{\mu\nu\alpha\beta} = g_{\mu\rho}R^{\rho}{}_{\nu\alpha\beta}$. The last two symmetries quoted above imply the anti-symmetry of the third and the fourth index:

$$R_{\mu\nu\alpha\beta} = -R_{\mu\nu\beta\alpha}\,. \tag{3.64}$$

Because of these symmetries, the number of independent components of the Riemann curvature tensor is reduced as follows:

dimension	1	2	3	4
number of components	1	16	81	256
independent components	0	1	6	20

3.4.4 Ricci tensor

Which physically relevant tensors can be generated by contraction from the curvature tensor? If you contract the first two indices, you get zero because of the antisymmetry. The same applies to a contraction of the indices 3-4. The only contractions that do not give zero are 1-3, 1-4, 2-3, and 2-4, and these four contractions are identical up to a minus signs due to the antisymmetry. In the literature it is customary to contract indices 1-3. The result is the so-called *Ricci tensor*

$$\boxed{R_{\mu\nu} \;:=\; R^{\rho}{}_{\mu\rho\nu}.} \tag{3.65}$$

This tensor is the only possible non-trivial contraction of the curvature tensor. In order to distinguish it from the Riemann curvature tensor $\mathbf{R}$ in the representation-free notation, it is also denoted as 'Ric ', i.e.,

$$\mathbf{Ric} = R_{\mu\nu}\,\mathrm{d}x^{\mu} \otimes \mathrm{d}x^{\nu}\,. \tag{3.66}$$

The Ricci Tensor can be contracted further to a *curvature scalar*

$$R = R^{\mu}{}_{\mu} \,. \tag{3.67}$$

As we will see below, this scalar plays an important role in the action integral of the gravitational field.

3.4.5 Interpretation of curvature tensors

Interpretation of the Riemann curvature tensor

You can always arrange coordinates in such a way that the metric tensor assumes a certain matrix representation in a particular point of the manifold. In particular, one can choose the coordinates such that $g_{\mu\nu} = \eta_{\mu\nu}$ in one particular point, so that the metric tensor takes the form of a flat Minkowski metric (just like our captain who can always draw a local coordinate system with Euclidean coordinates on the sea surface at his current location). In general relativity, such a coordinate system corresponds to the natural Minkowski coordinates, that is, a local inertial reference frame that would be used by a free-falling astronaut.

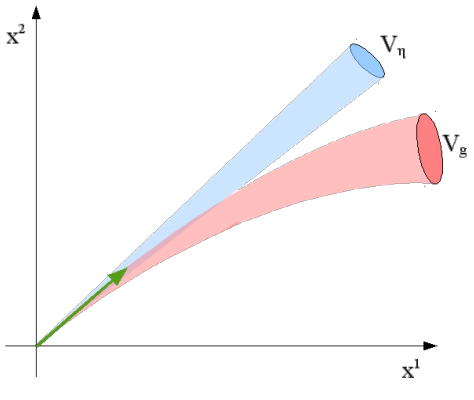
Having chosen a representation in such a way that the metric tensor at the origin $x^{\mu} = 0$ of the coordinate system is equal to the Minkowski metric, one can ask oneself how the components of this tensor vary to lowest order in the immediate vicinity of the origin. A calculation (here without proof) shows that the lowest-order corrections are *quadratic* and that they are described by Riemann's curvature tensor:

$$g_{\mu\nu}(\mathbf{x}) \;=\; \eta_{\mu\nu} \;+\; \frac{1}{3} R_{\mu\alpha\nu\beta}\, x^{\alpha} x^{\beta} + \mathcal{O}(|\mathbf{x}|^3) \,. \tag{3.68}$$

The Riemann curvature tensor thus describes how the metric changes to lowest order when going in a certain direction.

Interpretation of the Ricci tensor

To interpret the Ricci tensor, we now imagine a narrow cone (Schultüte) of geodesic lines starting at the origin and going into a particular direction. In a Minkowski metric, this cone will expand in a certain way with increasing distance and span a volume element $V_{\eta}(\mathbf{x})$. In a curved geometry, on the other hand, we obtain a volume element $V_g(\mathbf{x})$ which differs from $V_{\eta}(\mathbf{x})$. In the vicinity of the origin we expect this difference to be infinitesimally small. It turns out that the corresponding corrections are just given by the Ricci tensor:



Ricci tensor

$$V_g(\mathbf{x}) \;=\; \left(1 - \frac{1}{6} R_{\mu\nu} x^{\mu} x^{\nu} + \mathcal{O}(x^3)\right) V_{\eta}(\mathbf{x}) \tag{3.69}$$

4 Electrodynamics as a gauge theory

The general theory of relativity belongs to the group of so-called *gauge theories*. Not only general relativity, but basically all quantum field theories, in particular the standard model of electroweak interactions and QCD, are gauge theories. It seems that gauge theories deeply reflect the way how Nature is made.

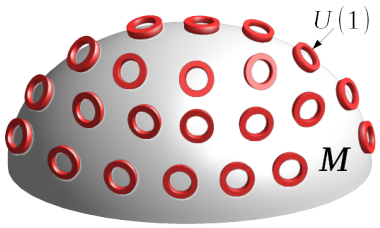
All gauge theories have in common that they are based on a certain *gauge group*. In general relativity, the gauge group is just the Lorentz group. All gauge theories are designed in a similar way. To get in touch with this principle we first consider the simplest of all gauge theories, namely, *electrodynamics*. Electrodynamics is a gauge theory based on the symmetry group $U(1)$.

Gauge symmetry group	Theory
$U(1)$	Electrodynamics
$SU(2)$	Weak interactions (spin)
$U(1) \times SU(2)$	unified electroweak QFT
$SU(3)$	Strong interactions (QCD)
Lorentz	General relativity

4.1 U(1) gauge theory

4.1.1 Intrinsic degrees of freedom

Apparently, the space-time realized in Nature is not just a locally 3+1-dimensional manifold but has a much more complicated structure. Namely, in addition to the spatiotemporal degrees of freedom, in which we can move forth and back, there are also certain *intrinsic degrees of freedom*, which can be thought of as small 'rolled-up dimensions' that are 'attached' at any point in spacetime. Although as a human being one cannot move in concrete terms in these compactified dimensions, as one is used to move in space-time dimensions, the physical effects of these intrinsic degrees of freedom can be experienced indirectly in the form of physical force fields.



Intrinsic degrees of freedom: Each point of the manifold $\mathcal{M}$ hosts an additional internal space, which in the case of electrodynamics is just a circle.

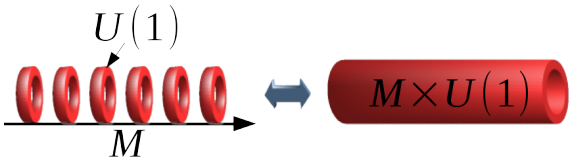
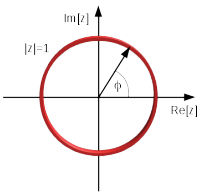


Figure 4.1: The circles form a continuum, as can be visualized in 1D, where the resulting topology is that of a cylinder.

Just as the 3+1-dimensional space-time possesses a certain structure, which can be described by the symmetry group of its tangent space (namely, the Lorentz- or Poincaré group), the intrinsic degrees of freedom are also characterized by a certain symmetry group. The simplest example of such a compactified dimension is a circle. Its symmetry group is the so-called *circle group* of the translations along the circle. Unlike translations in $\mathbb{R}$, which allow you to move as far as you like, on a circle you will eventually come back to the starting point. In mathematical terms, this means that the circle group is *compact*. This is common to all intrinsic degrees of freedom: they are *compactified*.

> **Remark:** Symmetries are the fundamental origin for the existence of any structure in Nature. Symmetries do not give freedom, they rather restrict freedom. Without symmetries, quantum physics would spread over the entire state space and create a featureless mixture of maximal entropy. With symmetries, however, strict rules in the form of conservation laws come into play. This is the reason why macroscopically observable properties of objects are always related to certain symmetries. For example, without translation invariance, the notions of positions and momenta would not exist. Theoretical physics therefore always starts with the identification of the underlying symmetry groups.

There are a variety of ways to represent translations on a circle. For example, the complex numbers $z \in \mathbb{C}$ with $|z| = 1$ form a circle in the complex plane. Translations along this circle can be expressed by multiplying z by a complex phase $e^{i\phi}$, where $\phi \in [0, 2\pi)$. Since such translations are formally unitary (=norm-preserving) and since the circle is one-dimensional, the circle group is denoted by $U(1)$, the group of *unitary transformations in one dimension*.



$U(1)$ in the complex plane.

> **Remark:** The term $U(n)$ stands for "*unitary* transformation in n dimensions" while $SU(n)$ stands for "*special* unitary transformations in n dimensions". These groups are the complex counterparts to the orthogonal groups $O(n)$ and $SO(n)$ of rotations in real-valued vector spaces with and without reflections, respectively. The group elements of $U(n)$ can be thought of as complex-valued rotations in $\mathbb{C}^n$ which preserve the standard scalar product. The group $U(1)$ accordingly describes norm-preserving rotations of complex numbers in the complex plane, and therefore it is isomorphic to the circle group. Since this group has only one generator, it is a commutative group.

We now consider a relativistic theory with an intrinsic $U(1)$ symmetry, neglecting gravitational effects. In such a theory, the space-time is a flat Minkowski space $\mathbb{R}^{3+1}$ with a circle attached at each point. In order to gain some intuitive understanding, let us first imagine space-time as one-dimensional, as if there were only the x-axis (see Fig. 4.1). We also want to think of this axis as a *discrete* sequence of points. In each of these points we attach a circle as a compactified intrinsic degree of freedom. Now the position of a particle is characterized not only by its spatial location on the x axis, but also by its corresponding position on the circle.

> **Remark:** Mathematically speaking, the internal space can be described in terms of a *fiber*

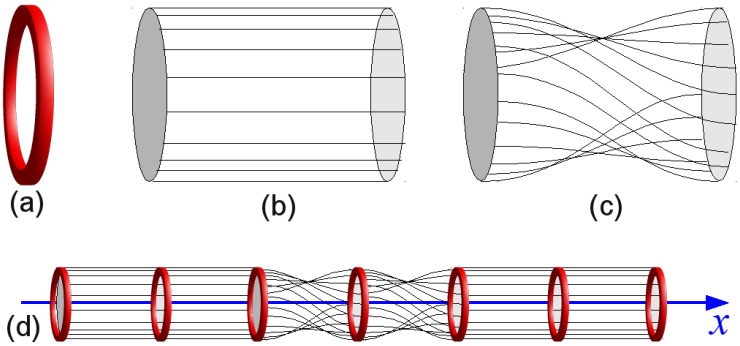


Figure 4.2: Building blocks of electrodynamics. (a) At each point of space-time, we attach a compactified one-dimensional circle. (b)-(c) The circles of neighboring points of space-time are glued by connecting elements. These can be 'straight' or' twisted'. (d) If the spacetime were one-dimensional, the gluing would give us a torus with locally varying twists.

bundle attached to the base space (manifold). This construction comes with a projection which maps each circle back to the space-time point p in which it is attached.

In order to move along the x axis, it is necessary to links these circles by connecting elements. We can think of these connections as some kind of tubes which can be used to move from one circle to the next. Roughly speaking, these connections define what it means to move straight forward in the internal space, telling us how a geodesic would look like in the tube. The connecting elements can either be straight (Fig. 4.2b) or twisted (Fig. 4.2c). For example, if all circles were connected by straight tubes, particles resting at the 'lowest point' of the circle would remain there when moving through space-time. In this case a macroscopic observer would not be able to notice the existence of the small circles. Real physical effects come about only in cases where the connecting tubes are twisted.

Remark: Within the framework of classical physics, the spacetime does of course not consist of equidistant points but of a continuum of points. The space shown in Fig. 4.2d is in fact not a sequence of circles but a continuous torus $\mathbb{R} \oplus S^1$. However, the discretized version is used here for the following reasons:

1. The tubes illustrate the mathematical concept of a so-called *connection*.
2. While we have no means to imagine the actual space $\mathbb{R}^{3+1} \oplus S^1$, the discretization at least partially illustrates the space $\mathbb{R}^2 \oplus S^1$ (see below).
3. In computer simulations of quantum field theories, so-called *lattice field theories* such as lattice-QCD, the continuum is actually discretized in the same way and useful numerical predictions can be made on this basis.
4. Many physicists believe that on the Planck scale of 10^{-35}m, the structure of space-time may exhibit some unknown microstructure that could amount to an effective discretization. One possible speculative approach is *quantum loop gravity*, as described, for example, in the book by Carlo Rovelli [**?**].

A straight connection (untwisted tube) is just an identical map from one circle to the next, while a twisted connection causes an additional translation along the circle. Thus the connecting elements themselves are nothing but symmetry transformations and thus they are group elements of the symmetry group $U(1)$. Here we already recognize a fundamental construction principle of gauge theories:

The intrinsic degrees of freedom are spatially connected by a group elements of the corresponding symmetry group.

However, as we are not really dealing with a discrete locations but with a continuum of points with circles attached to it, we can assume that the connections bridge only an infinitesimal distance. Thus, under normal circumstances we expect that connections between circles at an infinitesimal distance $\mathrm{d}x$ only exhibit *infinitesimal* twists. This means that the connecting elements $u \in U(1)$ are infinitesimal transformations, which differ only slightly from the identity and can therefore be expanded to first-order by

$$u = \mathbb{1} + \Sigma \, \mathrm{d}x \tag{4.1}$$

The quantity Σ is the generator of this infinitesimal transformation and as such it is an element of the *Lie algebra* of the symmetry group. More precisely, Σ is a Lie-algebra-valued field of 1-forms.

The Lie-Algebra $u(1)$ of the Lie group $U(1)$ is extremely simple because it is commutative and possesses only a single generator σ. This generator obeys the relation $\sigma^2 = -1$ and is most easily represented in the complex plane by the imaginary unit $\sigma = i$, in which case the corresponding group elements are represented by $u = e^{i\phi}$. Hence we can express Σ in this particular case as

$$\Sigma = \sigma \mathbf{A} \,, \tag{4.2}$$

where $\mathbf{A}$ is an ordinary (real-valued) 1-form which returns the twist rate of the $U(1)$-connections when moving in a particular direction.

4.1.2 Representation of intrinsic degrees of freedom

In order to do explicit calculations with the intrinsic degrees of freedom, it is again necessary to choose a suitable representation. It is customary to represent the $U(1)$ as a unit circle in the complex plane $z \in \mathbb{C}$ with $|z| = 1$, where the generator is represented by $\sigma = i$. The space $\mathbb{R} \oplus S^1$ considered here is thus represented by two coordinates x, z. In the definition of the coordinate system, however, one has the freedom to arbitrarily set the origin of the coordinate system $z = 1$ on each of the circles. This freedom is referred to as *gauge freedom*, so a gauge is nothing but a special choice of the coordinate system in the intrinsic space.

- **A gauge is the choice of a coordinate system in the intrinsic space.**

Once a particular coordinate system is selected, a particle located at a particular position in spacetime and on the circle, described by the coordinates $(\mathbf{x}, z)$, can be "transported in parallel" to the neighboring circle at $\mathbf{x}' = \mathbf{x} + \mathrm{d}\mathbf{x}$, where it then has the intrinsic coordinate

$$z_{\parallel}(x + \mathrm{d}x) = u(\mathbf{x}, \mathrm{d}\mathbf{x}) z \,. \tag{4.3}$$

The group element $u \in U(1)$ is also represented as a complex number on the unit circle. Since, according to Eq. (4.1), this transformation differs only slightly from the identity

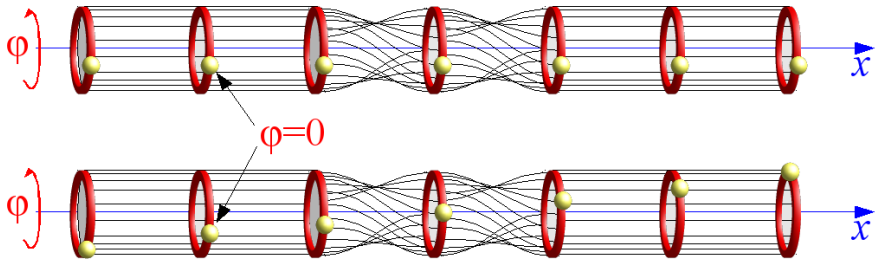


Figure 4.3: Gauge freedom in choosing the coordinate system for the intrinsic degrees of freedom. The yellow dots mark the positions on the circle where $\varphi = 0$ or $z = 1$, so to speak, marking the origin of the local coordinate system.

for infinitesimal displacements $\mathrm{d}x$, it will be given to lowest order by

$$u(\mathbf{x}, \mathrm{d}\mathbf{x}) = \mathbb{1} + i\mathbf{A}(\mathbf{x})\ \mathrm{d}\mathbf{x} = 1 + iA_\mu(\mathbf{x})\,\mathrm{d}x\,, \tag{4.4}$$

meaning that $z_\parallel(x + \mathrm{d}x) = \big(1 + i\,A_\mu(x)\ \mathrm{d}x\big)z(x)$. Here $u(x)$ is a representation of u and $iA(x)$ is a representation of Δ, that is, it maps the space coordinate onto the chosen *representation* of the Lie algebra of the symmetry group $U(1)$. The function $A(x)$ is called the *gauge field*. It quantifies the twist of the connecting elements in a given direction of space-time.

Let's assume that this zero point is always 'in the same place" as shown in the upper part of Fig. 4.3. In this case, 'straight'connections are represented by a vanishing gauge field $A(x) = 0$. However, you could also choose the zero points differently, as shown in the lower part of the figure. In this case a 'straight' connection would be described by a non-vanishing gauge field. It would even be possible to select the coordinates in such a way that the transformations caused by twisted connections are compensated by the coordinate representation, so that the gauge field would be equal to zero despite the actual rotation along the $U(1)$ circle. The representation of a gauge field thus results from both 'true' twists of the connections and 'apparent' twists coming from the choice of coordinates.

> **Remark:** To measure distances in space and time, we need man-made units such as meters and seconds. But do we also need new units for measuring distances in the intrinsic space? The answer is: textit no. In space-time, units are necessary only because these degrees of freedom are infinitely extended (non-compact) and thus do not characterize a natural length scale by themselves, neither on a small scale nor on a large scale. In contrast, the intrinsic degrees of freedom are compactified and thus provide a natural unit. For example, in the case of the circle, the distance along the whole circle is just 2π.

4.1.3 Gauge transformations

A *gauge transformation* is a coordinate transformation in the intrinsic spaces, i.e. a change of the representation that has no influence on the actual physics. As we can define the origin of the coordinate system on each of the circles independently, it is clear that each intrinsic space may be transformed differently. Thus, in the case of the $U(1)$ gauge theory one can imagine a gauge transformation as a location-dependent shift of the coordinate system along the circles, for example as a change from the upper

to the lower situation shown in Fig. 4.3. Such a transformation can be written as

$$z(x) \to \tilde{z}(x) = e^{if(x)} z(x)\,. \tag{4.5}$$

Here $e^{if(x)}$ is the (location-dependent) complex phase that shifts the origin of the coordinate system on the circles (but not the circles themselves). It can easily be shown that under this coordinate transformation the gauge field changes according to

$$A(x) \to \tilde{A}(x) = A(x) + \frac{d}{dx} f(x)\,. \tag{4.6}$$

Gauge fields which differ by such a transformation describe different representations but they are physically equivalent.

Proof: Because of Eq. (4.5) one has

$$z_{\parallel}(x+dx) = (1 + iA(x)dx)\, z(x)\,, \qquad \tilde{z}_{\parallel}(x+dx) = (1 + i\tilde{A}(x)dx)\, \tilde{z}(x),$$

and therefore

$$\begin{aligned} e^{if(x+dx)} z_{\parallel}(x+dx) &= (1 + i\tilde{A}(x)dx)\, e^{if(x)} z(x) \\ \Rightarrow e^{if(x)}\,(1 + if'(x)dx + \ldots)\,(1 + iA(x)\,dx)\, z(x) &= (1 + i\tilde{A}(x)dx)\, e^{if(x)} z(x) \end{aligned}$$

If we compare the expanded terms to first order we arrive at the statement given above.

In one dimension the situation is particularly trivial because for any pair of functions $A(x), \tilde{A}(x)$, by simple integration, one can find an appropriate $f(x)$ such that both gauge fields are physically equivalent. As a result, all possible gauges are physically equivalent and therefore have no physical significance. This is also plausible, because a beetle that lives on the torous without an embedding space has no possibility to detect a twist of the connecting elements.

4.1.4 Two-dimensional U(1) gauge theory

In higher dimensions, however, the situation is different. As an example let us consider a two-dimensional plane $\mathbb{R}^2$ with an intrinsic $U(1)$ symmetry. Although one cannot embed the space $\mathbb{R}^2 \oplus S^1$ in the $\mathbb{R}^3$ and therefore cannot imagine the space vividly, we can at least gain some intuition by studying a discretized variant. Fig. 4.4 sketches such a two-dimensional space, in which the intrinsic spaces are symbolized by planar circles, which are mutually connected with connecting elements in both spatial directions. Again, keep in mind that 'in reality' we are dealing here with a continuum of connections without a square lattice structure.

Unlike in the previously discussed 1D case, it is now possible to move in arbitrary directions in $\mathbb{R}^2$. A beetle living in this two-dimensional tunnel system now has the opportunity to measure distortions, even without being able to see them from the outside. For this purpose, the beetle has to wander along a closed path (a so-called loop). Arriving at the destination the beetle can compare its position on the circle with the previous position, detecting a possible physical net twist of the connections along the path. For example, if the beetle walks along the path A-B-E-D-A in Fig. 4.4, it will detect a twist, but if it moves along B-C-F-E-B, it will find no distortion, because the effects of

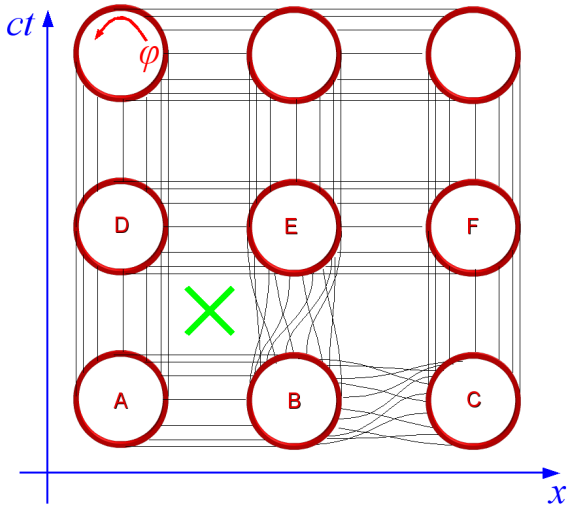


Figure 4.4: Two dimensional discrete ties space with intrinsic $U(1)$ circles (see text).

the twists between E-B and B-C compensate each other.

The distortion detected by the beetle along a closed path is independent of the chosen representation and therefore describes a *physical* property of the intrinsic degrees of freedom, namely - as we will see below - the presence of an electric or magnetic fields. Unlike in one dimension, where all gauge fields can be swept away and are therefore physically meaningless, in higher dimensions there are 'true' distortions along closed contours in this system of connections that cannot be eliminated by gauge transformations.

Remark: The 'real' distortions in the texture of this fabric, which cannot be removed by gauge transformation, manifest themselves physically as force fields (interactions). These force fields are completely determined by the symmetry group of their intrinsic degrees of freedom:

$U(1)$	electromagnetic interaction
$SU(2)$	weak interaction
$SU(3)$	strong interaction

These examples concern twists of the intrinsic degrees of freedom. Twisting the connections of space-time itself manifests itself as gravitation, where the tangent space plays the role of the inner degrees of freedom and the gauge field $A(x)$ is replaced by the Christoffel symbols. All four fundamental forces can thus be described in the same framework as gauge theories. The only significant difference between gravity and other gauge theories is that the tangent space is non-compact and directly linked to the underlying space-time, while theories like electrodynamics take place in intrinsic spaces attached to space-time, having no direct influence on space-time itself.

This is in fact exactly the reason why theory is with intrinsic degrees of freedom could be quantized successfully (such as the unified electroweak quantum theory and quantum chromodynamics) while a consistent quantum theory of gravity is still missing. If we quantize theories like electrodynamics, the intrinsic degrees of freedom become fuzzy because of quantum fluctuations while the supporting space-time remains classical and intact. However, if we make the attempt to quantize gravity, the tangent space and therewith space-time itself becomes fuzzy, so that, illustratively speaking, the ground slips away under our feet. In fact, a consistent formulation of quantum gravity is probably one of the most important open problem in physics. You are invited to solve it.

4.1.5 Covariant derivative

The equation $z_{\parallel}(x + \mathrm{d}x) = \big(1 + iA(x)\,\mathrm{d}x\big)\,z(x)$, which describes the change of the coordinate z under parallel transport, has to be generalized to higher dimensions with vectorial displacements $\mathrm{d}\mathbf{x}$. This means that the gauge field A is replaced by a 1-form with a lower index A_μ:

$$z_{\parallel}(\mathbf{x} + \mathrm{d}\mathbf{x}) = \big(1 + iA_\mu(\mathbf{x})\,\mathrm{d}x^\mu\big)\,z(\mathbf{x})\,. \tag{4.7}$$

We now consider a given phase field $z(\mathbf{x})$ and ask how this field changes when moving in the direction of the basis vector $\mathbf{e}_\mu = \partial_\mu$ with respect to this kind of intrinsic parallel transport. This change is carried out by the *covariant derivative*

$$\nabla_\mu z = \lim_{\lambda\to 0} \frac{z(\mathbf{x} + \lambda\mathbf{e}_\mu) - z_{\parallel}(\mathbf{x} + \lambda\mathbf{e}_\mu)}{\lambda} = \big(\partial_\mu - iA_\mu\big)\,z(x)\,, \tag{4.8}$$

which, as we have seen before, is just the representation of the abstract connection ∇. This relationship maps a tangent vector to an element of the Lie group, thus describing the actual rate of rotation on the circles when moving in a given spatial direction. The connection ∇ is thus a field of *Lie-group-valued 1-forms*. ∇ plays essentially the same role here as a connection in differential geometry, except that it now acts on the intrinsic degrees of freedom instead of the tangent space. So we recognize here another general construction principle of gauge theories:

> **The connection ∇ of the gauge theory is a linear map of a spatial-temporal tangent vector onto the Lie algebra of the corresponding symmetry group. It describes the rate at which the intrinsic degrees of freedom are transformed as we move in the given direction.**

Gauge transformations $z(\mathbf{x}) \to \tilde{z}(\mathbf{x}) = e^{if(\mathbf{x})}z(\mathbf{x})$ leave the abstract connection ∇ unchanged, but they do change its representation:

$$A_\mu(\mathbf{x}) \to \tilde{A}_\mu(\mathbf{x}) = A_\mu(\mathbf{x}) + \partial_\mu f(\mathbf{x}) \tag{4.9}$$

This confirms that gauge fields are physically equivalent if they differ only by the addition of the gradient of the scalar function $f(\mathbf{x})$.

4.1.6 Intrinsic curvature: The electromagnetic field

The resulting twist along the close contour can be interpreted as a curvature or distortion of the enclosed intrinsic space. The underlying space-time, however, is not affected. With the covariant derivative introduced above we obtain a curvature tensor $\mathbf{F}$, a 2-form which maps to tangent vectors $\mathbf{X}, \mathbf{Y}$ onto a number as follows:

$$\mathbf{F}(\mathbf{X}, \mathbf{Y}) = i\Big([\nabla_{\mathbf{X}}, \nabla_{\mathbf{Y}}] - \nabla_{[\mathbf{X},\mathbf{Y}]}\Big)\,. \tag{4.10}$$

Here we divided by the trivial generator i so that this 2-form renders just the twist angle in radians that the beetle would experience when moving along the close contour spanned by the two tangent vectors $\mathbf{X}$ and $\mathbf{Y}$. Since this 2-form is defined in a

representation-independent way, it is automatically independent of the coordinate system describing the internal degrees of freedom, meaning that it is automatically invariant under gauge transformations.

Now let us compute the components of this tensor in a given coordinate basis. Since the Lie bracket in the coordinate basis is always equal to zero, the last term vanishes, giving

$$F_{\mu\nu} \;=\; \mathbf{F}(\partial_\mu, \partial_\nu) \;=\; i[\nabla_\mu, \nabla_\nu] \;=\; \partial_\mu A_\nu - \partial_\nu A_\mu \,. \tag{4.11}$$

The tensor components represent the electromagnetic field and are by definition invariant under $U(1)$ gauge transformations. In the Minkowski space with an attached electromagnetic field $\mathbb{R}^{3+1} \oplus U(1)$ they are given by

$$F_{\mu\nu} = \begin{pmatrix} 0 & -E_x/c & -E_y/c & -E_z/c \\ E_x/c & 0 & B_z & -B_y \\ E_y/c & -B_z & 0 & B_x \\ E_z/c & B_y & -B_x & 0 \end{pmatrix} \tag{4.12}$$

Hence the fields $\vec{E}(\mathbf{x})$ and $\vec{B}(\mathbf{x})$ represent information content of the electromagnetic field which is independent of the chosen gauge. As we all know, these are just the physical fields that we can actually measure.

An electric field is a $U(1)$-distortion along a closed contour spanned by the time direction and a direction in space.

A magnetic field is a $U(1)$-distortion along a closed counter spanned by two different directions in space.

4.2 Electrodynamics in terms of differential forms

4.2.1 The electromagnetic field as a differential form

The commutator $[\nabla_\mu, \nabla_\nu]$ of a general covariant derivative $\nabla_\mu = \partial_\mu + \Gamma_\mu(\mathbf{x})$ acting on intrinsic degrees of freedom generally consists of three terms:

$$[\nabla_\mu, \nabla_\nu] \;=\; \partial_\mu \Gamma_\nu(\mathbf{x}) - \partial_\nu \Gamma_\mu(\mathbf{x}) + [\Gamma_\mu(\mathbf{x}), \Gamma_\nu(\mathbf{x})]. \tag{4.13}$$

Here the quantities $\Gamma_\mu(\mathbf{x})$ are the generators of the symmetry group, i.e., they are elements of the associated Lie algebra. The third term can be calculated using the commutation relations of Lie algebra which is nonzero in the case of noncommutative symmetry groups. However, the symmetry group of electrodynamics $U(1)$ is an exception in so far as it is commutative, meaning that this term disappears.

Since the last term cancels, the field tensor $\mathbf{F}$ in electrodynamics can be interpreted as an exterior derivative of the gauge field. As can be verified easily, the forms

$$\mathbf{A} = A_\mu dx^\mu \,, \qquad \mathbf{F} = \frac{1}{2} F_{\mu\nu}\, \mathrm{d}x^\mu \wedge \mathrm{d}x^\nu \tag{4.14}$$

satisfy the relation

$$\boxed{\mathbf{F} = \mathrm{d}\mathbf{A}\,.} \tag{4.15}$$

Because of $\mathrm{d}^2 = 0$ this leads immediately to the *homogeneous Maxwell equations*

$$\boxed{\mathrm{d}\mathbf{F} = 0} \tag{4.16}$$

or, in the coordinate representation

$$\partial_\rho F_{\mu\nu} + \partial_\nu F_{\rho\mu} + \partial_\mu F_{\nu\rho} = 0\,. \tag{4.17}$$

Remark: Because of the anti-symmetry of $\mathbf{F}$ only four out of $4^3 = 64$ possible index combinations are linearly independent. These four equations read:

$$\begin{aligned}
\partial_0 F_{12} + \partial_2 F_{01} + \partial_1 F_{20} &= 0\\
\partial_0 F_{13} + \partial_3 F_{01} + \partial_1 F_{30} &= 0\\
\partial_0 F_{23} + \partial_3 F_{02} + \partial_2 F_{30} &= 0\\
\partial_1 F_{23} + \partial_3 F_{12} + \partial_2 F_{31} &= 0\,.
\end{aligned}$$

Inserting the fields $\vec{E}$ und $\vec{B}$ according to Eq. (4.12) yields

$$\begin{aligned}
\partial_t B_z - \partial_y E_x + \partial_x E_y &= 0\\
-\partial_t B_y - \partial_z E_x + \partial_x E_z &= 0\\
\partial_t B_x - \partial_z E_y + \partial_y E_z &= 0\\
\partial_x B_x + \partial_y B_y + \partial_z B_z &= 0\,.
\end{aligned}$$

The first three and the last equation can be written as $\partial_t \vec{B} = -\nabla \times \vec{E}$ or $\nabla \cdot \vec{B} = 0$.

This highlights that the homogeneous Maxwell equations reflect the geometrical properties and the commutativity of the gauge group. They have no physical content in so far as they don't tell us something about the dynamics or the interaction with electric charges.

4.2.2 Equation of motions in differential forms

In classical physics, a particle does not have to be at rest, but it can move in various different ways, but constrained in so far as the action "consumed" along the path is minimal. The same applies to the connecting elements between the circles: they do not necessarily have to be straight, but they are allowed to twist, but only in such a way that the *action of the electromagnetic field* is extremal. By defining an action we specify at this point how the connecting elements function physically.

Clearly, the action is a scalar measuring the stress in the overall system, that is, it is somehow measures the total 'torsion' in the connection structure. For a system without twisted connections, the action should assume its minimum value. However, by specifying suitable boundary conditions (for example, by 'twisted' initial conditions), it is possible to exert a torsion on the connecting structure. The principle of least action must then adjust the other connections exactly in such a way that the torsion is minimized in the overall system. The result is a specific spatiotemporal configuration of connections, namely that of an electromagnetic field or an electromagnetic wave.

In finding the correct form of the action integral, physicists like to be guided by the

so-called *heuristic principle of simplicity*. This principle tells us that the action realized in nature is formally the simplest one consistent with the given symmetries of the system. In the case of the $U(1)$ gauge theory, this means the following: First, the action must be a spatiotemporal volume integral

$$S = \int \mathrm{d}^4x\, \mathcal{L} \tag{4.18}$$

of a scalar Lagrange density $\mathcal{L}$. Since the action has to be invariant unter gauge transformations, the Lagrange density should depend only on the 'true' fields $\vec{E}, \vec{B}$, i.e., it should be a function of the field tensor $\mathbf{F}$. The simplest scalar that can be constructed from $\mathbf{F}$ is the trace F^{μ}_{μ}, but unfortunately this trace is zero – so this attempt fails. The next possibility near at hand would be to contract the tensor with itself, i.e.,

$$\mathcal{L} = -\frac{1}{4} F^{\mu\nu} F_{\mu\nu}\,, \tag{4.19}$$

where the pre-factor $-1/4$ is just the convention. And, in fact, this is the right choice.

4.2.3 Equations of motion in components

Now let us vary the network of connections by an infinitesimal transformation $\mathbf{A} \to \mathbf{A} + \delta\mathbf{A}$ and let us study the corresponding change of the action $S \to S + \delta S$. Obviously we have

$$\begin{aligned} \delta S &= -\frac{1}{2}\int \mathrm{d}^4x\, F^{\mu\nu}\, \delta F_{\mu\nu} = -\frac{1}{2}\int \mathrm{d}^4x\, F^{\mu\nu}\, (\partial_\mu \delta A_\nu - \partial_\nu \delta A_\mu) \\ &= -\frac{1}{2}\int \mathrm{d}^4x\, (F^{\mu\nu} - F^{\nu\mu})\, \partial_\mu \delta A_\nu = -\int \mathrm{d}^4x\, F^{\mu\nu}\, \partial_\mu \delta A_\nu\,. \end{aligned} \tag{4.20}$$

Performing a partial integration we arrive at

$$\delta S = \int \mathrm{d}^4x\, \delta A_\nu\, \partial_\mu F^{\mu\nu}\,. \tag{4.21}$$

Since the components δA_ν can be varied independently, the action will be extremal if

$$\boxed{\partial_\mu F^{\mu\nu} = 0\,.} \tag{4.22}$$

These equations form the second set of Maxwell equations. With the contravariant representation of the field tensor

$$F^{\mu\nu} = \begin{pmatrix} 0 & E_x/c & E_y/c & E_z/c \\ -E_x/c & 0 & B_z & -B_y \\ -E_y/c & -B_z & 0 & B_x \\ -E_z/c & B_y & -B_x & 0 \end{pmatrix} \tag{4.23}$$

one can easily derive the Maxwell equations in the usual form $\nabla \cdot \vec{E} = 0$ and $\partial_t \vec{E} = \nabla \times \vec{B}$.

4.2.4 U(1) gauge symmetry

In the theory of electrodynamics the $U(1)$ symmetry of phases $z(\mathbf{x})$ is described by the connection

$$\nabla_{\mathbf{X}} \;=\; \mathbf{X} - i\mathbf{A}(\mathbf{X}) \tag{4.24}$$

where i is the generator of the corresponding Lie algebra and $\mathbf{A}$ is a 1-form which describes how the phase changes in the direction $\mathbf{X}$. Under gauge transformations $z(\mathbf{x}) \to z(\mathbf{x})e^{if(x)}$the connection $\mathbf{A}$ is transformed by

$$\mathbf{A} \to \mathbf{A} + \mathrm{d}\mathbf{f} \tag{4.25}$$

so that the exterior derivative

$$\mathbf{F} = \mathrm{d}\mathbf{A} \tag{4.26}$$

is a gauge-invariant quantity and thus describes the physical fields. Since $\mathbf{F}$ is an exact form, this implies the homogeneous Maxwell equations

$$\boxed{\mathrm{d}\mathbf{F} = 0\,.} \tag{4.27}$$

4.2.5 Action

The action S of the electromagnetic field is given by the 4-dimensional volume integral

$$S = \int \mathbf{L} \tag{4.28}$$

integrated over the 4-form

$$\mathbf{L}[\mathbf{A},\, \mathrm{d}\mathbf{A}] \;=\; -\frac{1}{2}\,\mathrm{d}\mathbf{A} \wedge \star\,\mathrm{d}\mathbf{A} + \mathbf{A} \wedge \star\mathbf{J}. \tag{4.29}$$

Here the 1-form $\mathbf{J}$ is the so-called *charge current density*[1]. By variation of the action one is led to the Langrange equations

$$\frac{\partial \mathbf{L}}{\partial \mathbf{A}} + \mathrm{d}\frac{\partial \mathbf{L}}{\partial(\,\mathrm{d}\mathbf{A})} = 0 \tag{4.30}$$

in the form of the inhomogeneous Maxwell equations

$$\boxed{\mathrm{d} \star \mathbf{F} = \star\mathbf{J}\,.} \tag{4.31}$$

Because of $\mathrm{d}^2 = 0$ this implies the continuity equation for the conservation of charges

$$\boxed{\mathrm{d} \star \mathbf{J} = 0.} \tag{4.32}$$

[1] In the literature the charge current density is often defined as a 3-form $\bar{\mathbf{J}}$. This 3-form quantifies the momentum which penetrates the three-dimensional hypersurface spanned by the three vectors. The 1-form $\mathbf{J} = \star\bar{\mathbf{J}}$ used in this lecture notes therefore describes the momentum which penetrates the hypersurface which is perpendicular to the vector applied to the form. The components of this form are the charge density σ and the charge current $\mathbf{j}$.

4.2.6 Wave equation

Applying the Hodge-Star operator to the inhomogeneous Maxwell equations $\mathrm{d}\star\,\mathrm{d}\mathbf{A} = \star\mathbf{J}$, we obtain

$$\mathrm{d}^\dagger\,\mathrm{d}\mathbf{A} = \mathbf{J}. \tag{4.33}$$

Since the Laplacian is given by $\Box = -\,\mathrm{d}^\dagger\,\mathrm{d} - \,\mathrm{d}\,\mathrm{d}^\dagger$, this implies $\Box\mathbf{A} = -\mathbf{J} - \,\mathrm{d}\,\mathrm{d}^\dagger\mathbf{A}$. With the Lorenz gauge

$$\mathrm{d}^\dagger\mathbf{A} = 0 \tag{4.34}$$

we obtain the usual wave equation

$$\boxed{\Box\mathbf{A} = -\mathbf{J}.} \tag{4.35}$$

4.2.7 Representation of electrodynamics

In a given reference frame with Minkowski coordinates, the differential forms used in the electronic dynamics can be represented as follows:

$$\begin{aligned}
\mathbf{A} &= \phi\,\mathrm{d}x^0 + A_x\,\mathrm{d}x^1 + A_y\,\mathrm{d}x^2 + A_z\,\mathrm{d}x^3 && (4.36)\\
\mathbf{F} &= E_x\,\mathrm{d}x^1\wedge\mathrm{d}x^0 + E_y\,\mathrm{d}x^2\wedge\mathrm{d}x^0 + E_z\,\mathrm{d}x^3\wedge\mathrm{d}x^0 && (4.37)\\
&+ B_x\,\mathrm{d}x^2\wedge\mathrm{d}x^3 + B_y\,\mathrm{d}x^3\wedge\mathrm{d}x^1 + B_z\,\mathrm{d}x^1\wedge\mathrm{d}x^2 \\
\mathbf{J} &= \rho\,\mathrm{d}x^0 + j_x\,\mathrm{d}x^1 + j_y\,\mathrm{d}x^2 + j_z\,\mathrm{d}x^3\,. && (4.38)
\end{aligned}$$

Dabei ist ϕ das Potential, $\vec{A} = (A_x, A_y, A_z)$ das Vektorpotential, $\vec{E} = (E_x, E_y, E_z)$ das elektrische Feld, $\vec{B} = (B_x, B_y, B_z)$ das magnetische Feld, ρ die Ladungsdichte und $\vec{j} = (j_x, j_y, j_z)$ die Stromdichte.

In this illustration, the corresponding Hodge duals are given by

$$\begin{aligned}
\star\mathbf{F} &= B_x\,\mathrm{d}x^0\wedge\mathrm{d}x^1 + B_y\,\mathrm{d}x^0\wedge\mathrm{d}x^2 + B_z\,\mathrm{d}x^0\wedge\mathrm{d}x^3 && (4.39)\\
&+ E_x\,\mathrm{d}x^2\wedge\mathrm{d}x^3 + E_y\,\mathrm{d}x^3\wedge\mathrm{d}x^1 + E_z\,\mathrm{d}x^1\wedge\mathrm{d}x^2 \\
\star\mathbf{J} &= \rho\;\mathrm{d}x^1\wedge\mathrm{d}x^2\wedge\mathrm{d}x^3 && (4.40)\\
&- j_x\,\mathrm{d}x^2\wedge\mathrm{d}x^3\wedge\mathrm{d}x^0 - j_y\,\mathrm{d}x^3\wedge\mathrm{d}x^1\wedge\mathrm{d}x^0 - j_z\,\mathrm{d}x^1\wedge\mathrm{d}x^2\wedge\mathrm{d}x^0
\end{aligned}$$

4.2.8 Charge conservation

Now consider a three-dimensional spatial domain G in this representation[2] and integrate the inhomogeneous Maxwell equations over this domain

$$\int_G \star\mathbf{J} \;=\; \int_G \mathrm{d}\star\mathbf{F} \;=\; \int_{\partial G} \star\mathbf{F} \tag{4.41}$$

[2] A spatial domain has no temporal extent. Note that this property is generally lost when changing the frame of reference

where on the right side Stokes theorem was applied. Since both G and ∂G are spatial domains, only those terms are included in the integrand that do not contain $\mathrm{d}t = \mathrm{d}x^0$. For example,

$$\int_G \star\mathbf{J} \;=\; \int_G \rho \; \mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3 \;=\; \int_G \rho(\vec{x})\, d^3x \;=\; Q \tag{4.42}$$

is the total charge contained in this domain. If we integrate over $\star F$ only the electrical components will contribute:

$$\int_{\partial G} \star\mathbf{F} \;=\; \int_{\partial G} E_x \, \mathrm{d}x^2 \wedge \mathrm{d}x^3 + E_y \, \mathrm{d}x^3 \wedge \mathrm{d}x^1 + E_z \, \mathrm{d}x^1 \wedge \mathrm{d}x^2 \tag{4.43}$$

The integrand $\sum_{i=1}^{3} E_i(\star \, \mathrm{d}x^i) = \vec{E} \cdot \vec{n} \, dS$ can be interpreted as the scalar product of the electric field with a normal vector $\vec{n}$ which is perpendicular on the area dS. This leads us to the theorem by Gauss for the charge

$$Q = \int_{\partial G} \vec{E} \cdot \vec{n} \, dS. \tag{4.44}$$

An analogous calculation for the homogeneous Maxwell equations results in

$$\int_{\partial G} \mathbf{F} = \int_{\partial G} \vec{B} \cdot \vec{n} \, dS = 0 \tag{4.45}$$

and tells us that there are no magnetic monopoles.

5 Field equations of general relativity

After the discovery of special relativity and its mathematical interpretation it was clear to Einstein and others that the concept of space-time opens the door for a radically new theory of gravity by considering a curved instead of a flat space-time. Everything seemed to be ready: Differential geometry was already well established in Mathematics. The simple task would be to define an appropriate action, apply the principle of least action, write down the Lagrange equations of motions, the soo-called *field equations*. General relativity appeared as a ripe fruit that you just have to pick.

Not only Einstein realized the feasibility of such a theory. Another heavyweight scientist was none other than the famous mathematician David Hilbert. Einstein and Hilbert soon saw each other in a frenetic race: Who would be the first to formulate General Relativity in a consistent form.

Unexpectedly, it took about 10 years (1905-1915) before Einstein could win the race. Before 1915, both men published a large variety of papers, partially withdrew and replaced their papers, struggling for a consistent formulation of the theory. The development was so chaotic that many scientist of that time lost interest and stopped to follow these publication. But finally Einstein made a substantial breakthrough, probably because he had the advantage of being a physicist with a good intuition. What happened?

5.1 Concept of General Theory of Relativity

5.1.1 Invariance under diffeomorphisms

What could such a "general theory of relativity" look like? As a necessary condition, the laws of physics have to apply not only to intertial systems but to *arbitrary* reference frames. The general theory of relativity must therefore be formulated in arbitrary coordinate systems, so their formulas must also be correct in the reference system of a roller coaster. Intuitively, it is clear that this can only be achieved if the 'normal' equations of motion are extended by correction terms that compensate for the acceleration effects in such frames, so we need a kind of acceleration gauge field. This gauge field is the gravitational field.

If the formulas of the general theory of relativity are correct in every frame of reference, that is, in every coordinate system, then they must be forminvariant under any coordinate transformations. In general relativity such maps are called *diffeomorphisms*.

> **Recall:** A diffeomorphism is a bijective, continuously differentiable map whose inverse map is also continuously differentiable. Illustratively speaking, a diffeomorphism is a topology-preserving map which does not tear the space apart.

After studying the differential geometry, Einstein soon realized that the Christoffel symbols determined from the metric tensor encode the gravitational field and that particles that are only subject to the gravitational force move on geodesic lines in the curved spacetime. This already fixed essential elements of the theory. But then Einstein encountered two fundamental problems. One of these was the required forminvariance when switching between arbitrary coordinate systems, i.e., covariance under passive diffeomorphisms. This problem can be described as follows:

In Sect. 1.3.3 on page 9 we discussed the difference between active and passive transformations. Suppose that we would decide to shift the zero meridian from Greenwich to Würzburg. Then, on a map, the coordinates of all cities would change through such a passive transformation. Alternatively, we could also demolish the cities and rebuild them further to the west, - the effect of such an active transformation on the coordinates would be identical. The inhabitants of Würzburg, however, would find themselves somewhere in France.

For the general theory of relativity, this applies in a similar way: For every passive coordinate transformation on the map, there must be a corresponding active transformation on the manifold (space-time), so that the same effect is achieved on the map. The equations of motion must therefore be invariant under both passive and active diffeomorphisms. In other words, any solution of the (yet to be found) field equations is mapped by a diffeomorphism into any other solution.

So far that is nothing new. Also in the special theory of relativity, the solutions of the equations of motion mapped Lorentz transformation are again solutions of the equation of motion. Butin the general theory of relativity the invariance group is much much larger, namely, it comprises all diffeomorphisms, that is, basically all thinkable maps that do not tear the space apart. But this is exactly where the problem is. Namely, it is possible to construct special diffeomorphisms that are identical in one part of the manifold but not in the other. As a cartoon we present and a 1+1-dimensional manifold, which is divided into two temporal domains by a space-like hypersurface (see Fig. 5.1). In the lower area, in which the initial condition is fixed, the diffeomorphism is identical, thus it does not modify the solution or the course of the trajectories. In the upper area, on the other hand, there are shifts in the trajectory. But because an active diffeomorphism maps a solution to another solution, both trajectories must be solutions to the same initial conditions. But how can there be several solutions to a single initial condition in a deterministic theory?

5.1.2 On the physical meaning of the manifold

Einstein spends several years for this “struggle for the meaning of coordinates”. He rejects earlier publications and works unsuccessfully on non-covariant approaches. Only in 1915 does he suddenly return to the covariant formulation and then everything goes very fast. Put in simple terms he realizes that the two solutions in the above example, although they look different on the manifold, nevertheless describe the *same* physical situation. The implication would be that

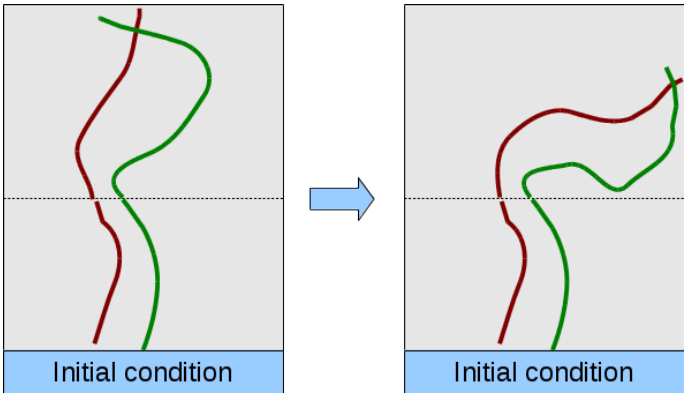


Figure 5.1: Einstein's problem: The manifold is divided into two areas by a space-like surface (dashed line). An active diffeomorphism then exclusively maps the upper part in a nontrivial way. Then the old and the new trajectory must both be solutions to the same equations. But since the initial condition remains unchanged, it is not clear why different trajectories can arise with the same initial conditions. See text.

The points on the manifold have no physical meaning.

The lesson is that we should not identify the points on the manifold with the physical events. So, if we read (in textbooks and also in these lecture notes) that the general theory of relativity lives on a curved space-time described by a differentiable manifold $\mathcal{M}$, then we usually infer the idea that $\mathcal{M}$ *is* space-time, but this idea is wrong!

But what then is the meaning of the manifold? It is a kind of projection screen for the theory in which the same physics can be represented in an infinite number of different ways. It is a kind of mathematical vehicle with which we can represent physics in a highly redundant way. But physically does not exist, it is merely some kind of bookkeeping tool.

But, you will argue, in the special theory of relativity we learned that the flat Minkowski space was indeed identical with the true space-time, in this case the Minkowski space was indeed considered as a physically existing reality. Where has this spacetime gone? The answer is that such a space-time is indeed a possible mode of description for the gravitational-free case, but that it loses its meaning in general relativity.

> **Example:** What happens if you could turn on a homogeneous static gravitational field in the universe? The galaxies would be deflected in this field, so would describe a different path on the manifold with the same initial conditions than without field. Nevertheless, we would notice nothing of this field. The conclusion: Neither the homogeneous field nor the manifold actually exist, but they prove to be redundant elements of the mathematical description, so to speak as "gauge freedom".

At this point, the ground slips away under our feet because we have to say goodbye to a cherished notion: Newton's absolute space. It does not exist, neither flat nor curved, instead there is only the gravitational field. Yes, there are no points in space, there are only connections.

Now we see where Newton was wrong: Newton mistakenly considered the vanishing gravitational field to be an absolute space. In order to be able to describe gravity in a world of zero gravity, he artificially introduced a second gravitational field on top of

the real one, an additional field which is implemented as an instantaneous interaction at a distance.

5.2 Field equations

5.2.1 The concept of the field equations

Einstein's field equations describe how matter bends space-time. By "matter" it means anything that is not gravity. This includes all forms of matter, charges, and radiation that are not gravitational in nature, that is, in present-day parlance, all elementary particles and gauge bosons except the graviton.

The field equations are derived - as always - from the principle of least action. It is assumed that the total action of the universe is the sum of a gravitational contribution and another contribution related to the matter content, i.e.

$$S = S_G + S_M. \tag{5.1}$$

The contributions to the action can be written as integrals over the entire manifold of the corresponding Lagrange 4-forms $\mathbf{L}_G$ and $\mathbf{L}_M$

$$S \;=\; \gamma \int \mathbf{L}_G \;+\; \int \mathbf{L}_M \tag{5.2}$$

or in a coordinate representation as integrals over Lagrange densities

$$S \;=\; \gamma \int \mathcal{L}_G \sqrt{-g}\; \mathrm{d}^4x \;+\; \int \mathcal{L}_M \sqrt{-g}\; \mathrm{d}^4x\,. \tag{5.3}$$

The fact that the action can be written as the sum of a gravitational and a non-gravitational part suggests at first glance that these two parts would not interact. In non-relativistic physics, where space-time is a static container, this way of thinking would be correct. In general relativity, however, space-time itself becomes a dynamic object, and a variation of the space-time in the first integral therefore usually leads to a change in the second integral, because it is integrated over space-time. In a coordinate representation, this coupling manifests itself by the fact that in the second integral, all partial derivatives must be replaced by covariant derivatives, and thus the value of the integral will implicitly depend on the Christoffel symbols.

Since the two components of the action are thus indirectly coupled via the geometry of the manifold, we have to introduce a coupling constant γ in order to specify how the two components are weighted relative to each other. This coupling constant has the quality of a new fundamental constant an equal footing with $\hbar, c$ and describes to what extent a given mass mass bends space-time.

Quantities to be varied

Which quantities are to be varied in the action integral in order to arrive at the correct equations of motion? Since a particle in the gravitational field always follows a geodesic line given by Eq. (3.27), its trajectory depends on the Christoffel symbols, and since the Christoffel symbols are given via Eq. (3.33) in terms of the metric tensor, we can conclude that it is the metric tensor $g_{\mu\nu}$ which has to be varied.

> **Remark:** There are several formal approaches that differ in which quantities are considered to be the 'gravitational field'. The traditional approach proposed by Einstein interprets the metric as the gravitational field, thus varying according to the components of the metric tensor. However, this method has the disadvantage that one cannot integrate fermionic quantum fields consistently, and for this reason it is increasingly being replaced by more modern variants, some of which we will discuss below. The most important variant used today interprets the local Minkowski base (Vierbein) as the gravitational field.

5.2.2 Action S_G of the gravitational field and the field equations in vacuum

What is the Langrange density of the gravitational field? Again, one is guided by the heuristic principle of simplicity, suggesting that we have to search for the simplest formula that is compatible with the required symmetries.

The simplest scalar that describes the curvature of the manifold is the scalar curvature obtained from the contraction of the Ricci tensor $R = R^{\mu}_{\mu}$. But there is an even simpler scalar, namely, a constant. Adding such a constant has a physical effect, as it would lead to a term in the Lagrangian that is proportional to the 4-volume (i.e., the volumetric form). Depending on the sign, such a constant would therefore cause a homogeneous expansion or a contraction of space-time, that is to say, it would act like a gravitative or antigravitative homogeneous 'ether', and it would equally affect the entire universe. If we call this constant 2Λ, then the action of the gravitational field would be given by

$$\boxed{S_G \;=\; \gamma \int (R - 2\Lambda)\sqrt{-g}\; \mathrm{d}^4x\,.} \tag{5.4}$$

The constant Λ is denoted as the *cosmological constant*. It has a colorful history, which we will return to later.

By varying the components of the metric tensor $g_{\mu\nu} \to g_{\mu\nu} + \delta g_{\mu\nu}$ and applying the standard variational calculus (here without proof) one gets

$$\delta S_G \;=\; \gamma \int \left(R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} + \Lambda g_{\mu\nu}\right) \delta g^{\mu\nu}\, \sqrt{-g}\; \mathrm{d}^4x \tag{5.5}$$

Without presence of matter the contribution S_G has to be extremal, i.e., $\delta S_G = 0$. Since all components of the metric tensor (up to symmetry) can be varied independently, the integrand must vanish. In this way we get the *vacuum field equations*

$$\boxed{R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} + \Lambda g_{\mu\nu} \;=\; 0\,.} \tag{5.6}$$

This specific combination of the Ricci tensor and the Ricci scalar is also referred to as *Einstein tensor*

$$E_{\mu\nu} \;:=\; R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu}\,, \tag{5.7}$$

which is also often denoted as $G_{\mu\nu}$ in the literature. Then the vacuum field equations take on the form $E_{\mu\nu} + \Lambda g_{\mu\nu} = 0$.

5.2.3 Action S_M of the matter field and the form of the field equations

From the perspective of a relativist, matter is anything that is not gravity, that is, essentially the entire particle and radiation content of the Standard Model of particle physics. As complicated as this Lagrangian may be (see figure), if we vary the metric, the variation of the action S_M always acquires the form

$$\delta S_M \;=\; -\frac{1}{2}\int T_{\mu\nu}\,\delta g^{\mu\nu}\,\sqrt{-g}\,\mathrm{d}^4x\,, \tag{5.8}$$

that is, we get a certain tensor field $\mathbf{T}$ of rank 2 which, when contracted with the variation $\delta\mathbf{g}$, just yields the local change of the action. This tensor is called *energy-momentum tensor*. Since we are varying here with respect to a symmetric quantity $g_{\mu\nu}$, it is clear that the tensor field $\mathbf{T}(\mathbf{x})$ will be symmetric as well. This tensor comprises the whole matter content of the theory. What this tensor field looks like in certain cases will be discussed in the following section.

$$\begin{aligned}
&-\tfrac{1}{4}\partial_\nu g^a_\mu \partial_\nu g^a_\mu - g_s f^{abc}\partial_\mu g^a_\nu g^b_\mu g^c_\nu - \tfrac{1}{4}g_s^2 f^{abc}f^{adr} g^b_\mu g^c_\nu g^d_\mu g^r_\mu + \\
&\tfrac{1}{2}ig_s^2(\bar{q}^\sigma_i\gamma^\mu q^\sigma_j)g^a_\mu + \bar{G}^a\partial^2 G^a + g_s f^{abc}\partial_\mu \bar{G}^a G^b g^c_\mu - \partial_\nu W^+_\mu \partial_\nu W^-_\mu - \\
&M^2 W^+_\mu W^-_\mu - \tfrac{1}{2}\partial_\nu Z^0_\mu \partial_\nu Z^0_\mu - \tfrac{1}{2c_w^2}M^2 Z^0_\mu Z^0_\mu - \tfrac{1}{2}\partial_\mu A_\nu \partial_\mu A_\nu - \tfrac{1}{2}\partial_\mu H \partial_\mu H - \\
&\tfrac{1}{2}m_h^2 H^2 - \partial_\mu\phi^+\partial_\mu\phi^- - M^2\phi^+\phi^- - \tfrac{1}{2}\partial_\mu\phi^0\partial_\mu\phi^0 - \tfrac{1}{2c_w^2}M\phi^0\phi^0 - \beta_h[\tfrac{2M^2}{g^2} + \\
&\tfrac{2M}{g}H + \tfrac{1}{2}(H^2+\phi^0\phi^0+2\phi^+\phi^-)] + \tfrac{2M^4}{g^2}\alpha_h - ig c_w[\partial_\nu Z^0_\mu(W^+_\mu W^-_\nu - \\
&W^+_\nu W^-_\mu) - Z^0_\mu(W^+_\mu\partial_\nu W^-_\nu - W^-_\mu\partial_\nu W^+_\nu) + Z^0_\nu(W^+_\mu\partial_\nu W^-_\mu - \\
&W^-_\mu\partial_\nu W^+_\mu)] - ig s_w[\partial_\nu A_\mu(W^+_\mu W^-_\nu - W^+_\nu W^-_\mu) - A_\nu(W^+_\mu\partial_\nu W^-_\mu - \\
&W^-_\mu\partial_\nu W^+_\mu) + A_\mu(W^+_\nu\partial_\nu W^-_\mu - W^-_\nu\partial_\nu W^+_\mu)] - \tfrac{1}{2}g^2 W^+_\mu W^-_\mu W^+_\nu W^-_\nu + \\
&\tfrac{1}{2}g^2 W^+_\mu W^-_\nu W^+_\mu W^-_\nu + g^2 c_w^2(Z^0_\mu W^+_\mu Z^0_\nu W^-_\nu - Z^0_\mu Z^0_\mu W^+_\nu W^-_\nu) + \\
&g^2 s_w^2(A_\mu W^+_\mu A_\nu W^-_\nu - A_\mu A_\mu W^+_\nu W^-_\nu) + g^2 s_w c_w[A_\mu Z^0_\nu(W^+_\mu W^-_\nu - \\
&W^+_\nu W^-_\mu) - 2A_\mu Z^0_\mu W^+_\nu W^-_\nu] - g\alpha[H^3 + H\phi^0\phi^0 + 2H\phi^+\phi^-] - \\
&\tfrac{1}{8}g^2\alpha_h[H^4 + (\phi^0)^4 + 4(\phi^+\phi^-)^2 + 4(\phi^0)^2\phi^+\phi^- + 4H^2\phi^+\phi^- + 2(\phi^0)^2H^2] - \\
&gMW^+_\mu W^-_\mu H - \tfrac{1}{2}g\tfrac{M}{c_w^2}Z^0_\mu Z^0_\mu H - \tfrac{1}{2}ig[W^+_\mu(\phi^0\partial_\mu\phi^- - \phi^-\partial_\mu\phi^0) - \\
&W^-_\mu(\phi^0\partial_\mu\phi^+ - \phi^+\partial_\mu\phi^0)] + \tfrac{1}{2}g[W^+_\mu(H\partial_\mu\phi^- - \phi^-\partial_\mu H) - W^-_\mu(H\partial_\mu\phi^+ - \\
&\phi^+\partial_\mu H)] + \tfrac{1}{2}g\tfrac{1}{c_w}(Z^0_\mu(H\partial_\mu\phi^0 - \phi^0\partial_\mu H) - ig\tfrac{s^2_w}{c_w}MZ^0_\mu(W^+_\mu\phi^- - W^-_\mu\phi^+) + \\
&igs_w MA_\mu(W^+_\mu\phi^- - W^-_\mu\phi^+) - ig\tfrac{1-2c_w^2}{2c_w}Z^0_\mu(\phi^+\partial_\mu\phi^- - \phi^-\partial_\mu\phi^+) + \\
&igs_w A_\mu(\phi^+\partial_\mu\phi^- - \phi^-\partial_\mu\phi^+) - \tfrac{1}{4}g^2W^+_\mu W^-_\mu[H^2 + (\phi^0)^2 + 2\phi^+\phi^-] - \\
&\tfrac{1}{8}g^2\tfrac{1}{c_w^2}Z^0_\mu Z^0_\mu[H^2 + (\phi^0)^2 + 2(2s_w^2 - 1)^2\phi^+\phi^-] - \tfrac{1}{2}g^2\tfrac{s^2_w}{c_w}Z^0_\mu\phi^0(W^+_\mu\phi^- + \\
&W^-_\mu\phi^+) - \tfrac{1}{2}ig^2\tfrac{s^2_w}{c_w}Z^0_\mu H(W^+_\mu\phi^- - W^-_\mu\phi^+) + \tfrac{1}{2}g^2 s_w A_\mu\phi^0(W^+_\mu\phi^- + \\
&W^-_\mu\phi^+) + \tfrac{1}{2}ig^2 s_w A_\mu H(W^+_\mu\phi^- - W^-_\mu\phi^+) - g^2\tfrac{s_w}{c_w}(2c_w^2 - 1)Z^0_\mu A_\mu\phi^+\phi^- - \\
&g^2 s_w^2 A_\mu A_\mu\phi^+\phi^- - \bar{e}^\lambda(\gamma\partial + m^\lambda_e)e^\lambda - \bar{\nu}^\lambda\gamma\partial\nu^\lambda - \bar{u}^\lambda_j(\gamma\partial + m^\lambda_u)u^\lambda_j - \\
&\bar{d}^\lambda_j(\gamma\partial + m^\lambda_d)d^\lambda_j + igs_w A_\mu[-(\bar{e}^\lambda\gamma^\mu e^\lambda) + \tfrac{2}{3}(\bar{u}^\lambda_j\gamma^\mu u^\lambda_j) - \tfrac{1}{3}(\bar{d}^\lambda_j\gamma^\mu d^\lambda_j)] + \\
&\tfrac{ig}{4c_w}Z^0_\mu[(\bar{\nu}^\lambda\gamma^\mu(1+\gamma^5)\nu^\lambda) + (\bar{e}^\lambda\gamma^\mu(4s_w^2 - 1 - \gamma^5)e^\lambda) + (\bar{u}^\lambda_j\gamma^\mu(\tfrac{8}{3}s_w^2 - \\
&1 - \gamma^5)u^\lambda_j) + (\bar{d}^\lambda_j\gamma^\mu(1 - \tfrac{8}{3}s_w^2 - \gamma^5)d^\lambda_j)] + \tfrac{ig}{2\sqrt{2}}W^+_\mu[(\bar{\nu}^\lambda\gamma^\mu(1+\gamma^5)e^\lambda) + \\
&(\bar{u}^\lambda_j\gamma^\mu(1+\gamma^5)C_{\lambda\kappa}d^\kappa_j)] + \tfrac{ig}{2\sqrt{2}}W^-_\mu[(\bar{e}^\lambda\gamma^\mu(1+\gamma^5)\nu^\lambda) + (\bar{d}^\kappa_j C^\dagger_{\lambda\kappa}\gamma^\mu(1+ \\
&\gamma^5)u^\lambda_j)] + \tfrac{ig}{2\sqrt{2}}\tfrac{m^2_e}{M}[-\phi^+(\bar{\nu}^\lambda(1-\gamma^5)e^\lambda) + \phi^-(\bar{e}^\lambda(1+\gamma^5)\nu^\lambda)] - \\
&\tfrac{g}{2}\tfrac{m^\lambda_e}{M}[H(\bar{e}^\lambda e^\lambda) + i\phi^0(\bar{e}^\lambda\gamma^5 e^\lambda)] + \tfrac{ig}{2M\sqrt{2}}\phi^+[-m^\kappa_d(\bar{u}^\lambda_j C_{\lambda\kappa}(1-\gamma^5)d^\kappa_j) + \\
&m^\lambda_u(\bar{u}^\lambda_j C_{\lambda\kappa}(1+\gamma^5)d^\kappa_j] + \tfrac{ig}{2M\sqrt{2}}\phi^-[m^\lambda_d(\bar{d}^\lambda_j C^\dagger_{\lambda\kappa}(1+\gamma^5)u^\kappa_j) - m^\kappa_u(\bar{d}^\lambda_j C^\dagger_{\lambda\kappa}(1- \\
&\gamma^5)u^\kappa_j] - \tfrac{g}{2}\tfrac{m^\lambda_u}{M}H(\bar{u}^\lambda_j u^\lambda_j) - \tfrac{g}{2}\tfrac{m^\lambda_d}{M}H(\bar{d}^\lambda_j d^\lambda_j) + \tfrac{ig}{2}\tfrac{m^\lambda_u}{M}\phi^0(\bar{u}^\lambda_j\gamma^5 u^\lambda_j) - \\
&\tfrac{ig}{2}\tfrac{m^\lambda_d}{M}\phi^0(\bar{d}^\lambda_j\gamma^5 d^\lambda_j) + \bar{X}^+(\partial^2 - M^2)X^+ + \bar{X}^-(\partial^2 - M^2)X^- + \bar{X}^0(\partial^2 - \\
&\tfrac{M^2}{c_w^2})X^0 + \bar{Y}\partial^2 Y + igc_w W^+_\mu(\partial_\mu\bar{X}^0 X^- - \partial_\mu\bar{X}^0 X^0) + igs_w W^+_\mu(\partial_\mu\bar{Y}X^- - \\
&\partial_\mu\bar{X}^+ Y) + igc_w W^-_\mu(\partial_\mu\bar{X}^- X^0 - \partial_\mu\bar{X}^0 X^+) + igs_w W^-_\mu(\partial_\mu\bar{X}^- Y - \\
&\partial_\mu\bar{Y}X^+) + igc_w Z^0_\mu(\partial_\mu\bar{X}^+ X^+ - \partial_\mu\bar{X}^- X^-) + igs_w A_\mu(\partial_\mu\bar{X}^+ X^+ - \\
&\partial_\mu\bar{X}^- X^-) - \tfrac{1}{2}gM[\bar{X}^+ X^+ H + \bar{X}^- X^- H + \tfrac{1}{2}\bar{X}^0 X^0 H] + \\
&\tfrac{1-2c_w^2}{2c_w}igM[\bar{X}^+ X^0\phi^+ - \bar{X}^- X^0\phi^-] + \tfrac{1}{2c_w}igM[\bar{X}^0 X^-\phi^+ - \bar{X}^0 X^+\phi^-] + \\
&igMs_w[\bar{X}^0 X^-\phi^+ - \bar{X}^0 X^+\phi^-] + \tfrac{1}{2}igM[\bar{X}^+ X^+\phi^0 - \bar{X}^- X^-\phi^0]
\end{aligned}$$

Lagrangian of the Standard Model

If we now carry out the variational calculation for the total action $S = S_G + S_M$, one gets in Eq. (5.6) the term $\frac{1}{2\gamma}T_{\mu\nu}$ on the right hand side. In the following chapter we will consider the approximation for weak gravitational fields and compare it with Newton's theory of gravity. This comparison will show that the coupling constant γ is given (up to geometrical factors) by the reciprocal Newtonian gravitational constant:

$$\gamma = \frac{c^4}{16\pi G} \tag{5.9}$$

Thus the field equations in full form read:

$$\boxed{R_{\mu\nu} - \frac{1}{2}R g_{\mu\nu} + \Lambda g_{\mu\nu} \;=\; \frac{8\pi G}{c^4}T_{\mu\nu}} \tag{5.10}$$

or with the Einstein tensor

$$E_{\mu\nu} + \Lambda g_{\mu\nu} \;=\; \frac{8\pi G}{c^4}T_{\mu\nu}\,, \tag{5.11}$$

where $G \simeq 6.67408 \times 10^{-11}\,\frac{\mathrm{m}^3}{\mathrm{kg\,s}^2}$.

Remark: Why is the coupling constant γ proportional to G^{-1} and not to G? In order to understand this, we can view the action integral in a way that the contribution of matter is 'compensated' by the contribution of the gravitational field in such a way that the overall action is minimized. The smaller γ is, the larger must be the compensating gravitational field. Since the gravitational field is directly coupled back into the covariant derivatives of the equations of motion for the particles, we expect that for small γ the particle trajectories will be more strongly curved by gravitational effects.

By contracting both sides of the above field equations with $g^{\mu\nu}$, we get the scalar relation

$$-R + 4\Lambda = \frac{8\pi G}{c^4} T\,, \tag{5.12}$$

where $R = R^{\mu}{}_{\mu}$ is the curvature tensor and $T = T^{\mu}{}_{\mu}$ is the trace of the energy-momentum tensor. This relation can be used to bring the second term in the field equations to the other side. We then arrive at the alternative form of the field equations, namely

$$\boxed{R_{\mu\nu} = \Lambda g_{\mu\nu} + \frac{8\pi G}{c^4}\left(T_{\mu\nu} - \frac{1}{2} T g_{\mu\nu}\right).} \tag{5.13}$$

Einstein's field equations enable us, in principle, to calculate the Ricci tensor for any given the energy-momentum distribution of matter. But doing so we do not yet know the full Riemann curvature tensor (the one with four indices), and the question arises as to whether this tensor contains more information than the rank-2 Ricci tensor, and if so, which one. The equations already give us a physical hint: in vacuum, with a vanishing cosmological constant, we get $R_{\mu\nu} = 0$. However, this does not imply that the spacetime is flat in the sense that $R^{\mu}_{\nu\alpha\beta} = 0$. As we shall see, the remaining freedom is exactly what we need for gravitational waves.

5.2.4 Form of the energy-momentum tensor

For a given Lagrangian $\mathcal{L}_M$ describing the matter content, the variation of the action integral $S_M = \int \mathrm{d}^4x\sqrt{-g}\mathcal{L}_M$ and the application of standard methods leads to

$$\delta S_M \;=\; \int \mathrm{d}^4x \left(\frac{\partial(\sqrt{-g}\mathcal{L}_M)}{\partial g^{\mu\nu}} - \Big[\frac{\partial(\sqrt{-g}\mathcal{L}_M)}{\partial g^{\mu\nu}{}_{,\lambda}}\Big]_{,\lambda} \right) \delta g^{\mu\nu}\,. \tag{5.14}$$

This implies that the energy-momentum tensor is given by

$$T_{\mu\nu} \;=\; -\frac{2}{\sqrt{-g}} \left(\frac{\partial(\sqrt{-g}\mathcal{L}_M)}{\partial g^{\mu\nu}} - \Big[\frac{\partial(\sqrt{-g}\mathcal{L}_M)}{\partial g^{\mu\nu}{}_{,\lambda}}\Big]_{,\lambda} \right). \tag{5.15}$$

In this expression, which is strongly reminiscent of the Lagrange equations of motion, the Lagrangian density is linear. We can thus determine the corresponding additive proportion of the energy-momentum tensor for each additive contribution to the Lagrangian density, i.e., we can calculate the respective contributions to $T_{\mu\nu}$ separately for the different forms of matter and radiation.

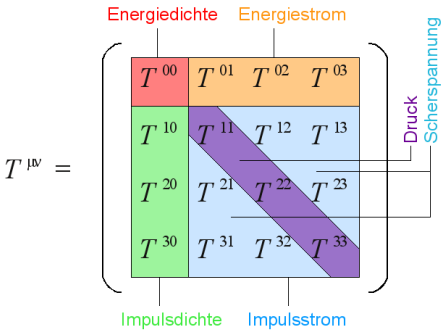


Figure 5.2: Interpretation of the components of the energy-momentum tensor

Interpretation of the energy-momentum tensor

Since $T^{\mu\nu}$ is defined as the local change of the Lagrangian density under variation of the (dimensionless) metric, this tensor must have the physical unit of an action-4-density. Since the dimension of an action equals energy·time, $T^{\mu\nu}$ has the dimension energy divided by a 3-volume, i.e., an energy density. But since the dimension of the action is also momentum·length, one can interpret the dimension of $T^{\mu\nu}$ as well as

$$[\mathbf{T}] = \frac{\text{momenteum}}{\text{time} \cdot \text{area}} .$$

The physical meaning of the tensor is most easily explained in the language of differential forms. While a single particle is described by its four-momentum p^μ, continuously distributed matter is characterized by a *4-momentum density*. From the viewpoint of differential geometry this is a vector-valued 3-form $\mathbf{P}^\mu$ which, applied to a 3-volume, returns the four-momentum p^μ contained therein. Such a 3-volume may be an ordinary spatial volume $\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z$, but may also be a spatio-temporal volume such as e.g. $\mathrm{d}t \wedge \mathrm{d}x \wedge \mathrm{d}y$. In the latter case, the result returned by the form can be understood as the 4-momentum passing through the surface $\mathrm{d}x \wedge \mathrm{d}y$ in the time $\mathrm{d}t$, i.e., it is a *surface current density*.

In the traditional formulation of general relativity, it is more common to use the Hodge dual instead of the vector-valued 3-form $\mathbf{P}^\mu$:

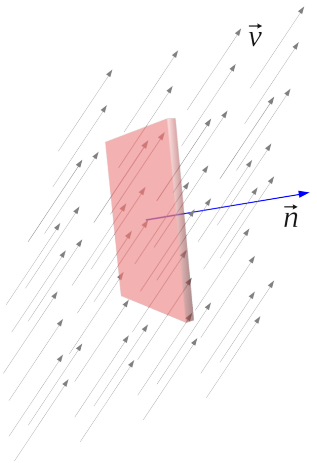
$$\mathbf{T}^\mu = \star\mathbf{P}^\mu \tag{5.16}$$

This *energy-momentum tensor* is a vector-valued 1-form and therefore it has only two instead of four indices. For this reason this tensor is the preferred one in the traditional formulation of ART. The energy-momentum tensor needs as input a single 4-vector, which is the normal to a certain 3-volume. The value returned by the form is the 4-momentum present in this 3-volume. For example, $\mathbf{T}^\mu(\mathrm{d}t)$ is the 4-momentum density while $\mathbf{T}^\mu(\mathrm{d}x)$ is the 4-momentum-current passing through a surface element placed in yz direction (see Fig. 5.2).

Why at all do we need a *tensor* to describe the energy-momentum content? Would not a vector be enough? To understand this, let us first imagine a homogeneous cloud of particles in $\mathbb{R}^3$ flying in parallel with velocity $\vec{v}$. Further, let us consider a surface

element whose size and orientation is determined by the normal vector $\vec{n}$. It is clear that the particle flux per unit time through this area element is equal to the scalar product $\vec{v} \cdot \vec{n}$. Since each particle carries a momentum $\vec{p}$, the momentum flux through the surface is given by $\vec{p}(\vec{v} \cdot \vec{n}) = m\vec{v}(\vec{v} \cdot \vec{n})$.

This map can be interpreted as a tensor T acting on $\vec{n}$ by $T(\vec{n}) = m\vec{v}(\vec{v} \cdot \vec{n})$. This tensor is just the dyadic product $T = m\vec{v} \circ \vec{v}$ or in Dirac notation $T = m|v\rangle\langle v|$, hence it projects the normal vector onto the velocity and returns the corresponding momentum.



Particles crossing a test area

The tensor **T** cannot always be written as a dyadic product. For example, if we consider a cloud of non-interacting particles, half of them flying upwards with velocity $\vec{v}_1$, the other flying to the right with velocity $\vec{v}_2$ (and not colliding as point particles), the corresponding tensor $T = \frac{1}{2}m(|v_1\rangle\langle v_1| + |v_2\rangle\langle v_2|)$ is the sum of the two parts. This sum can no longer be written as a dyadic product, meaning that this mixture would be distinguishable from the preceding example by measurement at the test area. A single vector would not be enough to make this difference.

Remark: You may know a similar situation from quantum theory. A statistical ensemble of quantum systems is described there by a density matrix. For a pure state, this matrix has the form of a dyadic product $|\psi\rangle\langle\psi|$, while general mixed states cannot be written that way. The density matrix contains the maximum information about the probability distribution of the states which can be extracted by measurement. Similarly, the energy-momentum tensor contains the maximum information of the probability distributions of the directions of flight extractable by measurement on test areas.

Single particles

The above considerations in $\mathbb{R}^3$ apply analogously to general relativity in four dimensions. The energy-momentum tensor $T^{\mu\nu}$ of an *individual* particle moving along the path $\mathbf{y}(\tau)$ is therefore proportional to $mu^\mu u^\nu$, where $u^\mu = \frac{\mathrm{d}}{\mathrm{d}\tau}y^\mu$ is the velocity of the particle, which in turn is defined as the derivative of the trajectory coordinate $y^\mu(\tau)$ with respect to the proper time τ. Since in four dimensions the particle is described by a trajectory (*world-line*) rather than a point, one has to integrate over this trajectory. So the energy-momentum tensor of a single particle moving on the path $y(\tau)$ is given by

$$T^{\mu\nu}(\mathbf{x}) = m\int \mathrm{d}\tau\, \delta^4(\mathbf{x} - \mathbf{y}(\tau))\, \frac{\mathrm{d}y^\mu(\tau)}{\mathrm{d}\tau}\, \frac{\mathrm{d}y^\nu(\tau)}{\mathrm{d}\tau} \tag{5.17}$$

If the particle is not subject to external forces, the following conservation law applies:

$$\partial_\mu T^{\mu\nu} = 0\,. \tag{5.18}$$

Proof: In forming the divergence, one applies the chain rule and obtains

$$\begin{aligned}\partial_\mu T^{\mu\nu} &= m\int \mathrm{d}\tau\, \frac{\mathrm{d}y^\nu(\tau)}{\mathrm{d}\tau}\,\frac{\mathrm{d}y^\mu(\tau)}{\mathrm{d}\tau}\,\frac{\partial}{\partial x^\mu}\delta^4(\mathbf{x}-\mathbf{y}(\tau))\\ &= -m\int \mathrm{d}\tau\, \frac{\mathrm{d}y^\nu(\tau)}{\mathrm{d}\tau}\,\frac{\mathrm{d}y^\mu(\tau)}{\mathrm{d}\tau}\,\frac{\partial}{\partial y^\mu}\delta^4(\mathbf{x}-\mathbf{y}(\tau))\\ &= -m\int \mathrm{d}\tau\, \frac{\mathrm{d}y^\nu(\tau)}{\mathrm{d}\tau}\,\frac{\mathrm{d}}{\mathrm{d}\tau}\delta^4(\mathbf{x}-\mathbf{y}(\tau))\end{aligned}$$

By partial integration we get

$$\partial_\mu T^{\mu\nu} = m\int \mathrm{d}\tau\, \delta^4(\mathbf{x}-\mathbf{y}(\tau))\,\frac{\mathrm{d}^2 y^\nu(\tau)}{\mathrm{d}\tau^2}\,.$$

In a force-free uniform motion, the second derivative is zero.

Perfect fluids

In physics, a *fluid* is meant to be a spatially extended substance which has no thermal conductivity and does not develop shear forces for low velocities, i.e., the viscosity is zero. The term *fluid* not only includes certain liquids, but also gases, plasmas, and even radiation.

A *perfect fluid*s understood to mean a fluid that is completely defined by a density distribution $\rho(\vec{x},t)$, a velocity field $\vec{v}(\vec{x},t)$ and an isotropic thermodynamic *pressure* $p(\vec{x},t)$. Such fluids evolve according to the *hydrodynamic equation of motion*

$$\rho\dot{\vec{v}} = -\nabla p \qquad \text{mit} \qquad \dot{\vec{v}} = \frac{\partial}{\partial t}\vec{v} + (\nabla\cdot\vec{v})\vec{v} \tag{5.19}$$

and obey the *continuity equation*

$$\frac{\partial}{\partial t}\rho = -\nabla(\rho\vec{v})\,. \tag{5.20}$$

The energy-momentum tensor can be derived rigorously from the Lagrangian density $\mathcal{L} = -\rho$ by variation. But here we want to take a more intuitive approach: Imagine that we are in the local rest frame of the fluid. Here, the *mean* velocity of the particles is zero, but the particles still move relative to each other in random directions. Now, if we consider a surface element, it is clear that about the same number of particles will pass from one side to the other and in the opposite direction through the surface. However, as the surface element is an oriented object, the particles are positively counted in *both directions*, because transporting a positive momentum from left to right has the same effect as transporting a negative momentum from right to left. The energy-momentum tensor will thus tell us in its spatial components which total momentum per unit of time passes through the test area. In physics this quantity is denoted as the *pressure* p.

Imagine that every particle in this gas has a randomly distributed velocity vector $\mathbf{u}$, which is distributed with a certain probability density $P(\mathbf{u})$. In the local rest system of the fluid, this distribution is expected to be rotationally symmetric. The energy-

momentum tensor is therefore given by

$$T^{\mu\nu} \;=\; \int D\mathbf{u}\, P(\mathbf{u})\, \rho u^\mu u^\nu \,, \tag{5.21}$$

where $D\mathbf{u}$ is a suitable integration measure. Since $P(\mathbf{u})$ is invariant under the reflection of one of its components $u^\mu \to -u^\mu$, it is clear that this tensor has to be diagonal in the rest frame of the fluid. More specifically, the temporal diagonal element is $T^{00} = \rho c^2$, while the spatial diagonal elements have to be proportional to the pressure. It turns out that the proportionality constant is equal to 1. The energy-momentum tensor of a perfect fluid therefore takes on the form

$$T^{\mu\nu} \;=\; \begin{pmatrix} \rho c^2 & & & \\ & p & & \\ & & p & \\ & & & p \end{pmatrix} \tag{5.22}$$

If one is *not* in the local rest frame, but in another inertial frame moving with the 4-speed $\mathbf{u}$ relative to the observer, one can obtain the corresponding tensor by a simple Lorentz transformation of the matrix given above (exercise). The result is:

$$\boxed{T^{\mu\nu} = (\rho + p) u^\mu u^\nu + p g^{\mu\nu} \,.} \tag{5.23}$$

How the pressure actually depends on the density is determined by the *equation of state* and the specific symmetries of the matter distribution. Some special cases are listed in the following table:

Dust:	$p = 0$
Non-relativistic gas:	$p \propto \rho$
Ultra-relativistic gas:	$p = \frac{1}{3}\rho$
Non-relativistic gas of fermions:	$p \propto \rho^{5/3}$
Ultra-relativistic gas of fermions:	$p \propto \rho^{4/3}$
Vacuum energy (cosmological constant):	$p = -\rho$

Electromagnetic field

The energy-momentum tensor of the electromagnetic field can be calculated with relatively little effort directly by applying the formula Eq. (5.15). First, it should be noted that the Lagrangian of the electromagnetic field

$$\mathcal{L}_{EM} \;=\; -\frac{1}{4} F_{\alpha\beta} F^{\alpha\beta} \;=\; -\frac{1}{4} g^{\alpha\beta} g^{\mu\nu} F_{\mu\alpha} F_{\mu\beta} \tag{5.24}$$

depends only on the components of the metric tensor, but not on its partial derivatives. Therefore we have

$$T_{\mu\nu} \;=\; -\frac{2}{\sqrt{-g}} \frac{\partial(\sqrt{-g}\mathcal{L}_{EM})}{\partial g^{\mu\nu}} \;=\; -\frac{2\mathcal{L}_{EM}}{\partial g^{\mu\nu}} - \frac{\mathcal{L}_{EM}}{g} \frac{\partial g}{\partial g^{\mu\nu}} \,. \tag{5.25}$$

Using Eq. (1.112) we can calculate the last term and we get

$$T_{\mu\nu} = -2\frac{\partial \mathcal{L}_{EM}}{\partial g^{\mu\nu}} + g_{\mu\nu}\mathcal{L}_{EM} \tag{5.26}$$

By inserting the Langrangian density (5.24) one is led to

$$\boxed{T_{\mu\nu} = F^{\alpha}{}_{\mu}F_{\alpha\nu} - \frac{1}{4}g_{\mu\nu}F_{\alpha\beta}F^{\alpha\beta}\,.} \tag{5.27}$$

This tensor is traceless (that is, $T^{\mu}{}_{\mu} = 0$), so that the pressure and energy density of the electromagnetic radiation are given by the equation of state $\rho = \frac{1}{3}p$. Electromagnetic radiation thus behaves like an ultra-relativistic gas. This is reasonable because electromagnetic radiation propagates relativistically with the velocity of light.

Summary: The most important formulas of general relativity in coordinate representation:

$$\Gamma^{\alpha}{}_{\mu\nu} = \frac{1}{2}g^{\alpha\beta}(g_{\beta\mu,\nu} + g_{\beta\nu,\mu} - g_{\mu\nu,\beta})$$
$$\ddot{x}^{\alpha} + \Gamma^{\alpha}{}_{\mu\nu}\dot{x}^{\mu}\dot{x}^{\nu} = 0$$
$$R^{\mu}{}_{\nu\alpha\beta} = \Gamma^{\mu}{}_{\nu\beta,\alpha} - \Gamma^{\mu}{}_{\nu\alpha,\beta} + \Gamma^{\rho}{}_{\nu\beta}\Gamma^{\mu}{}_{\rho\alpha} - \Gamma^{\rho}{}_{\nu\alpha}\Gamma^{\mu}{}_{\rho\beta}\,; \qquad R_{\mu\nu} = R^{\rho}{}_{\mu\rho\nu}$$
$$R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} + \Lambda g_{\mu\nu} = \frac{8\pi G}{c^4}T_{\mu\nu} \qquad R_{\mu\nu} = \Lambda g_{\mu\nu} + \frac{8\pi G}{c^4}\left(T_{\mu\nu} - \frac{1}{2}Tg_{\mu\nu}\right)$$
$$T^{\mu\nu} = (\rho + p)u^{\mu}u^{\nu} + pg^{\mu\nu}$$

5.2.5 Weak field approximation

A *week gravitational field* differs only slightly from the flat Minkowski metric. In this case we use the ansatz

$$g_{\mu\nu}(\mathbf{x}) = \eta_{\mu\nu} + h_{\mu\nu}(\mathbf{x})\,, \tag{5.28}$$

where the symmetric tensor field $h_{\mu\nu}(\mathbf{x})$ and its partial derivatives are assumed to be small. The aim is to linearized the field equations to first order in h. Of course, we have to do it in a way that the gauge invariance of the theory, i.e., the invariance under diffeomorphisms, is not violated.

Linearized field equations

To first order in h one obtains the Christoffel symbols

$$\Gamma^{\alpha}{}_{\mu\nu} = \frac{1}{2}\eta^{\alpha\beta}(h_{\beta\mu,\nu} + h_{\beta\nu,\mu} - h_{\mu\nu,\beta}) + \mathcal{O}(h^2)\,. \tag{5.29}$$

In Riemanns curvature tensor, only the first two terms contribute in a first-order approximation, giving an expression in terms of second-order partial derivatives:

$$R^{\mu}{}_{\nu\alpha\beta} = \Gamma^{\mu}{}_{\nu\beta,\alpha} - \Gamma^{\mu}{}_{\nu\alpha,\beta} + \mathcal{O}(h^2) = \frac{1}{2}\Big(h_{\rho\beta,\nu\alpha} - h_{\nu\beta,\rho\alpha} - h_{\rho\alpha,\nu\beta} + h_{\nu\alpha,\rho\beta}\Big) + \mathcal{O}(h^2)\,. \tag{5.30}$$

This results in the Ricci tensor

$$R_{\mu\nu} \;=\; R^{\rho}{}_{\mu\rho\nu} \;=\; \frac{1}{2}\Big(\underbrace{h^{\rho}{}_{\nu,\mu\rho}}_{=h^{\rho}{}_{\nu,\rho\mu}} - \underbrace{h_{\mu\nu,}{}^{\rho}{}_{\rho}}_{=\Box h_{\mu\nu}} - \underbrace{h^{\rho}{}_{\rho,\mu\nu}}_{=h_{,\mu\nu}} + \underbrace{h_{\mu\rho,}{}^{\rho}{}_{\nu}}_{=h^{\rho}{}_{\mu,\rho\nu}} \Big) + \mathcal{O}(h^2)\,, \tag{5.31}$$

hence

$$R_{\mu\nu} \;=\; \frac{1}{2}\left(h^{\rho}{}_{\mu,\rho\nu} + h^{\rho}{}_{\nu,\rho\mu} - \Box h_{\mu\nu} - h_{,\mu\nu}\right) + \mathcal{O}(h^2) \tag{5.32}$$

where $h = h^{\mu}{}_{\mu}$ and were

$$\Box \;=\; \eta^{\alpha\beta}\partial_\alpha\partial_\beta \;=\; \nabla - \partial_t^2 \tag{5.33}$$

is the so-called *d'Alembert operator*, also referred to as the wave operator and frequently denoted by the symbol *Quabla*. Hence the Ricci scalar is given by

$$R \;=\; h^{\mu\nu}{}_{,\mu\nu} - \Box h + \mathcal{O}(h^2) \tag{5.34}$$

The linearized Einstein field equations (without cosmological constant) therefore read

$$h^{\rho}{}_{\mu,\rho\nu} + h^{\rho}{}_{\nu,\rho\mu} - \Box h_{\mu\nu} - h_{,\mu\nu} - \eta_{\mu\nu}(h^{\alpha\beta}{}_{,\alpha\beta} - \Box h) \;=\; \frac{16\pi G}{c^4} T_{\mu\nu}\,, \tag{5.35}$$

where the factor $\frac{1}{2}$ has been moved to the right side. These field equations are gauge-invariant, so are in valid in *any* arbitrary coordinate system, provided that the weak field approximation remains valid.

Remark: Why is the trace $h = h^{\mu}{}_{\mu} \neq 0$? We know that $g^{\mu}{}_{\nu} = \delta^{\mu}_{\nu}$, hence

$$\underbrace{g^{\mu}{}_{\mu}}_{=4} \;=\; \underbrace{\delta^{\mu}{}_{\mu}}_{=4} + \underbrace{h^{\mu}{}_{\mu}}_{=0?}$$

But this conclusion would be premature because we neglect the higher orders of h in this place. All higher orders together would have a vanishing trace, but there is no need for individual orders to have vanishing trace. We know something analogous from quantum mechanics: $\exp(i\mathbf{H}t)$ is unitary, but its first order expansion $\mathbb{1} + i\mathbf{H}t$ is not.

Gauge invariance

As in the theory of electrodynamics, we can use the gauge freedom to bring the linearized field equations into the simplest possible form. For this we consider a diffeomorphism that differs only slightly from an identical map, i.e., an infinitesimal coordinate transformation

$$x^{\mu}(p) \to x^{\mu\prime}(p) = x^{\mu}(p) + \xi^{\mu}(p)\,, \tag{5.36}$$

where $p \in \mathcal{M}$is a space-time event and were the ξ^{μ} are so small that an approximation in first order is justified. Under this assumption we want to examine the transformation behavior expanded to first order both in $h_{\mu\nu}$ and int ξ^{μ}.

According to Eq. (1.100) the metric tensor is changes in a coordinate transformation according to

$$g_{\mu\nu} \;\to\; g'_{\mu\nu} = \frac{\partial x^{\alpha}}{\partial x^{\mu\prime}}\,\frac{\partial x^{\beta}}{\partial x^{\nu\prime}}\,g_{\alpha\beta}\,. \tag{5.37}$$

With $\frac{\partial}{\partial x^{\mu'}} = \frac{\partial}{\partial x^{\mu}} + \mathcal{O}(\xi)$ and with

$$\frac{\partial x^{\alpha}}{\partial x^{\mu'}} = \frac{\partial (x^{\alpha'} - \xi^{\alpha})}{\partial x^{\mu'}} = \delta^{\alpha}_{\mu} - \frac{\partial \xi^{\alpha}}{\partial x^{\mu'}} = \delta^{\alpha}_{\mu} + \xi^{\alpha}{}_{,\mu} + \mathcal{O}(\xi^2)$$

this implies the equation

$$\eta_{\mu\nu} + h_{\mu\nu} \;\to\; \eta_{\mu\nu} + h'_{\mu\nu} \;=\; \left(\delta^{\alpha}_{\mu} - \xi^{\alpha}{}_{,\mu}\right)\left(\delta^{\beta}_{\nu} - \xi^{\beta}{}_{,\nu}\right)\left(\eta_{\alpha\beta} + h_{\alpha\beta}\right), \tag{5.38}$$

hence

$$h'_{\mu\nu} \;=\; h_{\mu\nu} - \xi_{\mu,\nu} - \xi_{\nu,\mu}\,. \tag{5.39}$$

As expected, the components of $h_{\mu\nu}$ are not gauge-invariant. Therefore we can choose a particular gauge in such a way that the divergence of $h_{\mu\nu} - \frac{1}{2}h\eta_{\mu\nu}$ vanishes, i.e.,

$$h^{\mu}{}_{\nu,\mu} - \frac{1}{2}h^{\mu}{}_{\mu,\nu} = 0\,. \tag{5.40}$$

Proof: Suppose the left side is not equal to zero. Under an infinitesimal coordinate transformation and using Eq. (5.39), the left hand side turns into

$$g^{\mu\rho}\left(h_{\mu\nu,\rho} - \frac{1}{2}h_{\mu\rho,\nu} - \xi_{\mu,\nu\rho} + \frac{1}{2}\xi_{\mu,\rho\nu} - \xi_{\nu,\mu\rho} + \frac{1}{2}\xi_{\rho,\mu\nu}\right) \;=\; h^{\mu}{}_{\nu,\mu} - \frac{1}{2}h^{\mu}{}_{\mu,\nu} - \Box\xi_{\nu}\,,$$

were the terms marked in red cancel each other. In order to get rid of the sternum, we have to solve the wave equation $\Box\xi_{\nu} = h^{\mu}{}_{\nu,\mu} - \frac{1}{2}h^{\mu}{}_{\mu,\nu}$ for the displacement ξ.

In the so-called *Lorenz gauge*[1] The Ricci tensor simplifies to $R_{\mu\nu} = -\frac{1}{2}\Box h_{\mu\nu}$, i.e., the linearized field equations read

$$\boxed{-\Box h_{\mu\nu} + \frac{1}{2}\eta_{\mu\nu}\Box h = \frac{16\pi G}{c^4}T_{\mu\nu}} \qquad \text{bzw.} \qquad \boxed{\Box h_{\mu\nu} = -\frac{16\pi G}{c^4}\left(T_{\mu\nu} - \frac{1}{2}T\eta_{\mu\nu}\right).} \tag{5.41}$$

5.2.6 Newtonian limit

Let us now compare the approximated field equations with Newton's theory. To this end we first list the corresponding elements in both theories:

	Newtonian gravity	Einsteinian gravity
Field equations	$\nabla^2\Phi = 4\pi G\rho$	$\Box h_{\mu\nu} = -\frac{16\pi G}{c^4}\left(T_{\mu\nu} - \frac{1}{2}T\eta_{\mu\nu}\right)$
Equations of motion	$\ddot{\vec{x}} = -\nabla\Phi$	$\ddot{x}^{\mu} + \Gamma^{\mu}_{\alpha\beta}\dot{x}^{\alpha}\dot{x}^{\beta} = 0$

In the Newtonian limit we have $v \ll c$, hence the 4-velocity $\dot{x}^{\mu} = \gamma(c, \vec{v})$ is dominated by its temporal component. For fixed $i = 1, \ldots, 3$, the spatial equations of motion

[1] The Lorenz gauge is named after the Danish physicist Ludvig Lorenz, not to be confused with Hendrik A. Lorentz, who discovered the Lorentz transformation.

expanded to lowest order therefore take the form

$$\ddot{x}^i \;=\; -\Gamma^i{}_{00}\underbrace{\dot{x}^0\dot{x}^0}_{\approx c^2}, \tag{5.42}$$

where

$$\Gamma^i_{00} \;=\; \frac{1}{2}\eta^{i\beta}\left(2h_{\beta 0,0} - h_{00,\beta}\right) \tag{5.43}$$

Assuming that the gravitational field (i.e., the metric) is time-independent, the first term vanishes. This reduces the equation of motion as follows:

$$\ddot{x}^{\,i} = \frac{c^2}{2}\,h_{00,i}\,. \tag{5.44}$$

The right-hand side of this equation of motion is now linked to the energy-momentum tensor by using the field equation. To this end we consider the 00-component of the Ricci tensor:

$$R_{\mu 0\alpha 0} \;=\; \frac{1}{2}\left(h_{\mu 0,0\alpha} - h_{\mu\alpha,00} - h_{00,\mu\alpha} + h_{0\alpha,\mu 0}\right) \;=\; -\frac{1}{2}h_{00,\mu\alpha} \tag{5.45}$$

where we have only one term without time derivatives (time-independence). The 00-component of the Ricci tensor therefore reads

$$\Rightarrow R_{00} \;=\; -\frac{1}{2}h_{00,\ \alpha}^{\ \ \alpha} \;=\; -\frac{1}{2}\sum_{i=1}^{3}\frac{\partial^2}{\partial {x^i}^2}h_{00} \tag{5.46}$$

If we insert Eq. (5.44) into the last equation, we obtain

$$R_{00} \;=\; -\frac{1}{c^2}\sum_{i=1}^{3}\frac{\partial \ddot{x}^{\,i}}{\partial x^i} \;=\; -\frac{1}{c^2}\,\nabla\cdot\ddot{\vec{x}}\,. \tag{5.47}$$

For the right side of the field equation we want to assume a matter distribution in the form of dust whose pressure is zero. In the rest frame of the dust the energy-momentum tensor is therefore dominated by the element $T_{00} = \rho c^2$. The field equation $R_{00} = \frac{8\pi G}{c^4}(T_{00} - \frac{1}{2}\eta_{00}T)$ therefore reads

$$-\frac{1}{c^2}\,\nabla\cdot\ddot{\vec{x}} \;=\; T_{00} - \frac{1}{2}g_{00}T \;=\; \frac{1}{2}T_{00} \;=\; \frac{\rho c^2}{2} + \mathcal{O}(h)\,. \tag{5.48}$$

This implies the Newtonian equation of motion

$$\nabla\cdot\ddot{\vec{x}} \;=\; -4\pi G\rho\,, \tag{5.49}$$

and this finally justifies the choice of the coupling constants in Eq. (5.9).

6 Advanced Formulations of General Relativity

As you may have noticed in the last two last chapters, we did not use differential forms, but we derived the field equations in the conventional old-fashioned index notation. In many cases the index notation is in fact the basis for practical applications. However, differential forms become more and more important. For this reason, this chapter is concerned with advanced formulations of GR, which is also conceptually interesting.

6.1 Differential geometry without metric tensor

The traditional and intuitive way of introducing General Relativity starts with the definition of a metric g which is a field of symmetric 2-forms with the coordinate representation $g_{\mu\nu}(x)$. The components of the metric tensor are considered as the elementary degrees of freedom, so to say as the gravitational field itself. In fact, the Einstein field equations have been derived by varying the action (5.4) with respect to $g_{\mu\nu}$. Everything can be derived from the metric, in particular the connection coefficients, called *Christoffel symbols* (cf. Eq. (3.33) on page 95)

$$\Gamma^{\alpha}{}_{\mu\nu} = \frac{1}{2} g^{\alpha\beta}(g_{\beta\mu,\nu} + g_{\beta\nu,\mu} - g_{\mu\nu,\beta})\,. \tag{6.1}$$

which in turn determines the curvature tensor. As can be seen, this formula forces the Christoffel symbols to be symmetric in the lower indices μ, ν. However, as pointed out by various mathematicians, in particular by Élie Cartan, the metric is not as fundamental as it seems. In fact, it turns out that the connection is probably more fundamental than the metric itself. In fact, Cartan showed that there exist manifolds with certain connections (so-called *Cartan connections*) which are not compatible with a metric.

Élie Cartan (1869-1951) [Wikimedia]

6.1.1 Torsion

The generalized manifolds studied by Cartan are defined by connections that are generally *not* symmetric in the lower indices. It turns out that such spaces differ from ordinary metric ones in so far as they exhibit *torsion*. To understand the notion of tor-

sion in differential geometry is not easy because there is no obvious counterpart in our everyday experience.

To get some intuitive understanding, imagine that you are moving freely in space, following a geodesic line. Suppose that your friend is also moving freely at some distance, positioned a few meters to the right of you, going into the same direction at the same speed, meaning that you are initially flying in parallel.

How does gravitation (or more generally, the curvature of the manifold) affect the position of your friend seen from your perspective? You would expect the distance to increase for negative and to decrease for positive curvature. But you would not expect to find your friend at some point on the left side, swirling around you like a spiral. But why not? This is exactly what happens in spaces with *torsion*.

The impact of antisymmetric contributions

To get some more insight, let us start with an ordinary torsion-free connection ${\Gamma^\alpha}_{\mu\nu}$ in some metric space, which is symmetric in the lower indices. Now let us modify this metric by

$$ {\Gamma^\alpha}_{\mu\nu} \;\to\; {\tilde{\Gamma}^\alpha}_{\mu\nu} = {\Gamma^\alpha}_{\mu\nu} + {A^\alpha}_{\mu\nu} \tag{6.2} $$

with some totally asymmetric contribution ${A^\alpha}_{\mu\nu} = -{A^\alpha}_{\nu\mu}$. How does this contribution affect the geodesic lines on the manifold?

To answer this question, recall the geodesic equation

$$ \nabla_{\mathbf{u}}\mathbf{u} = 0 \qquad \Leftrightarrow \qquad \ddot{x}^\alpha + {\Gamma^\alpha}_{\mu\nu}\dot{x}^\mu\dot{x}^\nu = 0\,. \tag{6.3} $$

Since $x^\mu\dot{x}^\nu$ is symmetric in the indices, it is easy to see that adding the antisymmetric part added in (6.2) has absolutely no effect. Therefore, the two spaces described by the connection coefficients Γ and $\tilde{\Gamma}$ have exactly the same geodesic structure.

At first glance this seems to be surprising. Knowing all the geodesics does not fully determine the connection? Why? If we have a closer look, however, it is clear that geodesics are special in so far as they describe parallel transport of a tangent vector in its own direction which is special. In fact, parallel transport of any other vector *does* feel the antisymmetric contribution.

...to be continued

6.2 Tetrad fields

Imagine that space-time is permeated by an infinite number of fictitious (i.e. massless, non-interacting) observers, all of whom are in free fall. Each of these observers is located locally in an inertial system in which they can define a local coordinate system with a Minkowski metric that they carry with them during free fall.

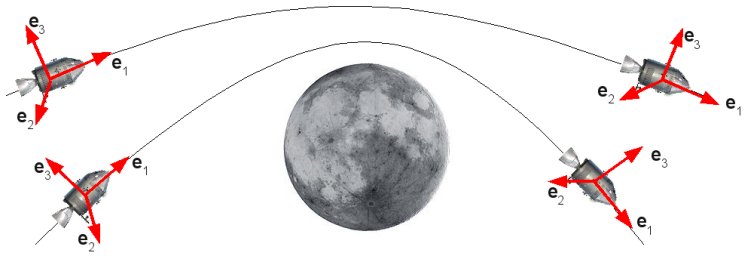


Figure 6.1: The tetrad frame field: Transport of a local orthogonal coordinate system in free fall (see text).

6.2.1 Tetrad basis

In other words, any astronaut in a non-powered spaceship can define a local ct, x, y, z coordinate system with a flat metric in his spaceship and carry it with him during the weightless flight (see reffigframe). Such a locally flat coordinate system is known in mathematics as a *frame*. The corresponding basic vector fields

$$\{\mathbf{e}_I\} \;=\; \{\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\} \qquad \mathbf{e}_I \in T\mathcal{M}, \qquad I = 0, \ldots, 3 \tag{6.4}$$

are called *frame fields* or *vierbein fields*. To distinguish them they are usually indexed with Latin letters, while Greek indices are reserved for the usual coordinate basis.

Convention:
Use Latin indices $a, b, c, d, \ldots$ for the tetrad basis.
Use Greek indices μ, ν, ρ... for the usual coordinate basis.

A tetrad is therefore defined in such a way that it specifies a local coordinate system for each trajectory, in which the metric tensor looks locally like a Minkowski metric:

$$\boxed{\mathbf{g}(\mathbf{e}_a, \mathbf{e}_b) = \eta_{ab}\,.} \tag{6.5}$$

The vector $\mathbf{e}_0$ is always oriented in such a way that it describes the observer's eigenzeit. The other vectors form an orthogonal tripod, which is determined up to spatial rotation. This means that the spatial tripod can thus be aligned differently on each trajectory, but only in such a way that the entire vector field remains continuously differentiable.

Tetrads in accelerated frames of reference

The concept of tetrads is not limited to free fall, but can also be meaningfully defined for accelerated observers. Even if the astronaut ignites the rocket propulsion, he can still define a locally flat coordinate system so that the metric tensor takes the form of a Minkowski metric in its location. The same applies to an earth inhabitant that rotates and is subject to gravitational acceleration.

Acceleration at a given point is determined *not* by the metric tensor, but by its derivatives.

In the following we can therefore imagine that space-time is interspersed with non-intersecting trajectories that can be of any shape. On each of these trajectories a tetrad is transported, which defines a local Minkowski metric (the laboratory coordinates of the astronaut, so to speak) at each point. The vector $\mathbf{e}_0$ points in the temporal direction, while the remaining spatial tripod is lives in the spatial $\mathbb{R}^3$ where it is determined up to rotation. Unlike a vector transported in parallel, the tripod can therefore rotate along the trajectory. In this case one speaks of a *spinning tetrad*.

With this more general interpretation it is e.g. possible to use tetrad formalism to describe a fictitious observer who hovers over the edge of a black hole with his spaceship, provided that his rocket propulsion is strong enough.

Hoovering spacecraft [Wikimedia]

Representation of tangent vectors in tetrad coordinates

The tetrad field $\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ provides a local basis in every point of spacetime. A tangent vector field $\mathbf{X} \in T\mathcal{M}$ can thus be represented by coordinates in this basis via

$$\mathbf{X} \;=\; X^a \mathbf{e}_a \,, \tag{6.6}$$

whereby the latin indices are added from 0 to 3 as usual. In order to calculate the components X^a of the vector field, we consider the tetrad dual 1-forms $\mathbf{e}^b$, called *coframes*, which fulfill the usual definition property

$$\mathbf{e}^b(\mathbf{e}_a) \;=\; \delta^b_a \,. \tag{6.7}$$

Then $\mathbf{e}^b(\mathbf{X}) = X^a \mathbf{e}^b(\mathbf{x}_a) = X^a \delta^b_a$, hence

$$X^a \;=\; \mathbf{e}^a(\mathbf{X}) \,. \tag{6.8}$$

As we will see below, the 1-forms $\mathbf{e}^a$ represent the gravitational field.

Switching between tetrad and ordinary coordinates

The tetrad basis $(\mathbf{e}^a, \mathbf{e}_a)$ and the usual coordinate basis $(\,\mathrm{d}x^\mu, \partial_\mu)$ are often used side by side. To distinguish them, the components are denoted by Latin or Greek letters. Since they are noting but two different choices of a basis, there will be a (position-dependent) basis transformation which allows the components to be converted.

First, as any vector field, the tetrad vector fields can be represented in a given coordinate base in components:

$$\mathbf{e}_a = e^\mu_a \partial_\mu \,, \qquad \mathbf{e}^b = e^b_\mu \,\mathrm{d}x^\mu \,. \tag{6.9}$$

Conversely, the coordinates basis vectors can also be also be represented by components in the tetrad basis. Because of $\mathbf{e}^b(\mathbf{e}_a) \;=\; \delta^b_a$ the transformation matrices are the same, i.e.

$$\partial_\mu = e^a_\mu \mathbf{e}_a \,, \qquad \mathrm{d}x^\mu = e^\mu_a \mathbf{e}^a \,, \tag{6.10}$$

where the two transformation matrices are mutually inverse:

$$e^a_\mu e^\mu_b = \delta^a_b\,, \qquad e^a_\mu e^\nu_a = \delta^\nu_\mu\,. \tag{6.11}$$

This makes it easy to switch between coordinate and tetrad representation:

Object	Coord. rep.	Tetrad rep.	Coordinate ↔ Tetrad	
Vector field $\mathbf{X}$	$\mathbf{X} = X^\mu \partial_\mu$	$\mathbf{X} = X^a \mathbf{e}_a$	$X^a = e^a_\mu X^\mu$	$X^\mu = e^\mu_a X^a$
1-form field α	$\alpha = \alpha_\mu \,\mathrm{d}x^\mu$	$\alpha = \alpha_a \mathbf{e}^a$	$\alpha_a = e^\mu_a \alpha_\mu$	$\alpha_\mu = e^a_\mu \alpha_i$

How the metric tensor is encoded in the tetrad

In the traditional formulation of GR, the action $S = \int (R - 2\Lambda)\,\mathrm{d}\omega$ is varied with respect to the components of the metric tensor, i.e. the components $g_{\mu\nu}$ are interpreted as the elementary degrees of freedom of the gravitational field. In tetrad formalism, however, it is the 1-forms $\mathbf{e}^a$ which are interpreted as elementary degrees of freedom. Reading textbooks about the subject, it is often said that these 1-forms are to a certain extent the square root of the metric tensor. What is it all about?

The starting point is the observation that $G = \{g_{\mu\nu}\}$ is a real symmetric (hermitean) matrix. It therefore has real eigenvalues $\lambda_0, \lambda_1, \lambda_2, \lambda_3$ and pairwise orthogonal eigenvectors, which we shall write here in Dirac notation $|0\rangle, |1\rangle, |2\rangle, |3\rangle$. Having solved the eigenvalue problem $G|K\rangle = \lambda|K\rangle$, we can express the metric tensor in the *spectral representation*

$$G = \sum_{K=0}^{3} \lambda_K\, |K\rangle\langle K|\,, \tag{6.12}$$

where we assumed the eigenvectors to be normalized. Because of the signature of the metric, one of the eigenvalues (say λ_0) is negative, while the other three eigenvalues are positive. One can now absorb the eigenvalues by introducing *non-normalized* eigenvectors

$$|e_K\rangle = \sqrt{|\lambda_K|}\,|K\rangle \tag{6.13}$$

such taht

$$G = \sum_{K=0}^{3} \mathrm{sign}(\lambda_K)\; |e_K\rangle\langle e_K|\,. \tag{6.14}$$

In this way, the matrix of the metric tensor (except for the sign) is expressed as a sum of projectors (dyadic products). The scalar product of two tangent vectors $\mathbf{X}$ and $\mathbf{Y}$ is then

$$\mathbf{g}(\mathbf{X}, \mathbf{Y}) = \langle X|G|Y\rangle = \sum_{K=0}^{3} \mathrm{sign}(\lambda_K)\; \langle X|e_K\rangle\langle e_K|Y\rangle\,. \tag{6.15}$$

The scalar products occurring in this expression can be interpreted as the result of applying the 1-forms $\mathbf{e}^K$, namely

$$\langle X|e_K\rangle = \langle e_K|X\rangle = \mathbf{e}^K(X)\,, \qquad \langle e_K|Y\rangle = \mathbf{e}^K(Y)\,, \tag{6.16}$$

hence

$$\mathbf{g}(\mathbf{X},\mathbf{Y}) = \sum_{K=0}^{3} \operatorname{sign}(\lambda_K)\ \mathbf{e}^K(X)\mathbf{e}^K(Y)\,. \tag{6.17}$$

Because of $\operatorname{sign}(\lambda_0) = -1$ und $\operatorname{sign}(\lambda_{1,2,3}) = +1$ we may also write

$$\boxed{\mathbf{g}(\mathbf{X},\mathbf{Y}) = \eta_{ab}\ \mathbf{e}^a(X)\mathbf{e}^b(Y)\,.} \tag{6.18}$$

From this calculation we can see that the tetrad vectors or the dual 1-forms are essentially the eigenvectors of the metric tensor.

Tetrad basis $\simeq$ Eigenvectors of the metric tensor.

6.2.2 GR formulated in the tetrad formalism

6.2.3 GR formulated in the tetrad formalism

The gravitational field

In the tetrad formalism, the gravitational field is a four-vector-valued 1-form

$$\mathbf{e}^a \;=\; \mathbf{e}^a_\mu\ \mathrm{d}x^\mu\,. \tag{6.19}$$

This 1-form can be used to represent tangent vectors $\mathbf{X} \in T\mathcal{M}$ in a local Minkowski space (free fall frame) by $X^a = \mathbf{e}^a(\mathbf{X})$. Note again that the Latin indices $a, b, c, \ldots = 0, 1, 2, 3$, also called *Lorentz indices*, denote the components of the tangent vector in the flat Minkowski basis. Therefore, they are raised and lowered with η_{ab}. This is in contrast with the components of the same tangent vector in the coordinate basis denoted by Greek indices, which are raised and lowered with $g_{\mu\mu}$

> **Remember:** Tentrad components (Latin indices) are raised/lowered with η^{ab} and η_{ab}.
> Ordinary components (Greek indices) are raised/lowered with $g^{\mu\nu}$ and $g_{\mu\nu}$.

Tetrad connection

The connection $\nabla_{\mathbf{X}}$ applied to $\mathbf{Y}$ is defined as the rate of change of the vector field $\mathbf{Y}$ relative to a vector carried in parallel when moving in the direction of $\mathbf{X}$ (see ref-pagesec:connections). This connection is represented in the local tetrad base as a so(3.1)-Lie-algebra-valued 1-form

$$\omega^a{}_b \;=\; \omega_\mu{}^a{}_b\ \mathrm{d}x^\mu \tag{6.20}$$

such that

$$\mathbf{e}^a \nabla_{\mathbf{X}} \mathbf{Y} \;=\; [\nabla_{\mathbf{X}} \mathbf{Y}]^a \;=\; \omega^a{}_b \mathbf{X} Y^b = \omega_\mu{}^a{}_b X^\mu Y^b\,. \tag{6.21}$$

This so-called *spin connection* in the tetrad representation does exactly the same job as the Christoffel symbols in the coordinate representation. The spin connection is anti-

symmetric in the Lorentz indices a, b, provided that they both placed either above or below:

$$\omega^{ab} = -\omega^{ba}\,, \qquad \omega_{ab} = -\omega_{ba}\,. \tag{6.22}$$

This antisymmetry is a fundamental property of generators in the Lie algebra of the Lorentz group, similar to the antisymmetry of generators of the rotation group in $\mathbb{R}^3$.

Covariant derivative

The spin connection defines a *covariant derivative* D or likewise a *covariant partial derivative* D_μ, which acts on tensors with Lorentz indices. For example, if $\mathbf{X}$ is a vector field in Lorentz representation X^a, then

$$\begin{aligned} D\mathbf{X} &= \mathrm{d}\mathbf{X} + \omega\mathbf{X} \\ DX^a &= \mathrm{d}X^a + \omega^a{}_b X^b \end{aligned} \tag{6.23}$$

Applying this Lorentz-vector-valued 1-form to the coordinate basis vector $\mathbf{e}_\mu$ we get

$$D_\mu X^a = \partial_\mu X^a + \omega_\mu{}^a{}_b X^b\,. \tag{6.24}$$

Analogously, the covariant partial derivative of a rank-2 tensor is given by

$$D_\mu T^{ab} = \partial_\mu T^{ab} + \omega_\mu{}^a{}_c T^{cb} + \omega_\mu{}^b{}_c T^{ac}\,. \tag{6.25}$$

Similarly one defines a covariant derivative acting on *forms* with Lorentz indices. For example, if α^a is a 4-vector-valued 1-form, then

$$D\alpha^a := \mathrm{d}\alpha^a + \omega^a{}_b \wedge \alpha^b\,. \tag{6.26}$$

and

$$D_\mu \alpha^a := \mathrm{d}\alpha^a(\partial_\mu) + \omega_\mu{}^a{}_b \wedge \alpha^b\,. \tag{6.27}$$

Torsion

The *torsion tensor* is a vector-valued 2-form $\mathbf{T}$ acting on two vector fields $\mathbf{X}, \mathbf{Y}$ by

$$\mathbf{T}(\mathbf{X}, \mathbf{Y}) = \nabla_{\mathbf{X}}\mathbf{Y} - \nabla_{\mathbf{Y}}\mathbf{X} - [\mathbf{X}, \mathbf{Y}]. \tag{6.28}$$

where $[\mathbf{X}, \mathbf{Y}] = \mathbf{X}\cdot\mathbf{Y} - \mathbf{Y}\cdot\mathbf{X}$ is the *Lie bracket* (c.f. Sect. 2.4.7 on page 73).

Representing the result vector in the tetrad basis, one can interpret the torsion tensor as a 4-component 2-form $\mathbf{T}^a$ with $a = 0, \ldots, 3$. One can show that

$$\mathbf{T}^a = D\mathbf{e}^a\,. \tag{6.29}$$

The torsion tensor provides information as to whether a tangential vector rotates in the tangential space during parallel transport. The spacetime of conventional GR is torsion-free, i.e.

$$\mathbf{T} = 0.$$

It can be shown that there is exactly one torsion-free spinor connection for a given tetrad field, which in this formalism represents the counterpart to the Levi-Civitá connection (see page **??**). Explicitly it is given by the rather complicated formula

$$\omega_\mu{}^{ab} = 2\mathbf{e}^{\nu[a}\partial_{[\mu}\mathbf{e}^{b]}_{\nu]} + \mathbf{e}_{\mu c}\mathbf{e}^{\nu a}\mathbf{e}^{\sigma b}\partial_{[\sigma}\mathbf{e}^{c}_{\nu]}\,, \tag{6.30}$$

where the square brackets stand for cyclic permutation. The connection is therefore involves up to the fourth power of the gravitational field.

Remark: The torsion-free spin connection can be calculated from the Christoffel symbols by

$$\omega_\mu{}^a{}_b = \mathbf{e}^\nu_j\left(\partial_\mu \mathbf{e}^a_\nu - \Gamma^\rho_{\mu\nu}\mathbf{e}^a_\rho\right).$$

Curvature

The curvature tensor is a Lorentz-algebra-valued 2-form

$$R^a{}_b \;=\; R^a{}_{b\mu\nu}\,\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu \tag{6.31}$$

defined by

$$\boxed{R^a{}_b \;=\; \mathrm{d}\omega^a{}_b + \omega^a{}_c \wedge \omega^c{}_b\,.} \tag{6.32}$$

This curvature tensor maps two tangent vectors (connected via the indices μ, ν) onto a 4×4 matrix. This matrix describes the rate at which a tangential vector, represented in the tetrad base, changes when it is transported on a closed path spanned by the two vectors.

Remark:
As an exercise show that $D^2u^a = R^a{}_b \wedge u^b$ and that a vanishing torsion implies the relation $R^a{}_b \wedge \mathbf{e}^b = 0$.

Action and the field equations in vacuum

For vanishing cosmological constant, the action of GR in the tetrad formalism reads

$$S[\mathbf{e}, \omega] = \frac{1}{16\pi G}\int \epsilon_{abcd}\mathbf{e}^a \wedge \mathbf{e}^b \wedge R^{cd} \tag{6.33}$$

where R implicitly depends on the spin connection ω. The four field equations are then

$$\epsilon_{abcd}\,\mathbf{e}^a \wedge R^{bc} \;=\; 0 \tag{6.34}$$

These field equations can be brought into a more traditional form by defining the Ricci tensor

$$R^a_\mu = R^{ab}{}_{\mu\nu}\mathbf{e}^\nu_b \tag{6.35}$$

and the Ricci scalar

$$R = R^a_\mu \mathbf{e}^\mu_a. \tag{6.36}$$

This leads to

$$R^a_\mu - \frac{1}{2} R e^a_\mu = 0\,. \tag{6.37}$$

7 Applications in Astrophysics

7.1 Schwarzschild solution

We owe Karl Schwarzschild (1873-1916) the simplest but perhaps most important exact solution of Einstein's field equations. When World War I broke out in 1914, he – like many German Jews of the time – volunteered for the German army. During his service at the front he managed to find time to work on exciting problems in Physics, and what he found in 1915 is the *Schwarzschild metric* and the *Schwarzschild radius* named after him. Only a little bit later he returned to Germany as an invalid and eventually died in 1916.

Karl Schwarzschild

The *Schwarzschild solutions* are based primarily on the assumption of radial symmetry and are therefore suitable for describing stars, neutron stars and black holes, but they are also used for simple cosmological models. Similar to Newtonian theory, in which the course of the gravitational field inside and outside a star is considered separately, there is an *exterior* and an *interior* Schwarzschild metric. We will first have a look at the exterior Schwarzschild metric.

7.1.1 Schwarzschild metric in vacuum

The *exterior Schwarzschild metric* is a radially symmetric solution of Einstein's field equations in vacuum $R_{\mu\nu} = 0$. Starting point is the observation that the flat Minkowski metric $\eta_{\mu\nu}$ can be written in spherical coordinates $t, \tilde{r}, \theta, \phi$ as the line lement

$$\mathrm{d}s^2 = -\,\mathrm{d}t^2 + \,\mathrm{d}\tilde{r}^2 + \tilde{r}^2(\,\mathrm{d}\theta^2 + \sin^2\theta\,\,\mathrm{d}\phi^2)\,. \tag{7.1}$$

A possible approach would be to multiply each of these terms by a function which depends only on the radius $\tilde{r}$:

$$\mathrm{d}s^2 = -f(\tilde{r})\,\mathrm{d}t^2 + g(\tilde{r})\,\mathrm{d}\tilde{r}^2 + h(\tilde{r})\tilde{r}^2(\,\mathrm{d}\theta^2 + \sin^2\theta\,\,\mathrm{d}\phi^2)\,. \tag{7.2}$$

Only two of these three functions are independent, since you can rescale the radial coordinate by[1]

$$r = \tilde{r}\sqrt{h(\tilde{r})}\,. \tag{7.3}$$

[1] In order for the signature to be preserved, the function h must meet certain requirements that are ignored here.

So we can use the following ansatz (setting $c = 1$)

$$\mathrm{d}s^2 = -B(r)\,\mathrm{d}t^2 + A(r)\,\mathrm{d}r^2 + r^2(\,\mathrm{d}\theta^2 + \sin^2\theta\;\mathrm{d}\phi^2)\,, \tag{7.4}$$

where $A(r)$ and $B(r)$ are positive functions that must be determined by solving the field equations.

Solution of the field equations

In the following it will be useful to represent the two functions by exponential functions as $A(r) = e^{\alpha(r)}$ and $B(r) = e^{\beta(r)}$. The metric tensor is diagonal with the components

$$g_{00} = g_{tt} = -e^{\beta}\,, \qquad g_{11} = g_{rr} = e^{\alpha}\,, \qquad g_{22} = g_{\theta\theta} = r^2\,, \qquad g_{33} = g_{\phi\phi} = r^2\sin^2\theta \tag{7.5}$$

and its inverse is simply given by $g^{\mu\nu} = (g_{\mu\nu})^{-1}$. Because of the symmetry, the second and third terms in the Christoffel symbols (cf. Eq. (3.33) on page 95) cancel each other. The only non-vanishing Christoffel symbols are therefore

$$\begin{gathered}
\Gamma^0{}_{01} = \Gamma^0{}_{10} = \frac{1}{2}\beta'\,, \qquad \Gamma^1{}_{00} = \frac{1}{2}\beta' e^{\beta-\alpha}\,, \qquad \Gamma^1{}_{11} = \frac{1}{2}\alpha'\,, \\
\Gamma^1{}_{22} = -re^{-\alpha}\,, \qquad \Gamma^1{}_{33} = -re^{-\alpha}\sin^2\theta\,, \qquad \Gamma^2{}_{12} = \Gamma^2{}_{21} = 1/r\,, \\
\Gamma^2{}_{33} = -\sin\theta\cos\theta\,, \qquad \Gamma^3{}_{13} = \Gamma^3{}_{31} = 1/r\,, \qquad \Gamma^3{}_{23} = \Gamma^3{}_{32} = 1/\tan\theta.
\end{gathered} \tag{7.6}$$

This results in the curvature tensor (see Eq. (3.59) on page 101) with the non-vanishing elements

$$\begin{aligned}
R^0{}_{101} &= -\frac{1}{2}\beta'' - \frac{1}{4}\beta'^2 + \frac{1}{4}\alpha'\beta' \\
R^0{}_{202} &= -\frac{1}{2}re^{-\alpha}\beta' \\
R^0{}_{303} &= -\frac{1}{2}re^{-\alpha}\beta'\sin^2\theta \\
R^1{}_{212} &= -\frac{1}{2}re^{-\alpha}\alpha' \\
R^1{}_{313} &= -\frac{1}{2}re^{-\alpha}\alpha'\sin^2\theta \\
R^2{}_{323} &= (1-e^{-\alpha})\sin^2\theta\,.
\end{aligned} \tag{7.7}$$

The Ricci Tensor reads

$$\begin{aligned}
R_{00} &= e^{\beta-\alpha}(\frac{1}{2}\beta'' + \frac{1}{4}\beta'^2 - \frac{1}{4}\alpha'\beta' + \frac{1}{r}\beta')\,, \\
R_{11} &= -\frac{1}{2}\beta'' - \frac{1}{4}\beta'^2 + \frac{1}{4}\alpha'\beta' + \frac{1}{r}\alpha'\,, \\
R_{22} &= 1 + e^{-\beta}(-\frac{1}{2}r\alpha' + \frac{1}{2}r\beta' - 1)\,, \\
R_{33} &= R_{22}\sin^2\theta
\end{aligned} \tag{7.8}$$

and therefore the Ricci scalar is given by

$$R = \frac{e^{-\alpha}}{2r^2}\left(4(e^{\alpha}-1)+r\big((\alpha'-\beta')(4+r\beta')-2r\beta''\big)\right). \tag{7.9}$$

In vacuum we have $R_{\mu\nu}=0$. The first two equations for R_{00} and R_{11} imply

$$\alpha'+\beta'=0 \qquad \Rightarrow \qquad \alpha+\beta=const. \tag{7.10}$$

We now require the Schwarzschild metric to describe a massive object in the center of the coordinate system (e.g. a star or black hole) which means that the Schwarschild metric should tend towards the flat Minkowski metric at large distance, i.e.

$$\lim_{r\to\infty}\alpha=\lim_{r\to\infty}\beta=0 \qquad \Rightarrow \qquad \alpha=-\beta \qquad \Rightarrow \qquad const=0. \tag{7.11}$$

The condition $R_{22}=0$ thus implies the differential equation

$$1-e^{-\beta}+r\beta' = 0 \tag{7.12}$$

with the solution

$$e^{\beta(r)} = 1-\frac{r_s}{r}, \tag{7.13}$$

where r_s is an integration constant. This means that the exterior Schwarzschild metric will be given by

$$\boxed{\mathrm{d}s^2 = -\left(1-\frac{r_s}{r}\right)\mathrm{d}t^2+\left(1-\frac{r_s}{r}\right)^{-1}\mathrm{d}r^2+r^2(\,\mathrm{d}\theta^2+\sin^2\theta\,\mathrm{d}\phi^2)\,.} \tag{7.14}$$

Remark:

- It can be shown that the rotational invariance largely determines the form of the metric tensor. In the present form, the metric is static, that is, the components of the metric tensor are time-independent. However, this does not mean that the problem as such is time-independent, it only means that the metric looks static in the selected coordinates. In fact, the Schwarzschild metric is also capable to describe collapsing and even radially oscillating objects. Note that subsequent coordinate transformation (diffeomorphism) may lead to representations of the Schwarzschild metric with different types of time dependencies.
- The so-called *Birkhoff theorem* states that the external gravitational field of a body with a radially symmetrical mass distribution, similar to Newtonian theory, only depends on the total mass M and that the outer Schwarzschild metric is the only spherically symmetrical, asymptotically flat solution of that kind. In other words, the exterior Schwarzschild metric is exactly the same for a mass point (black hole) or a spherically symmetric mass distribution (star).

Schwarzschild radius

The integration constant r_s is denoted as the *Schwarzschild radius*. To determine it quantitatively, we consider the Schwarzschild metric at a large distance from a star with the mass M located in the center. From the weak field approximation we know that at large

distance the component of the metric

$$g_{00} \approx 1 + h_{00} \approx 1 + \frac{2\Phi}{c^2} \tag{7.15}$$

will be dominated by Newton's gravitational potential $\Phi = GM/r$. This gives the elementary formula for the Schwarzschild radius

$$\boxed{r_s = \frac{2GM}{c^2}} \tag{7.16}$$

Here are a few examples:

mass	Schwarzschild radius
Electron mass $9.1 \cdot 10^{-31}$ kg	$1.3 \cdot 10^{-60}$m (below the Planck length)
Planck mass $2 \cdot 10^{-8}$ kg	Planck length $1.6 \cdot 10^{-35}$m
Human mass scale 1 kg	$1.5 \cdot 10^{-27}$m, less than the resolution at CERN
Earth mass $5.9 \cdot 10^{24}$ kg	7 mm
Solar mass $2.0 \cdot 10^{30}$ kg	3 km
Total mass of the Universe $1.6 \cdot 10^{55}$ kg	10^{28} m $\cong$ Visual horizon of the universe

Of course, these examples have no specific meaning, since the Schwarzschild radii are in all cases smaller than the objects under consideration, and we ignored the fact that the metric is only valid outside the objects. They are only intended to give a vague idea of the order of magnitude.

Gravitational redshift

The exterior Schwarzschild metric

$$\mathrm{d}s^2 = -\left(1 - \frac{r_s}{r}\right)\mathrm{d}t^2 + \left(1 - \frac{r_s}{r}\right)^{-1}\mathrm{d}r^2 + r^2(\,\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)$$

exhibits two singularities, namely

$r = r_s$: an apparent singularity at the Schwarzschild radius
$r = 0$: a real physical singularity at the center

The apparent singularity is a coordinate-related singularity, similar to how spherical coordinates at the north pole of a sphere are singular, although in reality there is no singularity at the north pole. The apparent singularity in the Schwarzschild metric is a consequence of the time coordinate, which is chosen here in such a way that it corresponds to a clock located at infinite distance. As can be seen directly from the g_{00}-component of the Schwarzschild metric, seen from this perspective, clocks near the center slow down by a factor of $\sqrt{1 - r_s/r}$ and eventually stop at the Schwarzschild radius.

Remark: In their own reference frame, the clocks do of course not stop at the Schwarzschild radius, they only seem to stop when they are observed from infinite distance. Clocks within

the Schwarzschild radius cannot be seen from an observer at infinite distance since they are causally disconnected.

This gravitational time dilation leads to the effect of the so-called *gravitational redshift*. If an object near a heavy star emits light with the frequency ν_e, an observer at infinite distance will perceive a red-shifted frequency $\nu_r = \nu_e\sqrt{1 - r_s/r}$. In astrophysics, the *redshift* is defined as the dimensionless quantity

$$z = \frac{\lambda_r - \lambda_e}{\lambda_e}. \tag{7.17}$$

With this definition the gravitational red shift is given by

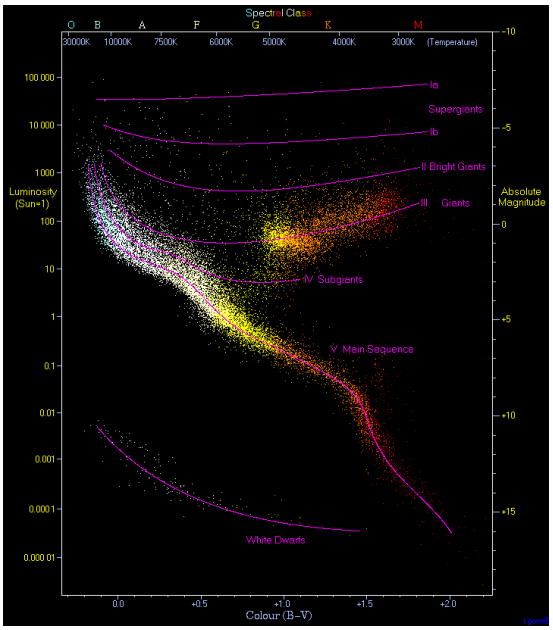
$$z = \frac{1}{\sqrt{1 - \frac{r_s}{r}}} - 1 \approx \frac{r_s}{2r}. \tag{7.18}$$

For example, the radiation emitted by the sun is red-shifted by $z \approx 2 \cdot 10^{-6}$, corresponding to a time difference between clocks on the suns surface and at infinity of approximately 19 hours per 1000 years. In other words, the gravitational redshift of sunlight is quite small.

On the other hand, even the much smaller gravitational redshift of our Earth matters e.g. when it comes to GPS satellites[2]. GPS satellites carry atomic clocks and your cellphone detects its position by comparing these clocks. GPS would not work if such corrections were not properly taken into account.

7.2 Celestial bodies with radial symmetry

Gravity is an attractive interaction, causing matter to clump and finally even collapse. If we were up to gravity alone, gravity would contract all existing matter locally to one point. However, as the density of the material increases, other physical counterforces begin to take effect, which under certain circumstances can stop the complete collapse and stabilize a massive object. As a result we observe radially symmetrical celestial bodies of various forms with (in comparison to the universe) high matter density. Depending on the type of the stabilizing counterforce, we can classify different kinds of celestial bodies.



The main sequence (Hertzsprung–Russell diagram) [Wikimedia]

[2]GPS = Global Positioning System, see Wikipedia

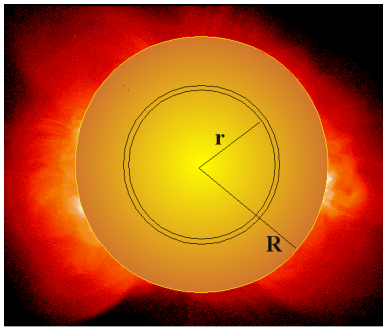


Figure 7.1: Equilibrium of stars: The gravitational force acting on a thin mass shell extending form r to $r + \mathrm{d}r$ is balanced by a pressure gradient.

The simplest classification criterion is the *luminosity*.[3] Plotting the luminosity double-logarithmically as function of the mean wavelength of the emission spectrum, giving the so-called *Hertzsprung-Russell diagram* (see adjacent figure). As can be seen, the diagram shows different groups of celestial bodies. Ordinary stars lie on the diagonal, the so-called *main sequence*. Next to it are the groups of white dwarfs and red giants. All these objects are in a (quasi) static equilibrium, which is characterized on the one hand by a balance of forces and on the other hand by a specific thermodynamic equation of state. In the following we will briefly discuss the most important cases.

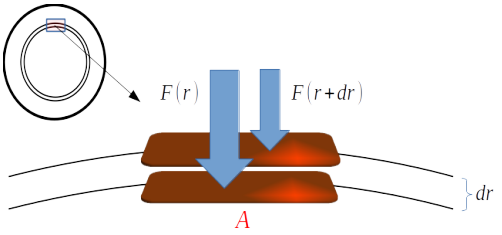
7.2.1 Stars

Simple classical approximation

With a red shift of $\approx 10^{-6}$, ordinary stars like the sun can be considered in good approximation as non-relativistic objects. Therefore, let us first examine the conditions for a star equilibrium based on Newtonian physics, which is described by the interplay of pressure $P(r)$ and density $\rho(r)$. To this end we consider a spherical surface with radius r. Clearly, the total mass of the celestial body below this spherical shell is

$$M(r) \;=\; 4\pi \int_0^r \mathrm{d}r\, r^2\, \rho(r)\,. \tag{7.19}$$

Let us now consider a thin shell between two spherical surfaces with thickness $\mathrm{d}r$ (see Fig. 7.1). Suppose that we investigate two surface elements with area A on these spheres. Then the weight resting on the lower surface is slightly larger than the weight resting on the upper surface. More



Force difference $\mathrm{d}F = F(r) - F(r + \mathrm{d}r)$ between two test areas.

[3] The luminosity $L = 4\pi r^2 F$ is the observed luminous intensity of a celestial body normalized by the distance. For a black body with radius R surface temperature T is given by $L = 4\pi R^2 \sigma T^4$, where $\sigma = \pi^2 k_B^4 / 60\hbar^3 c^2$ is the Stefan-Boltzmann constant.

specifically, the force F caused by the material on the surfaces differs by the amount

$$\mathrm{d}F \;=\; G\frac{M(r)\,\mathrm{d}M}{r^2} \;=\; G\frac{M(r)\rho(r)A\,\mathrm{d}r}{r^2}\,, \tag{7.20}$$

where $\mathrm{d}M = \rho(r)A\,\mathrm{d}r$ is the mass between the two test surfaces. This leads to a gradient $\mathrm{d}P = \mathrm{d}F/A$ of the pressure $P = F/A$:

$$\frac{\mathrm{d}P(r)}{\mathrm{d}r} \;=\; -\frac{1}{A}\frac{\mathrm{d}F}{\mathrm{d}r} \;=\; -\frac{G\,\rho(r)\,M(r)}{r^2}\,. \tag{7.21}$$

When solving this differential equation, the integration constant must be selected so that the pressure on the surface of the celestial body disappears, i.e. $P(R) = 0$. To solve this, thermodynamic equations are also required, which combine the pressure $P(r)$, the density $\rho(r)$ and the temperature $T(r)$.

Approximation of constant mass density

As a rough approximation, let us assume that the mass density $\rho(r)$ is constant inside the star. Then we have $M(r) = \frac{4}{3}\pi r^3\rho$, simplifying the differential equation (7.21) to

$$\frac{\mathrm{d}P(r)}{\mathrm{d}r} = -\frac{4}{3}\pi r G\rho^2\,. \tag{7.22}$$

The solution of this differential equation with the boundary condition $P(R) = 0$ is

$$P(r) \;=\; \frac{2\pi}{3}G\rho^2\left(R^2 - r^2\right). \tag{7.23}$$

With a constant density, the Schwarzschild radius (7.16) is given by [4]

$$r_s = \frac{8\pi G\rho R^3}{3c^2}\,. \tag{7.24}$$

If we combine both equations, we finally get

$$\frac{r_s}{R} = \frac{4P(0)}{\rho c^2}\,. \tag{7.25}$$

This is a really nice result: The current pressure divided by the 'relativistic pressure' ρc^2 is proportional to the dimensionless ratio r_s/R.

Star equilibrium

Stars emerge in regions of space with increased dust density, often triggered by shock waves from a supernova explosion. The dust cloud contracts under the effect of gravity and forms a so-called *protostar*. Protostars are about 1000 times larger than the solar

[4] Note that r_s lies *within* the celestial body, where the outer Schwarzschild metric is not valid.

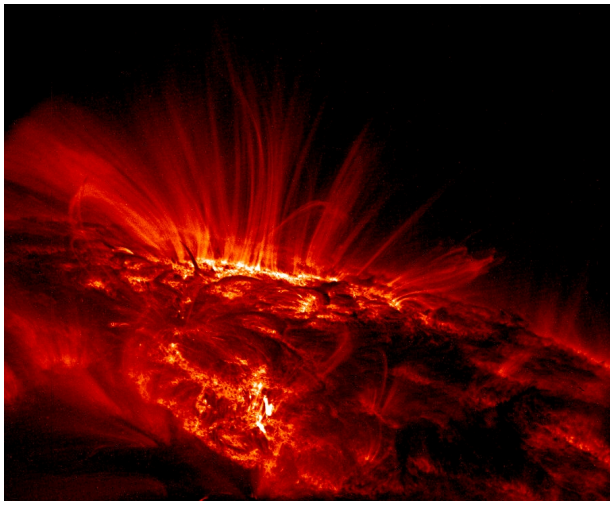


Figure 7.2: Surface of the sun seen from the satellite TRACE. (NASA - Wikimedia Commons)

system, still having a low density and a temperature of only a few Kelvin. The onset of the gravitational collapse leads to a steadily increasing temperature, which eventually leads to the ionization of the hydrogen and finally to the ignition of a nuclear fusion H $\rightarrow$ He, known as *hydrogen burning*. If the counterpressure caused by this nuclear reaction is able to stop the gravitational collapse, a star is formed. And since it constantly produces energy, it is shining.

The hydrogen burning starts in the center of the star and moves slowly outwards like a spherical shell. The waste product helium is further compressed in the core under the influence of gravity with extremely high pressure until a new level of nuclear fusion, the so-called *helium burning*, ignites. The hydrogen fusion zone again moves from the inside to the outside, thereby forming beryllium and carbon. If the stars are sufficiently heavy, the center will re-ignite another time, causing *carbon burning*. Depending on the mass of the star, this process can be continued with the formation of ever heavier elements down to iron.

As a result of these successive *firing cycles*, the star continues to inflate as its surface temperature decreases. Stars in such a late stage are called *red giants*. Eventually, the fusion reactions break down, causing the red giant to collapse under its own weight. Since it now consists of heavy nuclei such as iron, the collapse cannot be stopped by fusion reactions. It depends on the total mass whether this collapse creates a white dwarf, a neutron star or even a black hole (see below).

The sun is an average star in its best years. It consists of three quarters of hydrogen and one quarter of helium. The hydrogen burning takes place in the fusion zone in the center, which extends up to approximately $r = R/4$. The heat generated is transported to the outside by convection through the surrounding layers. Nuclear fusion stabilizes itself through negative feedback: If too much energy is produced, the star initially expands. This reduces the gravitational influence and the pressure in the center decreases. As a result, the conditions for nuclear fusion become less favorable and energy production is automatically reduced. This explains why stars shine so steadily.

Estimation of r_s/R

In order to relate pressure and density, we roughly assume that the sun is an ideal gas, i.e. $PV = Nk_BT$. In a rough approximation, let us identify P with the pressure in the center $P(0)$ and estimate the number of particles by $N = M/m_p$, where m_p is the proton mass. The ideal gas equation of state thus takes the form

$$P(0) = \frac{\rho\, k_B T}{m_p}. \tag{7.26}$$

Inserted into (7.25) we obtain the estimate

$$\boxed{\frac{r_s}{R} \approx \frac{4k_BT}{m_pc^2}}. \tag{7.27}$$

The Schwarzschild radius of the sun is 2.952 km, corresponding to $r_s/R = 4.24 \cdot 10^{-6}$. Since the proton rest mass is approximately 1 GeV, the resulting temperature of the plasma is $k_BT \approx 1$ keV $\approx 10^8$K. In the literature one finds values of about 15 million Kelvin, so our estimate differs by less than an order of magnitude.

We can also calculate the pressure in the center of the sun:

$$P(0) = \rho c^2 \frac{1\text{ keV}}{1\text{ GeV}} \approx 10^{-6}\rho c^2 = 10^{-6} \cdot 1408 \frac{\text{kg}}{\text{m}^3} c^2 \approx 1.2 \cdot 10^{14}\text{ Pa} \approx 1.2 \cdot 10^9\text{bar}. \tag{7.28}$$

The literature value is roughly 200 billion bar, which is significantly higher. This illustrates the limitations in assuming a constant density. In fact, the actual density in a star varies greatly and reaches values of around 150.000 kg/m^3 at the core, compared with an average density of the entire sun of only 1408 kg/m^3.

> **Remark:** Nuclear fusion occurs at energies of some 10 MeV and not yet at 1 keV. So in the sun it is far too cold for allowing nuclear fusion of hydrogen to helium. In fact, nuclear fusion in the sun is not comparable to what is happening in a thermonuclear explosion, otherwise the sun would literally explode like a supernova. Rather, nuclear fusion occurs only occasionally due to the *tunnel effect* at an extremely low rate (approx. 10^-20) . The sheer size of the sun ensures that enough energy is still produced to stabilize the celestial body.
>
> Incidentally, the same is true for our Earth. In fact, it would have cooled long ago if nuclear fission processes did not take place inside at a low rate due to the tunnel effect.

7.2.2 White dwarfs

White dwarfs are comparatively small star-like objects, which can be found below the main sequence in the Hertzsprung-Russel diagram. They represent the final stage of low-mass stars and are the remnants of red giants that have shed their outer shell and collapsed. As we will see, a white dwarf can only exist if it is lighter than 1.44 solar masses. In its iterior it usually consists of the burned-out core of a star.

White dwarfs are about the size of the earth, but contain approximately the mass of the sun. Their surface temperature is initially between 10 000 and 100 000 K. The resulting white or bluish color explains their attribute of being “white”. No nuclear

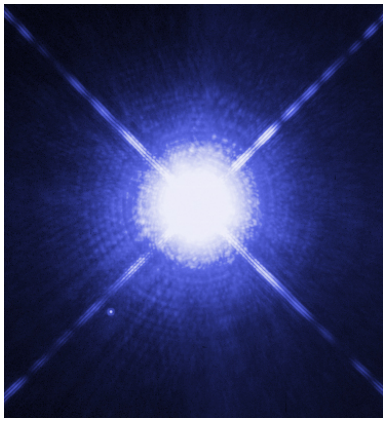


Figure 7.3: Sirius A and B, image taken by the Hubble telescope. The white dwarf can be seen in this overexposed picture as a small dot at the bottom left. (NASA/ESA - Wikimedia Commons)

fusion or any other kind of energy production takes place inside a white dwarf, so it continuously cools down and eventually becomes a 'brown' or even a 'black dwarf'. Sadly, our sun also awaits this fate.

The closest white dwarf is *Sirius B*, which rotates in a gravitational double star system with an ordinary star named Sirius A. For a long time, Sirius B was only detectable by orbital anomalies of Sirius A, the reason being that the absolute luminosity of an ordinary star is much higher than that of a white dwarf due to the larger radiation area. A direct image confirming its existence was only possible with the Hubble space microscope.

Fermi pressure

White dwarfs are stabilized by the *Fermi pressure* of the electrons. To understand this process, consider a compressed Fermi gas from N fermions of rest mass m, which are contained in a volume V. Since each particle occupies an average volume V/N, its typical location uncertainty is given by $\Delta x = (V/N)^{1/3}$. According to the Heisenberg uncertainty principle, this results in a momentum uncertainty of

$$\Delta p = \frac{\hbar}{\Delta x} = \hbar\left(\frac{V}{N}\right)^{-1/3}. \tag{7.29}$$

This results in an average kinetic energy

$$E_{kin} = N\left(\sqrt{m^2c^4 + (\Delta p)^2c^2} - mc^2\right). \tag{7.30}$$

In a non-relativistic approximation we therefore have

$$E_{kin} \approx N\frac{(\Delta p)^2}{2m} \approx \frac{N^{5/3}\hbar^2}{2mc^2V^{2/3}}, \tag{7.31}$$

As one can see, the energy increases with *decreasing* the rest mass of the particle. This is because the uncertainty principle fixes the momentum, - the smaller the mass, the

higher the corresponding kinetic energy $p^2/2m$. In a white dwarf consisting of ionized protons and electrons, the Fermi pressure of the light electrons therefore dominates.

To determine the equilibrium of the white dwarf, we minimize the sum of the kinetic energy E_{kin} and the gravitational energy $E_{grav} \approx -GM^2/R$, where $V = \frac{4}{3}\pi R^3$ is the volume, $N = M/m_n$ denotes the number of particles, and m_n, m_e are the nucleon are electron masses:

$$E = E_{kin} + E_{grav} = \frac{M}{m_n}\left(\sqrt{m_e^2c^4 + \frac{3^{2/3}c^2\hbar^2M^{2/3}}{2\sqrt[3]{2}\pi^{2/3}m_n^{2/3}R^2}} - m_ec^2\right) - \frac{GM^2}{R} \tag{7.32}$$

This expression is now minimized for R by solving the equation $\mathrm{d}E/\mathrm{d}R = 0$. The solution is

$$R_0 = \sqrt{\frac{36^{3/2}c^2\hbar^4 - 2(6\pi)^{2/3}G^2\hbar^2M^{4/3}m_n^{8/3}}{8\pi^{4/3}c^2G^2M^{2/3}m_e^2m_n^{10/3}}}\,. \tag{7.33}$$

In the limit of small masses $M \to 0$, the first term dominates in the numerator, meaning that the radius of the white dwarf scales like $R_0 \sim M^{-1/3}$. This implies that

A white dwarf shrinks with increasing mass!

Typical radii of white dwarfs are between 7000 km and 14000km, which is comparable to the size of the Earth.

Chandrasekhar limit

With increasing mass, the white dwarf becomes smaller and smaller until the Fermi pressure is no longer sufficient to withstand the gravitational collapse. The critical limit mass at which this happens can be calculated using the above formula by setting $R_0 = 0$. The result reads

$$M_c = \sqrt{\frac{3}{4\pi}}\left(\frac{c\hbar}{G}\right)^{3/2}\frac{1}{m_n^2} \tag{7.34}$$

and does not depend on the electron mass. This formula differs from the result of a fully relativistic treatment only in the pre-factor. The actual correct result is

$$M_c = \frac{2.01824\sqrt{3\pi}}{2}\left(\frac{c\hbar}{G}\right)^{3/2}\frac{1}{\eta^2m_n^2}\,, \tag{7.35}$$

where η is the molecular weight per electron, therewith taking into account the specific composition of the white dwarf.

The mass M_c is the so-called *Chandrasekhar mass*. Apart from the pre-factors, it only depends on fundamental constants $(\hbar, c, G)$ and the nucleon mass m_n. Since the fundamental *Planck mass* is given by $m_p = \sqrt{\hbar c/G}$, the Chandrasekhar mass can be written (up to a pre-factor of the order 1) by

$$M_c \propto \frac{m_p^3}{m_n^2} \approx \frac{(2.176\cdot10^{-8}\mathrm{kg})^3}{(1.673\cdot10^{-27}\mathrm{kg})^2} \approx 3.68\cdot10^{30}\mathrm{kg}\,. \tag{7.36}$$

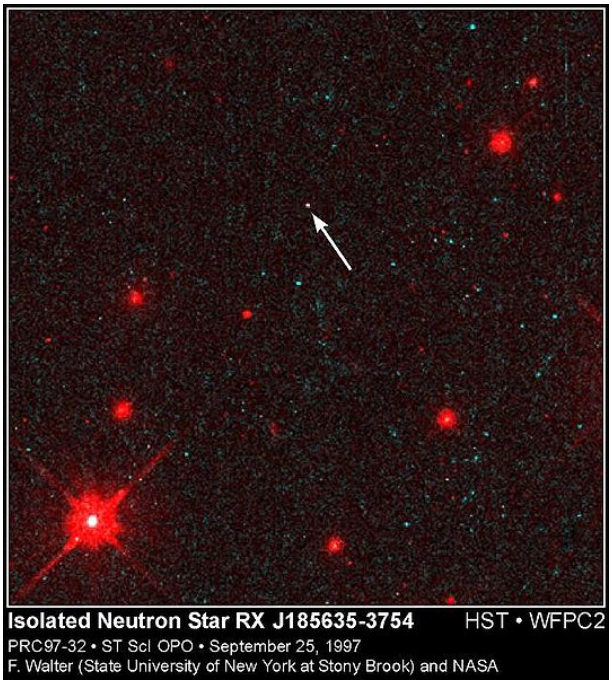


Figure 7.4: Neutron star (NASA - Wikimedia Commons)

For comparison: the solar mass is approximately $2 \cdot 10^{30}$ kg. In fact, the masses of the observable stars vary only a little, it is possible to find stars with masses in the range of about 0.07 to 120 solar masses, but most of them are found within two orders of magnitude. The Chandrasekhar mass is pretty much in the middle of this range. However, it is defined primarily in terms of *microscopic* physics, namely the uncertainty principle and the nucleon mass in combination with the gravitational constant.

The typical mass of a star is approximately m_p^3/m_n^2.

7.2.3 Neutron stars

If a very heavy star with more than about 10 solar masses collapses, it first goes through the temporary stage of a white dwarf. However, if there are particle momenta higher than $1.5 m_e c^2$, the so-called *inverse β-decay*

$$p + e^- \rightarrow n + \nu_e \tag{7.37}$$

will set in. With a huge burst of neutrinos the electrons and the protons merge to neutrons. Since the electrons are no longer available to stabilize the white dwarf, another gravitational collapse sets in.

This gravitational collapse continues until the Fermi pressure of the newly formed neutrons becomes strong enough to stabilize the body. The formulas derived above should therefore remain valid, all what we have to do is to replace the electron mass by the neutron mass. Because this mass enter only as a pre-factor in Eq. (7.33) and since

a neutron is about 2000 times heavier than an electron, it is clear that a neutron star will also be 2000 times smaller than a white dwarf. In fact, a neutron star has a radius of less than 10 km, but it contains more than a solar mass and thus reaches densities of about 100 billion tons per cubic centimeter. Neutron stars are characterized by a Schwarzschild ratio $r_s/R \approx 0.3$ and are therefore highly relativistic objects. Therefore, the approximations from the previous section are at best qualitatively correct, and a more careful treatment will lead to a modified equation of state.

Here, too, there is a critical limit mass, the so-called *Oppenheimer-Volkoff limit mass*, which differs from the Chandrasekhar limit mass only by a factor.

Neutron stars do not generate any energy, so they cool down slowly and are then stable (provided they do not collect any more matter from surrounding objects). To date, more than 2000 neutron stars have been identified in the Milky Way. 5 % of them are part of a binary system, i.e. they form a gravitationally bound pairs with another neutron star or white dwarf.

7.3 Dynamic solutions of the field equations

We now want to investigate what happens, when and how a gravitational collapse occurs. First the *inner Schwarzschild metric* is used to determine the stability limit beyond which a collapse is inevitable. In order to describe the collapse as a time-dependent process, we need a dynamic radially symmetrical solution of the field equations. This solution can be used not only to describe collapsing stars, but also collapsing galaxies (with stars as particles) and even for the entire universe (with galaxies as particles).

7.3.1 Inner Schwarzschild metric

We now calculate the representation of the metric tensor *in the interior* of a radially symmetrical mass distribution. Again we use the approach (7.4)

$$\mathrm{d}s^2 = -A(r)\,\mathrm{d}t^2 + B(r)\,\mathrm{d}r^2 + r^2(\,\mathrm{d}\theta^2 + \sin^2\theta\;\mathrm{d}\phi^2)\,, \tag{7.38}$$

with real positive functions $A(r)$ and $B(r)$, which are determined by the field equations, but now with a non-vanishing energy momentum tensor. The celestial body is modeled by a perfect fluid, i.e. we start from Eq. (5.23)

$$T^{\mu\nu} = (\rho + p)u^\mu u^\nu + p g^{\mu\nu}\,. \tag{7.39}$$

We also want to assume that we are dealing with a *static* celestial body, i.e. the spatial components of the 4-velocity u^1, u^2, u^3 are assumed to vanish. Because of $u^\mu u_\mu = u^0 u_0 = c^2$ we can conclude that

$$u^0 = c/\sqrt{A(r)}\,, \qquad u_0 = -c\sqrt{A(r)}\,. \tag{7.40}$$

The field equations can now be solved using elementary methods. The most important result is the *Oppenheimer-Volkoff equation* - an integro differential equation for the

pressure as a function of the radius:

$$\boxed{\frac{\mathrm{d}p(r)}{\mathrm{d}r} = -\frac{GM(r)\rho(r)}{r^2}\left(1+\frac{p(r)}{\rho(r)c^2}\right)\left(1+\frac{4\pi r^3 p(r)}{M(r)c^2}\right)\left(1-\frac{2GM(r)}{c^2 r}\right)^{-1}.} \tag{7.41}$$

Here

$$M(r) = 4\pi\int_0^r r'^2\rho(r')\,\mathrm{d}r' \tag{7.42}$$

denotes as before the mass within the radius r. This equation can only be solved in combination with a suitable *equation of state* that relates pressure and density. If successful, the functions $A(r)$ and $B(r)$ can be calculated (see proof below):

$$A(r) = \exp\left[-\frac{2G}{c^2}\int_r^\infty \frac{\mathrm{d}r'}{r'^2}\left(M(r')+\frac{4\pi r'^3 p(r')}{c^2}\right)\left(1-\frac{2GM(r')}{c^2 r'}\right)^{-1}\right] \tag{7.43}$$

and

$$B(r) \;=\; \left(1-\frac{2GM(r)}{c^2 r}\right), \tag{7.44}$$

If you put a constant value for $M(r)$ in the last equation (as if all the mass was concentrated in the center), you would get exactly the same expression for $B(r)$ as in the case of the outer Schwarzschild metric, which is exactly what we expect.

Proof: According to the ansatz, the metric tensor is given by

$$g_{\mu\nu} = \mathrm{diag}\big(-A(r),\, B(r),\, r^2,\, r^2\sin^2\theta\big)\,.$$

This results in the following representation of the energy momentum tensor

$$T_{\mu\nu} = \mathrm{diag}\big(\rho c^2 A(r),\, pB(r),\, pr^2,\, p(r^2\sin^2\theta)\big)\,.$$

Since all tensors are still diagonal, there are in principle four field equations. Since the equation for R_{33} equation depends linearly on the one for R_{22}, only the equations with the indices 00, 11 and 22 remain. By clever addition one can show that

$$\frac{R_{00}}{2A}+\frac{R_{11}}{2B}+\frac{R_{22}}{r^2} = -\frac{B'}{rB^2}-\frac{1}{r^2}+\frac{1}{r^2 B} = -\frac{8\pi G}{c^2}\rho$$

which results in a differential equation for $B(r)$:

$$\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{r}{B(r)}\right) \;=\; 1-\frac{8\pi G}{c^2}\rho r^2\,.$$

Together with the condition that $B(0)$ is finite one is led to the solution (7.44). The condition that the divergence of the energy momentum tensor vanishes (continuity equation) gives another differential equation, namely

$$-\frac{A'(r)}{A(r)} \;=\; -\frac{2p'(r)}{\rho(r)c^2+p(r)}\,.$$

Combining this with the third field equation for R_{22} we arrive at the Oppenheimer-Volkoff equation (7.41) and by integration finally at Eq. (7.43).

7.3.2 Absolute stability limit

In the following we will show that there is a critical threshold for M beyond which no physical mechanism can prevent a gravitational collapse. In this case, the whole object

must eventually collapse to a singularity, namely, a black hole.

The starting point is the *Oppenheimer-Volkoff equation* (7.41), which describes the internal pressure $p(r)$ for a radially symmetrical relativistic object as a function of the radius r. To solve this equation, we have to know the *equation of state* of the object which relates pressure and density. As the simplest approximation we want to assume that the matter is incompressible and that the celestial body has a constant density $\rho = \rho_0$. With the abbreviations

$$X = \sqrt{1 - \frac{r_s}{R}}, \qquad Y = \sqrt{1 - \frac{r_s r^2}{R^3}} \tag{7.45}$$

the two functions of the inner Schwarzschild metric are then given by

$$A(r) = \frac{1}{4}(3X - Y)^2, \qquad B(r) = Y^{-2}, \tag{7.46}$$

which leads us to the solution

$$P(r) = \rho_0 c^2 \frac{Y - X}{3X - Y}. \tag{7.47}$$

As expected, the pressure

$$P(0) = \rho_0 c^2 \frac{1 - X}{3X - 1} \tag{7.48}$$

is maximal in the center of the celestial body. Amazingly, the expression becomes divergent for $X = 1/3$, i.e. for $r_s/R = 1 - 1/9 = 8/9$. However, since there is no physical mechanism that can withstand infinite pressure, we can conclude that for

$$R < \frac{9}{8} r_s \tag{7.49}$$

every star *must* collapse! The pre-factor $\frac{9}{8}$ comes from the assumption of incompressibility, but the above inequality remains qualitatively correct even for realistic stars. In astrophysics, it also leads to the conclusion that the observable gravitational redshift z is limited to values in the range

$$z = \frac{\lambda_r}{\lambda_e} - 1 \;=\; \frac{1}{\sqrt{1 - \frac{r_s}{R}}} - 1 \;<\; 2 \tag{7.50}$$

since objects with a larger red shift would not be stable.

7.3.3 Passing the Schwarzschild radius in free fall

Before we deal with the gravitational collapse itself, let us again consider the exterior Schwarzschild metric:

$$\mathrm{d}s^2 \;=\; -\Big(1 - \frac{r_s}{r}\Big)\,\mathrm{d}t^2 + \Big(1 - \frac{r_s}{r}\Big)^{-1}\mathrm{d}r^2 + r^2(\,\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\,. \tag{7.51}$$

One can show (see exercise) that the equations of motion for a freely falling particle, namely, the geodesic equation

$$\ddot{x}^\alpha + \Gamma^\alpha{}_{\mu\nu}\dot{x}^\mu\dot{x}^\nu = 0 \tag{7.52}$$

is in this case given by

$$\frac{\mathrm{d}t}{\mathrm{d}\tau}\Big(1-\frac{r_s}{r}\Big) = 1 \tag{7.53}$$

$$r^2\Big(\frac{\mathrm{d}\phi}{\mathrm{d}\tau}\Big)^2 = L \tag{7.54}$$

$$\Big(\frac{\mathrm{d}r}{\mathrm{d}\tau}\Big)^2 - \frac{r_s c^2}{r} + \frac{L^2}{r^2} - \frac{r_s L^2}{r^3} = Q \tag{7.55}$$

where Q, L are constants of motion and it is assumed that the whole trajectory takes place in the equatorial plane $\theta = \pi/2$ of the Schwarzschild metric. We can identify the constant L as the angular momentum that vanishes in the case examined here. Furthermore, we can insert $r_s = \frac{2GM}{c^2}$, simplifying the equations of motion:

$$\frac{\mathrm{d}t}{\mathrm{d}\tau} = \frac{1}{1-\frac{r_s}{r}} = \frac{1}{1-\frac{2GM}{rc^2}} \tag{7.56}$$

$$\Big(\frac{\mathrm{d}r}{\mathrm{d}\tau}\Big)^2 = \frac{2GM}{r} + Q\,. \tag{7.57}$$

How long does it take for a particle to move from radius r_0 to r_s and what is the distance between the two points?

- **The flight duration seen from an observer at infinity diverges:**

$$\Delta t = \int_{t_0}^{t_s}\mathrm{d}t = \int_{r_s}^{r_0}\frac{\mathrm{d}t}{\mathrm{d}\tau}\frac{\mathrm{d}\tau}{\mathrm{d}r}\,\mathrm{d}r = \int_{r_s}^{r_0}\frac{1}{\sqrt{\frac{2GM}{r}+Q}}\frac{1}{\Big(1-\frac{r_s}{r}\Big)}\,\mathrm{d}r = \infty \tag{7.58}$$

  since the integral has a pole at the upper limit. In fact, since the red shift diverges as well, the particle becomes unobservable after short time.

- **Flight time seen from the perspective of the particle:**
  This requires to integrate the eigenzeit τ of the particle:

$$\Delta\tau_1 = \int_{\tau_0}^{\tau_s}\mathrm{d}t = \int_{r_s}^{r_0}\frac{\mathrm{d}\tau}{\mathrm{d}r}\,\mathrm{d}r = \int_{r_s}^{r_0}\frac{1}{\sqrt{\frac{2GM}{r}+Q}}\,\mathrm{d}r \;<\; \infty \tag{7.59}$$

- **Distance traveled up to the Schwarzschild radius:**
  Her one also gets a finite integral

$$\Delta s = \int_{r_s}^{r_0}\frac{\mathrm{d}r}{1-\frac{r_s}{r}} = \frac{1}{2}\left(r_s\log\left(\frac{r_s}{r_0}\right) - 2\left(\sqrt{r_0(r_0-r_s)} + r_s\log\left(\sqrt{1-\frac{r_s}{r_0}}+1\right)\right)\right) \tag{7.60}$$

- **Time needed to fly from the horizon to the center seen from the perspective of**

the particle:

$$\Delta\tau_2 = \int_0^{r_s} \frac{1}{\sqrt{\frac{2GM}{r} + Q}} < \infty \tag{7.61}$$

These examples show that the Schwarzschild radius is physically significant in that an observer at infinite distance cannot receive any information from areas within the Schwarzschild radius, i.e., the Schwarzschild sphere separates the causally accessible exterior of the black hole from the causally accessible interior seen from infinite distance. However, if a particle crosses the Schwarzschild radius, it will not notice any singular spatio-temporal structure. The singularity at the Schwarzschild metric is therefore a *coordinate singularity* that arises from the choice of the reference system at infinity.

7.3.4 Gravitational collapse

Gaussian normal coordinates

To investigate the collapse of a star, i.e. the free fall of a radially symmetrical mass distribution, we need a different coordinate system that does not diverge at the Schwarzschild radius. As we will see below, it is convenient to choose a coordinate system that moves with the collapsing matter. Such coordinates are called *Gaussian normal coordinates* and are defined by

$$x^0 = c\tau\,, \qquad x^1, x^2, x^3 = const\,. \tag{7.62}$$

In such a coordinate system, the collapsing particles would have a trivial 4-speed $u^\mu = (c, 0, 0, 0)$, i.e. even though the star collapses, the particles seem to be at rest in the co-moving coordinate system.

It can be shown that in the isotropic case such a metric can be written in the form

$$\mathrm{d}s^2 = -c^2\,\mathrm{d}t^2 + U(r,t)\,\mathrm{d}r^2 + V(r,t)(\,\mathrm{d}\theta^2 + \sin^2\theta\;\mathrm{d}\phi^2)\,. \tag{7.63}$$

With this ansatz one calculates the Christoffel symbols

$$\begin{aligned} \Gamma^1{}_{01} = \Gamma^1{}_{10} &= \frac{\dot{U}}{2U} \\ \Gamma^0{}_{11} &= \frac{\dot{U}}{2} \\ \Gamma^2{}_{02} = \Gamma^2{}_{20} = \Gamma^3{}_{03} = \Gamma^3{}_{30} &= \frac{\dot{V}}{2V} \\ \Gamma^0{}_{22} &= \frac{\dot{V}}{2} \\ \Gamma^0{}_{33} &= \frac{\dot{V}}{2}\sin^2\theta \end{aligned} \tag{7.64}$$

and

$$\begin{aligned} \Gamma^1{}_{11} &= \frac{U'}{2U} \\ \Gamma^1{}_{22} &= -\frac{V'}{2U} \\ \Gamma^1{}_{33} &= -\frac{V'}{2U}\sin^2\theta \\ \Gamma^2{}_{12} = \Gamma^2{}_{21} = \Gamma^3{}_{13} = \Gamma^3{}_{31} &= \frac{V'}{2V} \\ \Gamma^2{}_{33} &= -\sin\theta\cos\theta \\ \Gamma^3{}_{23} = \Gamma^3{}_{32} &= \cot\theta \end{aligned} \tag{7.65}$$

where the dot and the dash stand for the respective partial derivatives and all the Christoffel symbols not listed above are zero. We first use the field equation to check whether a constant four-speed $u^\mu = (c, 0, 0, 0)$ is consistent with this ansatz. It is found that $\Gamma^\mu_{00} = 0$ for all μ, meaning that the trajectory equation $\frac{\mathrm{d}u^\mu}{\mathrm{d}\tau} = -\Gamma^\mu{}_{\nu\rho}u^\nu u^\rho$ is indeed fulfilled.

The non-vanishing components of the Ricci tensor can be calculated directly from the Christoffel symbols given above:

$$R_{00} = \frac{\ddot{U}}{2U} + \frac{\ddot{V}}{V} - \frac{\dot{U}^2}{4U^2} - \frac{\dot{V}^2}{2V^2} \tag{7.66}$$

$$R_{11} = -\frac{\ddot{U}}{2} + \frac{\dot{U}^2}{4U} - \frac{\dot{U}\dot{V}}{2V} + \frac{V''}{V} - \frac{V'^2}{2V^2} - \frac{U'V'}{2UV} \tag{7.67}$$

$$R_{22} = -1 - \frac{\ddot{V}}{2} - \frac{\dot{U}\dot{V}}{4U} + \frac{V''}{2U} - \frac{V'U'}{4U^2} \tag{7.68}$$

$$R_{01} = R_{10} = \frac{\dot{V}'}{V} - \frac{\dot{V}V'}{2V^2} - \frac{\dot{U}V'}{2UV} \tag{7.69}$$

$$R_{33} = R_{22}\sin^2\theta \tag{7.70}$$

Radially symmetric collapse

Let us now assume that the collapsing matter consists of dust, i.e. it has no internal pressure. Furthermore, let us assume that its density is spatially constant, i.e. $\rho(r,t) = \rho(t)$. Possible applications are:

- Star collapse at $r > \frac{9}{8}r_s$
- Formation of a new star from a cloud of dust
- Collapse of a galaxy (with stars as dust particles)
- Collapse of the universe (with galaxies as dust particles)

With $p = 0$ and $u^\mu = (c, 0, 0, 0)$ the energy momentum tensor takes the form

$$T^{\mu\nu} = (\rho + p)u^\mu u^\nu + p g^{\mu\nu} = \begin{pmatrix} \rho(t)c^2 & & & \\ & 0 & & \\ & & 0 & \\ & & & 0 \end{pmatrix} . \tag{7.71}$$

This gives us the field equations

$$R_{\mu\nu} = T_{\mu\nu} + \frac{1}{2} g_{\mu\nu} T = \frac{\rho(t)c^2}{2} \begin{pmatrix} 1 & & & \\ & U(r,t) & & \\ & & V(r,t) & \\ & & & V(r,t)\sin^2\theta \end{pmatrix} . \tag{7.72}$$

The occurring equations contain sums of spatial and temporal derivatives. This suggests to try a separation ansatz:

$$U(r,t) = R(t)^2 f(r) , \qquad V(r,t) = S(t)^2 g(r) . \tag{7.73}$$

The field equation for $R_{01} = R_{10}$ implies

$$\frac{\dot{S}}{S} = \frac{\dot{R}}{R} , \tag{7.74}$$

so that S and R can differ at most by constants which can be absorbed in f, g, so that $S = R$. The remaining field equations for R_{11} and R_{22} (R_{33}) read

$$-\frac{f'}{rf^2} = \frac{\ddot{R}R}{+} 2\dot{R}^2 - \frac{4\pi G}{c^2}\rho(t)R^2 \tag{7.75}$$

$$-\frac{1}{r^2} + \frac{1}{r^2 f} - \frac{f'}{2rf^2} = \frac{\ddot{R}R}{+} 2\dot{R}^2 - \frac{4\pi G}{c^2}\rho(t)R^2 \tag{7.76}$$

Since the left sides of these equations depend only on r while the right sides depend only on t, we can conclude that both sides have to be constant, i.e.

$$-\frac{f'}{rf^2} = -2k \tag{7.77}$$

$$-\frac{1}{r^2} + \frac{1}{r^2 f} - \frac{f'}{2rf^2} = -2k , \tag{7.78}$$

so that

$$f(r) = \frac{1}{1 - kr^2} . \tag{7.79}$$

Therefore, the metric inside the collapsing dust cloud is

$$\mathrm{d}s^2 = -c^2\,\mathrm{d}t^2 + R(t)^2\Big(\frac{\mathrm{d}r^2}{1-kr^2} + r^2(\,\mathrm{d}\theta^2 + \sin^2\theta\;\mathrm{d}\phi^2)\Big) . \tag{7.80}$$

Using the fact that the energy momentum tensor is divergence-free, one can show that

$\frac{1}{\sqrt{g}}\partial_0(\sqrt{g}T^{00}) = 0$, implying the conservation of mass:

$$\rho(t)R(t)^3 \;=\; \rho(0)\,. \tag{7.81}$$

This result is inserted into the field equation for R_{00}, giving

$$\ddot{R}R = -\frac{4\pi G}{3c^2}\,\frac{\rho(0)}{R}\,. \tag{7.82}$$

This implies that

$$\frac{\mathrm{d}\dot{R}}{\mathrm{d}(ct)} \;=\; 2\ddot{R}\dot{R} \;=\; -\frac{8\pi G\rho(0)}{3c^2}\,\frac{\dot{R}}{R^2} \tag{7.83}$$

with the solution

$$\dot{R}^2 = const + \frac{8\pi G}{3c^2}\,\frac{\rho(0)}{R}\,. \tag{7.84}$$

By inserting this into the field equation for R_{11} one can show that $const = -k$. With the initial conditions $R(0) = 1$ and $\dot{R}(0) = 0$ one obtains $k = 8\pi G\rho(0)/3c^2$, leading to the simplified differential equation

$$\dot{R}^2 \;=\; k\frac{1-R}{R}\,. \tag{7.85}$$

This differential equation describes a parameterized cycloid

$$ct = \frac{1}{2\sqrt{k}}(\lambda + \sin\lambda)\,, \qquad R = \frac{1}{2}(1+\cos\lambda) \tag{7.86}$$

When the curve parameter takes the value $\lambda = \pi$, the dust cloud collapses into a single point. This takes the time span

$$T \;=\; t\Big|_{\lambda=\pi} \;=\; \frac{\pi}{2c\sqrt{k}} \;=\; \frac{\pi}{2}\sqrt{\frac{3}{8\pi G\rho(0)}}\,. \tag{7.87}$$

Amazingly, this time does not depend on the absolute size of the dust cloud, but only on its density. At second glance, however, this is plausible, since the gravitational force acting on a dust particle is only caused by the components of the cloud *within* the sphere with the radius that corresponds to its distance. Assuming an earthly density of 1g / cm^3, this time is very short: the gravitational collapse would only take about 35 minutes.

7.3.5 Supernovae

> *Zhihe era, first year, seventh lunar month, 22nd day. [...] Yang Weide declared: "I humbly observe that a guest star has appeared; above the star there is a feeble yellow glimmer. If one examines the divination regarding the Emperor, the interpretation [of the presence of this guest star] is the following: The fact that the star has not overrun Bi and that its brightness must represent a person of great value. [...] Previously, during the first year of the Zhihe era, during the fifth lunar month, it had appeared at dawn, in the direction of the east, under the watch of Tiānguān (Zeta Tauri). It had been seen in daylight, like Venus. It had rays stemming in all*

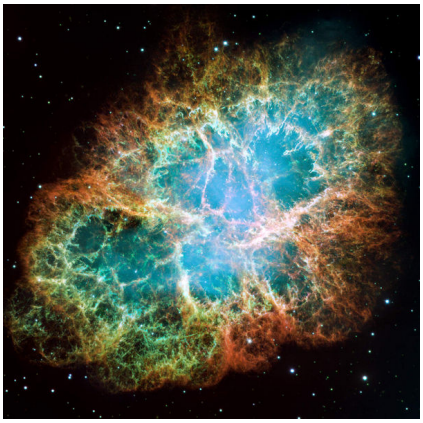
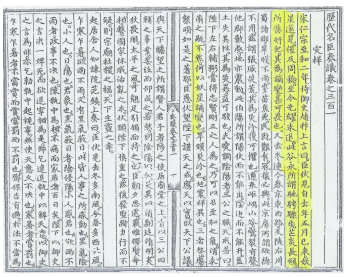


Figure 7.5: The Cancer Nebula: Remnants of a Supernova.

> *directions, and its colour was reddish white. Altogether visible for 23 days.*
>
> *Zhihe era of the reign, first year, fifth lunar month, jichou day. A guest star has appeared to the south-east of Tianguan, perhaps several inches away. After a year or more, it gradually disappeared.* [5]

As various documents report, Chinese astronomers observed a "guest star" in 1054, which initially shines brighter than Venus and is even visible during the day, but then slowly looses intensity. Today, the crab nebula, the remnant of a supernova, is located at the location precisely named in the Chinese source. In its middle we find a pulsar, a rapidly rotating neutron star, which X-ray astronomers like to use as a calibration source.



Chinese document reporting SN1054

But we don't have to look back so far. In February 1987 (I can remember very well) there was another spectacular Supernova, named SN1987A, which took place in the Magellanic Cloud, our neighboring galaxy. At that time there existed already very powerful neutrino detectors. At the time of the outbreak, 19 neutrinos were suddenly registered worldwide at the same time. This seemed to be a clear indication of a type II supernova in which a white dwarf collapses. If the stabilization by the electrons fails, the inverse β decay overtakes and the white dwarf collapses into a neutron star. However, the Hubble Space Telescope has taken images of the supernova regularly since August 1990 without a clear detection of a neutron star.

In our galaxy, a supernova is statistically expected to take place every 40 years. What exactly happens in a supernova is the subject of current research. What is certain is that the process of such a star explosion is initiated by a gravitational collapse. Only in the rarest of cases will this collapse be radially symmetrical, but usually in will carry a residual angular momentum that is more and more noticeable in the late stage of the collapse due to the pirouette effect.

A star collapse can be imagined roughly as follows: After the thermonuclear burning phases are getting shorter and shorter, an iron core forms in the middle of the star.

[5] Geschichte der Sung-Dynastie, China, zitiert nach J. J. L. Duyvendak

As soon as this core exceeds the Chandrasekhar limit mass (about 0.9 solar masses for iron), the core in itself begins to collapse. This happens very quickly - within milliseconds, while the outer layers fall into the center as a gravitational shock wave. As soon as the inner part of the core reaches densities at the nuclear level, it consists almost entirely of neutrons. If the critical Oppenheimer-Volkoff limit mass of a neutron star (approx. 3 solar masses) is not exceeded, the nucleus becomes incompressible due to the Fermi pressure of the neutrons, which suddenly stops the collapse and an enormous pressure increase takes place in the center. This creates a gigantic pressure wave which, after leaving the iron core, continues to gain energy as a result of complicated physical processes and fusion reactions that begin again as it spreads.

What remains - depending on the mass - is a neutron star or a black hole [6]. The often high speed of rotation creates a magnetic field that interacts with the particles of the repelled gas nebula and thus generates signals that can be received on Earth.

7.3.6 Black holes

A black hole is a mass accumulation that is so large that it is concentrated completely within its own Schwarzschild radius. The Schwarzschild horizon has the quality of a light-like dividing surface: If you were to emit a light beam horizontally with a flashlight, you would send the light into a circular orbit, so to speak. The Schwarzschild horizon is therefore a two-dimensional surface of geodesic lines that separates the inside and the outside of the black hole:

In the framework of classical GR studied her, a black hole is described by the exterior Schwarzschild metric combined with a singularity in the center or - more generally - by a variant of this metric that includes angular momentum, the so-called *Kerr metric*, which is given by the following line element:

$$\begin{aligned} ds^2 \;=\; & -\left(1-\frac{r_s r}{\rho^2}\right)c^2\mathrm{d}t^2 - \frac{2r_s r a\sin^2\theta}{\rho^2}\,c\,\mathrm{d}t\,\mathrm{d}\phi + \frac{\rho^2}{\Lambda^2}\mathrm{d}r^2 \\ & + \rho^2\mathrm{d}\theta^2 + \left(r^2+a^2+\frac{r_s r a^2}{\rho^2}\sin^2\theta\right)\sin^2\theta\;\mathrm{d}\phi^2\,. \end{aligned} \tag{7.88}$$

There are also other variants of the metric that take the electrical charge of a black hole into account.

Do these exotic black holes really exist? When I was studying, this question was still debated. Today, more than ten black holes have been identified in our galaxy. The black holes are classified according to their mass:

- *Stellar black holes* with up to 10 solar masses can be formed in the collapse of a star. They have a (Schwarzschild) radius of up to 30 km.
- *Moderately heavy black holes* arise from star collisions. They have about 1000 solar masses and have a radius of up to 1000 km. The existence of moderately heavy

[6] In the meantime, there have been speculations about another intermediate form, so-called *quark stars*, which consist of pure quarks.

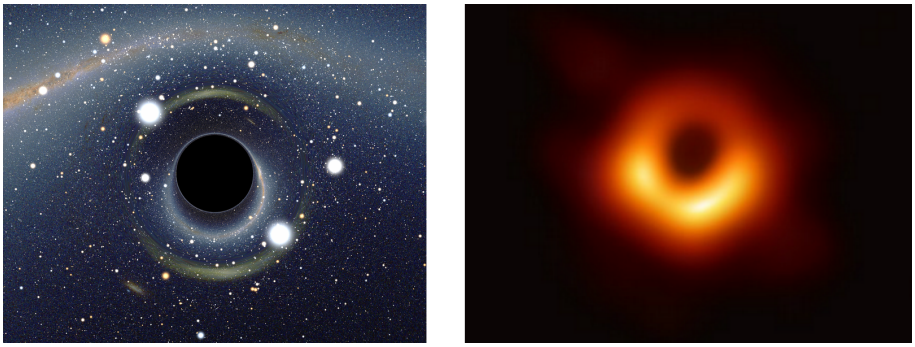


Figure 7.6: Gravitational lens. Left: Simulation of the optical impact of a black hole that would pass in front of the Magellanic Cloud [Wikimedia]. Right: Actual "photograph" of a black hole in 2019 [Event Horizon Telescope (EHT)].

black holes has not yet been proven beyond any doubt, but there are concrete candidates of this type.

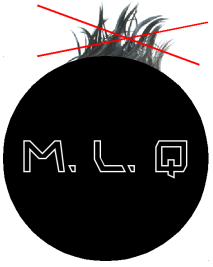
- *Supermassive black holes* with 10^5 up to 10^9 solar masses and can be found in the center of galaxies. It is believed that there is generically at least one supermassive black hole in the center of any galaxy.
- *Primordial black holes* are space-time singularities that could have formed immediately after the Big Bang and have a radius of one tenth of a millimeter. However, the existence of such micro-black-holes has not yet been proven.
- *Naked singularities* are very special black holes without an event horizon. They are mathematically possible but have not yet been observed so far.

How do you detect black holes? As illustrated in Fig. 7.6, black holes are not completely black, but they exhibit complex physical phenomena in the vicinity of the Schwarzschild radius. Black holes can, for example, emit jets or radio waves.

An important example is the supermassive black hole of our own galaxy, with has a weight of about 4 million solar masses. It is located in the Sagittarius constellation and is called *Sagittarius A**. The black hole is encircled by another star named S2 and thus allows a precise determination of the mass. Since the extremely heavy central body is not visible, it is assumed that it is a black hole.

Théorème de calvitie

It is J. A. Wheeler whom we owe the famous "*no hair theorem*". According to this theorem, a black hole is completely described by only three numbers, namely its mass M, its angular momentum L, and its electrical charge Q (however, the charge should quickly neutralize itself by preferential attraction of opposite charge carriers). The reason is that the Schwarzschild radius is an insurmountable information barrier, so it is basically impossible to learn anything about the "inner life" of a black hole.



No hair theorem

This theorem is remarkable because here a macroscopic object behaves exactly like an elementary particle, which can also be fully characterized by a few numbers. Is there a deeper connection between black holes and elementary particles? The no-hair theorem also poses a fundamental problem. Since a black hole has no information other than M, L, Q, its entropy is practically zero. However, this violates the 2nd law of thermodynamics: An object with a positive entropy $H > 0$ that moves through the Schwarzschild radius is irreversibly swallowed by the black hole. Since the black hole itself has no entropy, this process would lower the overall entropy.

Appendix: Frequently used symbols

$\circ$	Subsequent execution, concatenation
$\cong$	isomorphic
$\iota_X \omega$	Contraction of X with ω in the exterior algebra
$\lhd$	Normal subgroup
$\oplus$	Direct sum
$\otimes$	Tensor product
$\star$	Hodge star operator
*	belonging to the co-vector space
$\flat$	Musical isomorphism $V \to V^*$, lowering the index
$\sharp$	Musical isomorphism $V^* \to V$, raising the index
$\times$	Cartesian product
$\wedge$	Wedge product (antisymmetric tensor product)
$\mathbf{A}$	Vector potential of the electromagnetic field
GR	General Relativity
$\mathcal{C}$	Contraction of tensors
d	Exterior derivative of a differential form
D	Covariant derivative
$\mathbf{F}$	Fiels strength tensor of the electromagnetic field
$\mathbf{g}$	Metric tensor
$\mathbf{g}^*$	Metric tensor in the co-vector space
g	Determinant of the metric tensor
G	Newton's gravitational constant
$\mathbf{J}$	Current density of the electromagnetic field
$O(n)$	Orthogonal group in n diemsions
P_n	Permutation group of n objects
$SO(n)$	Special orthogonal group in n dimensions
SR	Special Relativity
s	Sign of the determinant of the metric tensor (-1)
ω	Volume form $\star(1)$
Z_2	Reflection group
Z_n	Cyclic group of n objects

Bibliography

[1] P. M. Schwarz and J. H. Schwarz, *Special Relativity*, Cambridge University Press, Cambridge, UK (2005) [TB Physik 750/UH 8200 S411]

[2] H. Goenner, *Spezielle Relativitätstheorie*, Spektrum, Elsevier, München (2004) [TB Physik 750/UH 8200 G595]

[3] S. M. Carroll, *Spacetime and geometry. An introduction to general relativity*, Sean Carroll. San Francisco, CA, USA: Addison Wesley (2004)

[4] T. Frankel, *The geometry of physics : an introduction*, Cambridge University Press, Cambridge, UK (2004).

[5] O. Grøn and S. Hervik, Einstein's General Theory of Relativity, Springer, New York (2007).

[6] J. B. Hartle, *Gravity : an introduction to Einstein's general relativity*, Addison Wesley, San Francisco, CA, USA (2003).

[7] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation*, Freemann, San Francisco (1973-2005) [UB UH 8500 M678]

[8] M. Nakahara, *Geometry, topology and physics*, Taylor & Francis, Boca Raton, USA (2003)

[9] T. Padmanabhan, *Gravitation: Foundations and Frontiers*, Cambridge University Press, Cambridge, UK (2010) [UB UH 8500 P123]

[10] B. Schutz, *Geometrical Methods of Mathematical Physics*, Cambridge University Press, Cambridge, UK (1980)

[11] R. U. Sexl and H. K. Urbantke, *Gravitation und Kosmologie*, Spektrum Akad. Verlag, Heidelberg (2002) [20/UH 8500 S518(5)]

[12] N. Straumann, *General Relativity: With Applications to Astrophysics*, Springer, New York (2004).

[13] R. d'Iverno, *Einführung in die Relativitätstheorie*, WILEY-VCH Verlag, Weinheim, 2.Auflage (2009).

[14] C. Rovelli, *Quantum Gravity*, Cambridge University Press, Cambridge, UK (2004).

[15] V. I. Arnold, Mathematical Methods of Classical Mechanics, Springer, New York (1989).

[16] H. Flanders, Differential Forms with Applications to the Physical Sciences, Aca-

demic Press, New York (1963).

[17] G. Lugo, *Differential Geometry and Physics*, Lecture Notes 2004, [http://people.uncw.edu/lugo/COURSES/DiffGeom/dg1.htm].

[18] B.-Y. Hou, *Differential geometry for physicists*, World Scientific, Singapore 1997.

[19] W. Kühnel, *Differentialgeometrie*, 5.Auflage, Vieweg und Teubner Verlag, Wiesbaden 2008.

Index