Introduction to General Relativity

Rainer F. Hauser
rainer.hauser@gmail.com

November 8, 2019

Abstract

Gravitation is the earliest recognized fundamental force of nature and found in Einstein's General Relativity so far its last complete theory despite the fact that General Relativity describing gravitation and Quantum Mechanics plus the Standard Model of Particle Physics describing the other three fundamental forces are still incompatible. Alex Flournoy from the Colorado School of Mines held lectures in 2019 covering 27 topics. His lectures were available on YouTube at the time this transcript has been assembled and may still be available today as "General Relativity (2019)" in 29 separate videos (two covering one topic and one from an earlier year). The first 18 lectures cover the theoretical side with manifolds, tensors and curvature, and the remaining 11 lectures touch on the practical side with the Schwarzschild solution, black holes and cosmology. A thorough basis of modern physics including classical mechanics and Special Relativity is required or, at least, very helpful.

1 Introduction

1.1 The Place of General Relativity in Physics

General Relativity addresses some of the big questions in physics such as cosmology and the nature of spacetime. It is one half of Quantum Gravity which is probably the most perplexing issue confronting theoretical physics. Despite its name it is, however, not a generalization of Special Relativity.

Correspondence principles show how one can move from a more powerful theory to a more limited theory which is easier to work with. The most limited mechanics is Newton's approach.

The following frameworks[1] of physics are somehow related to General Relativity:

- Newtonian Mechanics is the oldest and most limited framework. It is used for particles and fields restricted to $S \gg \hbar$ and $v \ll c$.
- Quantum Mechanics is used for particles and fields restricted to $v \ll c$. According to Bohr, Quantum Mechanics is true for large systems but the discreteness (spacing) is relatively small. A simple harmonic oscillator, for example, has energy levels $E_n = (n + \frac{1}{2})\hbar\omega$ where ω is the frequency of the oscillator, while the classical energy is $E = \frac{1}{2}m\omega^2 A$ where m is the mass and A the amplitude of the oscillation. For a big classical oscillator with $m = 1\,\mathrm{kg}$ and $A = 1\,\mathrm{m}$ oscillating relatively slowly with $\omega = 1\,\mathrm{Hz}$ it gives $n = 4.7 \cdot 10^{23}$ interpreted quantum-mechanically. With $\frac{\Delta E}{E} \approx 0$ the energy spectrum looks continuous.
  Another aspect is that classical behavior arises for many particles due to decoherence of wavefunctions between many degrees of freedom. (Exceptions are condensates.)
  Feynman brought Quantum Mechanics into the form of path integrals

$$e^{\frac{i}{\hbar}S} \qquad\qquad S = \int L\,dt \qquad\qquad L = E_{\text{kin}} - E_{\text{pot}}$$

  which is based on the so-called Lagrangian L.

[1] A framework is used for describing the evolution of a system, and a theory applies a chosen framework to a physical context. (For details see Script *Introduction to the Standard Model of Particle Physics – Part 1.*)

- Special Relativity is used for particles and fields restricted to $S \gg \hbar$. It replaces Galilean relativity. The difference between Newtonian Mechanics and Special Relativity is best explained through relativistic addition of velocities
$$v_3 = \frac{v_1 + v_2}{1 + \dfrac{v_1 v_2}{c^2}}$$
which reduces to simple addition of velocities $v_3 = v_1 + v_2$ for $v \ll c$.
- Quantum Field Theory is only used for fields while particles are just small fluctuations. Quantum Mechanics and Special Relativity are at odds with each other because Quantum Mechanics requires wavefunction normalization to conserve particle number and Special Relativity for particles allows creation and annihilation leading to changing particle number. One could combine them without fields, but it would be ugly and does not incorporate effects such as the Higgs mechanism.

In other words, Quantum Mechanics extends Newtonian mechanics to small systems, Special Relativity extends Newtonian mechanics to fast systems, and both are extended by Quantum Field Theory to small and fast systems.

For a theory one needs a framework plus degrees of freedom and some interactions. General Relativity is a theory and not a framework. It is the theory of the gravitational interaction. Thus there is a correspondence principle relating General Relativity to Special Relativity, and General Relativity is a generalization of Newtonian gravity needed when the ratio of mass m to characteristic length R and the ratio of the speed of light c squared to the gravitational constant G satisfies the inequality $m/R \geq c^2/G$. Special Relativity is one particular solution of General Relativity for a flat spacetime with no gravity acting at all. General Relativity was the first theory in physics starting with an unknown spacetime and not with a predefined spacetime. Another beautiful connection between Special and General Relativity is that one can start from Special Relativity, use a gauge principle similar to the ones in Particle Physics and ends up with General Relativity.

1.2 Fields and Test Particles in General Relativity

To relate General Relativity to something known, electrodynamics is used. Maxwell's equations
$$\vec{\nabla} \cdot \vec{E} = \frac{\rho}{\epsilon_0} \qquad \vec{\nabla} \times \vec{B} = \mu_0 \vec{j} + \mu_0 \epsilon_0 \frac{\partial \vec{E}}{\partial t} \qquad \vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \qquad \vec{\nabla} \cdot \vec{B} = 0$$
tell how sources $(\rho, \vec{j})$ create fields $(\vec{E}, \vec{B})$ with some topological constraints. The first two equations are the equations of motion of the electromagnetic fields, while the second two are the Bianchi identity and are geometric conditions.

One half of electromagnetism are Maxwell's equation telling how sources create fields, and the other half is how fields affect particles. The equation for the Lorentz force
$$\vec{F}_{\text{EM}} = q(\vec{E} + \vec{v} \times \vec{B}) \qquad \qquad \vec{F} = m\vec{a}$$
with Newton's law tells how particles react to the electromagnetic fields.

There is this split used a lot in physics where one uses big things creating backgrounds and then one uses small things as test particles to see what they do. A proton may be the massive thing which does not get influenced noticeable, and an electron may be the light thing placed in the electromagnetic field of the proton.

General Relativity mirrors this split. Einstein's equation (corresponding to Maxwell's equations) tells how sources create gravitational fields respectively curvature and represents one half of the theory. The geodesic equation (corresponding to the Lorentz force) tells how a test particle responds to curvature and represents the second half. The earth moving around the sun or a satellite moving around the earth are examples for test particles. (The gravitational field is not a field in the sense of the Newtonian gravity, but in describing the curvature of a spacetime the metric field is used which is perhaps the single most important element of General Relativity.) The metric field describes the geometry of spacetime including its curvature. One solves Einstein's equation to get the geometry of spacetime and looks for the extremal path (largest or smallest) by solving the geodesic equation.

1.3 Relativity

Relativity means that the laws of physics should take the same form to all observers in inertial frames. There are three symmetries of space and time which are relevant:

- Isotropy of space: If one takes the laboratory and rotates it, nothing should change.
- Homogeneity of space: If one takes the laboratory and translates it, nothing should change.
- Homogeneity of time: The laws of physics yesterday are the same as today and tomorrow.

The Galilean relativity assumes an absolute time everybody agrees on, and Einstein's relativity assumes that the speed of light is what everybody agrees on, but absolute time and constancy of speed of light are incompatible such that independent space and time have to be replaced by spacetime, and these three symmetries have to be replaced by

- Homogeneity of spacetime
- Isotropy of spacetime

because space and time are no longer separate entities. (Actually the fact that the speed of light is constant is because only light was known as massless at that time, but the constant speed is the one of any massless particle.)

The isotropy in spacetime is an important symmetry because much of Special Relativity is just rotations in four dimensions but there are other symmetries in physics. Symmetry is an incredibly powerful tool to help simplify calculations, and symmetry means that the action $S = \int L\, dt$ in classical mechanics or $S = \int \mathcal{L}\, d^4x$ relativistically (where L is a Lagrangian and $\mathcal{L}$ is a Lagrangian density) looks in both situations the same. A translation or a rotation in space or spacetime does not change the action S. A consequence of the invariance of the action is covariance of the equation of motion.

1.4 Symmetry, Groups and Representations

Symmetries can be static or dynamical. (A dynamical symmetry is one that leaves a Lagrangian unchanged.) They can also be global or local, discrete or continuous, finite or infinite, compact or non-compact, internal or spacetime. (If one coordinatize spacetime, then spacetime transformations also change coordinates while internal transformations do nothing to the coordinates.) Special Relativity is associated with spacetime symmetries, while the strong, weak and electromagnetic forces are associated with internal symmetries. To describe symmetry transformations mathematically, groups and representations are used[2].

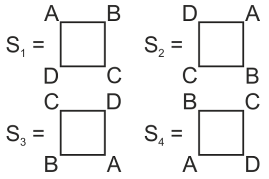
A *group* is a collection of elements $G = \{A, B, ...\}$ with a composition $\bullet$ that satisfies:

1. Closure: If $A, B \in G$ then $A \bullet B \in G$
2. Identity: There is some $I \in G$ such that $I \bullet A = A$ for any $A \in G$
3. Inverse: For any $A \in G$ there is an $A^{-1} \in G$ such that $A^{-1} \bullet A = I$
4. Associativity: $A \bullet (B \bullet C) = (A \bullet B) \bullet C$

The axioms identity and inverse will be very important in building invariants. A group with commutativity $A \bullet B = B \bullet A$ in addition to the above axioms is called *abelian*, but many groups needed in physics do normally not commute and are therefore non-abelian. If a subset of G satisfies all the above four axions then it is called a subgroup. Not all sets G with a composition $\bullet$ have an identity element. The cross product of vectors in three dimensions, for example, has no identity because the non-zero cross product is always orthogonal to any element multiplied and only the zero vector is orthogonal to any vector including itself.

Groups are often abstractly defined objects such as $\mathbb{Z}_2 = \{I, A\}$ which is specified through $I \bullet A = A \bullet I = A$ and $I \bullet I = A \bullet A = I$. (Finite groups can be defined by their multiplication table.) An abstract group can have many different concrete representations. Rotations in a plane about angles which are multiples of 180°, addition of odd and even numbers, and multiplication of 1 and -1 are examples of representations of $\mathbb{Z}_2$. Representations show how the group acts on things. In physics one works with representations of groups. Faithful representations give all information about a group.

[2] For a more detailed treatment see Script *Introduction to the Standard Model of Particle Physics – Part 1.*



Squares with labeled corners A, B, C, D, as an example, build a faithful representation of clockwise rotations about an angle which is a multiple of 90°. If opposite corners are labeled with the same letter then rotations about 180° are no longer distinguishable, and squares with opposite corners labeled the same build therefore not a faithful representation of the group of rotations about 90° in a plane. Unfaithful representations are also called degenerate. The most degenerate representation is the so-called identity representation where all four corners are labeled with the same letter A. One might argue that the identity representation is not very useful, but scalars, for example, are invariant under rotations and build therefore a representation with only one value. Also other objects which are not affected by a specific transformation are represented this way. The electron, for example, is not influenced by the strong force and is therefore invariant under rotations in color space.

In physics one often works with representations where the transformations are linear operators represented by matrices. This is, however, just a representation. In the example of the group of rotations by multiples of 90° the definitions

$$\mathrm{S}_1 = \begin{pmatrix}1\\0\\0\\0\end{pmatrix} \qquad \mathrm{S}_2 = \begin{pmatrix}0\\1\\0\\0\end{pmatrix} \qquad \mathrm{S}_3 = \begin{pmatrix}0\\0\\1\\0\end{pmatrix} \qquad \mathrm{S}_4 = \begin{pmatrix}0\\0\\0\\1\end{pmatrix}$$

$$\mathrm{R}_{0^\circ} = \begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{pmatrix} \quad \mathrm{R}_{90^\circ} = \begin{pmatrix}0&0&0&1\\1&0&0&0\\0&1&0&0\\0&0&1&0\end{pmatrix} \quad \mathrm{R}_{180^\circ} = \begin{pmatrix}0&0&1&0\\0&0&0&1\\1&0&0&0\\0&1&0&0\end{pmatrix} \quad \mathrm{R}_{270^\circ} = \begin{pmatrix}0&1&0&0\\0&0&1&0\\0&0&0&1\\1&0&0&0\end{pmatrix}$$

can be used where a column matrix is assigned to each of the four squares and a 4 × 4 matrix as an element of the group is assigned to each rotation. These matrices representing rotations behave as expected under matrix multiplication such that $\mathrm{R}_{90^\circ}\mathrm{R}_{180^\circ} = \mathrm{R}_{270^\circ}$ and so on. This is a faithful four-dimensional representation, but there are representations of this group with other dimensionalities. The smallest dimension giving a faithful representation is one where the group elements are complex numbers $e^{i\theta}$ with $\theta \in \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$.

1.5 Dual Representation, Invariance and Metric

Knowing how elements of a representation transform under an element of the group does not mean that one also knows how to build an invariant. A possible idea is that combining two objects which transform oppositely may give an invariant. That is exactly the correct way because the transformation combined with the opposite transformation cancel. For any matrix representation r one can form the *dual representation* $\tilde{r}$ defined such that if $A \in G$ then $\tilde{r} \to (A^{-1})^T\tilde{r}$ follows from $r \to Ar$. (This is analogous to the dot product $\vec{v} \cdot \vec{w}$ which results in a number. The transpose is reflected since $\vec{v}$ is a row vector and $\vec{w}$ is a column vector.) It follows because of

$$\tilde{r}^T r \to \left((A^{-1})^T\tilde{r}\right)^T Ar = \tilde{r}^T \left((A^{-1})^T\right)^T Ar = \tilde{r}^T A^{-1} Ar = \tilde{r}^T r$$

that $\tilde{r}^T r$ is invariant.

To find the dual representation $\tilde{r}$ for a given representation r one can use the *metric* g which is a mapping from an element of a representation r to a corresponding element of the dual representation $\tilde{r}$ by $\tilde{r} = gr$. The metric is always a symmetric matrix such that $g = g^T$. It follows from

$$\tilde{r}^T r = (gr)^T r = r^T g^T r = r^T gr \qquad\qquad r^T gr \to (Ar)^T gAr = r^T A^T gAr = r^T gr$$

for any $A \in G$ that

$$A^T gA = g \tag{1.1}$$

is the condition for $r^T gr$ to be invariant under the transformations $r \to Ar$.

Thus, one can use this property in two directions. On one side, given the group G with all the transformations A, one can find the metric g using (1.1) and the dual representation with the invariants $\tilde{r}^T r$.

On the other side, given a representation r and a metric g, one can use the condition (1.1) to find the transformations A which leave $r^T gr$ invariant. The second way is typically how one encounters symmetries in physics. One starts with stuff such as particles, fields, and other dynamical quantities which form a representation and finds using the metric g a set of transformations that are symmetries of $\tilde{r}^T r$. This way will be used below to uncover Special Relativity.

Before this is done, the first way is used to study rotations in three dimensions. The rotations are passive and thus transform coordinates and not objects. The corresponding group $G = \{R_x(\theta), R_y(\phi), R_z(\psi)\}$ is compact, continuous and non-abelian.

Given two objects in space whose positions are specified in a given rectangular coordinate system by (x_A, y_A, z_A) and (x_B, y_B, z_B), the distance between them should be invariant under a change of coordinates through a rotation. The distance is expressed as

$$\Delta s = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2} = \sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}$$

between these two positions. This can be interpreted as the dot product of a row vector and a column vector both with the coordinates Δx, Δy, Δz.

If a rotation is represented by a linear operator R then

$$\begin{pmatrix}\Delta x\\ \Delta y\\ \Delta z\end{pmatrix} \to \begin{pmatrix}\Delta x'\\ \Delta y'\\ \Delta z'\end{pmatrix} = R\begin{pmatrix}\Delta x\\ \Delta y\\ \Delta z\end{pmatrix} \qquad \Rightarrow \qquad (\Delta x, \Delta y, \Delta z) \to (\Delta x', \Delta y', \Delta z') = \left[R\begin{pmatrix}\Delta x\\ \Delta y\\ \Delta z\end{pmatrix}\right]^T$$

and

$$\begin{aligned}\Delta s^2 &= (\Delta x, \Delta y, \Delta z)\begin{pmatrix}\Delta x\\ \Delta y\\ \Delta z\end{pmatrix} = (\Delta x', \Delta y', \Delta z')\begin{pmatrix}\Delta x'\\ \Delta y'\\ \Delta z'\end{pmatrix} = \left[R\begin{pmatrix}\Delta x\\ \Delta y\\ \Delta z\end{pmatrix}\right]^T R\begin{pmatrix}\Delta x\\ \Delta y\\ \Delta z\end{pmatrix}\\ &= (\Delta x, \Delta y, \Delta z)R^T R\begin{pmatrix}\Delta x\\ \Delta y\\ \Delta z\end{pmatrix}\end{aligned}$$

such that $R^T R = I$ is the condition for Δs^2 to be invariant. Comparing to (1.1) shows that $R^T IR = I$ and $g = I$. The metric and the dot product are therefore

$$g = \begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix} \qquad\qquad \Delta s^2 = (\Delta x', \Delta y', \Delta z')\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}\begin{pmatrix}\Delta x\\ \Delta y\\ \Delta z\end{pmatrix}$$

in the case of rotations in the three-dimensional Euclidean space.

Now one can understand why g is called a metric. For spacetime transformations it plays a crucial role in defining distance. In fact it will help tremendously to think of coordinates as labels distinguishing points in space (or spacetime) with no intrinsic notion of distance. The distance is encoded in the metric.

In the case of polar coordinates r, θ in two dimensions the infinitesimal distance is

$$ds^2 = dr^2 + r^2 d\theta^2 = (dr, d\theta)\begin{pmatrix}1 & 0\\ 0 & r^2\end{pmatrix}\begin{pmatrix}dr\\ d\theta\end{pmatrix}$$

and not, as one might expect, $ds^2 = dr^2 + d\theta^2$.

2 Special Relativity

2.1 Basic Principle and Minkowski Metric

The premise of Special Relativity is that physical laws should not change under transformations continuously connected to the identity which preserve spacetime intervals $\Delta s^2 = -c^2\Delta t^2 + \Delta x^2 + \Delta y^2 + \Delta z^2$.

This immediately implies invariance under translations $t \to t' = t + \Delta t$, $x \to x' = x + \Delta x$ and so on. The metric and the formula for spacetime intervals are

$$\eta = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \qquad \Delta s^2 = (c\Delta t, \Delta x, \Delta y, \Delta z) \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} c\Delta t \\ \Delta x \\ \Delta y \\ \Delta z \end{pmatrix} \tag{2.1}$$

where η is the specific metric of *Minkowski space* in Cartesian coordinates. Any transformation Λ that satisfies $\Lambda^T \eta \Lambda = \eta$ corresponding to (1.1) will leave the interval invariant. Thus basically everything in Special Relativity including the constant speed of light follows from this equation.

2.2 Lorentz Transformations

There is obviously an infinite number of transformations satisfying $\Lambda^T \eta \Lambda = \eta$ because it is a continuous group, but similarly to rotations in the three-dimensional space one should be able to organize them into a small set of independent transformations each labeled by a continuous parameter. However the rotations in the three-dimensional space should not be interpreted as a rotation about an axis but as a rotation in a plane which is also true in higher dimensions. The rotation $R_x(\theta)$ is therefore the rotation$R_{yz}(\theta)$ and analogously for the other two axes. In three dimensions there is a one-to-one mapping between a plane and the direction of a normal vector, but this is no longer the case in four dimensions. The number of planes in a space with N dimensions is

$$\binom{N}{2} = \frac{N!}{(N-2)!\,2!} = \frac{1}{2} N(N-1)$$

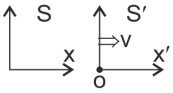
such that there are three planes in three dimensions and six planes in four dimensions.

One has to expect six independent transformations R_{xy}, R_{yz}, R_{zx}, R_{tx}, R_{ty}, R_{tz} each labeled with a continuous parameter. The first three R_{xy}, R_{yz}, R_{zx} are just *rotations* in space and do not do anything to time t. The other three R_{tx}, R_{ty}, R_{tz} are called *boosts*. Two of the corresponding transformations Λ as examples of *Lorentz transformations* are

$$\Lambda_{xy}(\theta) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\theta) & \sin(\theta) & 0 \\ 0 & -\sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \qquad \Lambda_{tx}(\phi) = \begin{pmatrix} \cosh(\phi) & -\sinh(\phi) & 0 & 0 \\ -\sinh(\phi) & \cosh(\phi) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{2.2}$$

corresponding to R_{xy} and R_{tx}. One can easily check that both matrices satisfy $\Lambda^T \eta \Lambda = \eta$ because $\sin(\theta)^2 + \cos(\theta)^2 = 1$ and $\cosh(\phi)^2 - \sinh(\phi)^2 = 1$. The matrix for Λ_{xy} looks familiar because it is just an ordinary rotation which leaves the z and t coordinates untouched. The matrix for $\Lambda_{tx}(\phi)$ however looks strange and has to be brought into a known form first.

Consider two frames S and S' where S' moves with respect to frame S in the x-direction with constant velocity v. From



$$\begin{pmatrix} c\Delta t \\ \Delta x \\ \Delta y \\ \Delta z \end{pmatrix} \to \begin{pmatrix} c\Delta t' \\ \Delta x' \\ \Delta y' \\ \Delta z' \end{pmatrix} = \begin{pmatrix} \cosh(\phi) & -\sinh(\phi) & 0 & 0 \\ -\sinh(\phi) & \cosh(\phi) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} c\Delta t \\ \Delta x \\ \Delta y \\ \Delta z \end{pmatrix} = \begin{pmatrix} \cosh(\phi)\, c\,\Delta t - \sinh(\phi)\,\Delta x \\ -\sinh(\phi)\, c\,\Delta t + \cosh(\phi)\,\Delta x \\ \Delta y \\ \Delta z \end{pmatrix}$$

and the fact that the origin O in S' is at rest in S' follows that $\Delta x' = 0 = -\sinh(\phi)\, c\,\Delta t + \cosh(\phi)\,\Delta x$ or $\frac{\Delta x}{\Delta t} = c\,\tanh(\phi)$. In S where the coordinates x and t are used, the origin O of S' moves with $v_x = v$ such that $\frac{\Delta x}{\Delta t} = c\,\tanh(\phi) = v$ or $\tanh(\phi) = \frac{v}{c}$. It follows

$$\cosh(\phi) = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \qquad \sinh(\phi) = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \frac{v}{c} \qquad \text{because } \cosh(\phi) = \frac{1}{\sqrt{1 - \tanh(\phi)^2}}$$

such that $\sinh(\phi)$ and $\cosh(\phi)$ can be replaced by expressions in v and c.

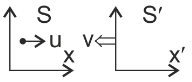
The transformation $\Lambda_{tx}(\phi)$ can now be written as

$$\Lambda_{tx}(\phi) = \begin{pmatrix} \gamma & -\gamma\frac{v}{c} & 0 & 0 \\ -\gamma\frac{v}{c} & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \begin{pmatrix} c\Delta t \\ \Delta x \\ \Delta y \\ \Delta z \end{pmatrix} \rightarrow \begin{pmatrix} c\Delta t' \\ \Delta x' \\ \Delta y' \\ \Delta z' \end{pmatrix} = \Lambda_{tx}(\phi) \begin{pmatrix} c\Delta t \\ \Delta x \\ \Delta y \\ \Delta z \end{pmatrix} = \begin{pmatrix} \gamma(c\Delta t - \frac{v}{c}\Delta x) \\ \gamma(\Delta x - v\Delta t) \\ \Delta y \\ \Delta z \end{pmatrix} \tag{2.3}$$

bringing the right side of (2.2) into a better known form where the quantity

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \tag{2.4}$$

is defined as usual to simplify the formulas.



If frame S' moves with respect to frame S in the direction of $-x$ with velocity v and an object in frame S moves with velocity u in the direction of x, then similar steps as above with $v \rightarrow -v$ give

$$\frac{\Delta x'}{\Delta t'} = \frac{\gamma \Delta x + \gamma v \Delta t}{\gamma \Delta t + \gamma \frac{v}{c^2}\Delta x} = \frac{\gamma \Delta x + \gamma v \Delta t}{\gamma \Delta t + \gamma \frac{v}{c^2}\Delta x} \frac{\frac{1}{\Delta t}}{\frac{1}{\Delta t}} = \frac{\gamma \frac{\Delta x}{\Delta t} + \gamma v}{\gamma + \gamma \frac{v}{c^2}\frac{\Delta x}{\Delta t}} = \frac{\frac{\Delta x}{\Delta t} + v}{1 + \frac{v}{c^2}\frac{\Delta x}{\Delta t}}$$

but $\frac{\Delta x}{\Delta t} = u$ and $\frac{\Delta x'}{\Delta t'} = u'$ such that

$$u' = \frac{v + u}{1 + \frac{vu}{c^2}} = \frac{\Delta x'}{\Delta t'} \tag{2.5}$$

which is the velocity addition formula of Special Relativity.

Time dilatation and length contraction are often introduced as the important properties of Special Relativity looking at time and space separately. However, it is better to think in terms of four dimensions where the quantity $\Delta s^2 = -c^2\Delta t^2 + \Delta x^2 + \Delta y^2 + \Delta z^2$ is invariant and does therefore not get changed by Lorentz transformations.

2.3 Orthogonal Groups

To do physics in a four-dimensional space alone does not make it Special Relativity. One can use the 4×4 identity matrix as the metric and call one coordinate t, but that does not give three rotations and three boosts but six rotations because this is not Minkowski space $\mathbb{M}^4$ but Euclidean space $\mathbb{R}^4$. The metric η is important with the fact that the three spacial coordinates and the one temporal coordinate get different signs. Two separate metrics one with a 1 in the first diagonal element and 0 in the others and the other one with a 0 in the first diagonal element and 1 in the others is called an $\mathbb{R}^3$ bundle over R^1 and corresponds to the Galilean spacetime where space and time are independent and which also has three boosts and three rotations.

Rotations R in the three-dimensional Euclidean space $\mathbb{R}^3$ satisfy $R^TR = I$ or $R^TIR = I$ which is an orthogonality condition. The group of transformations in $\mathbb{R}^3$ with $R^TR = I$ represented by matrices is called O(3). These matrices are called orthogonal, and they build together an *orthogonal group.*

Lorentz transformations Λ in the four-dimensional Minkowski space $\mathbb{M}^4$ satisfy $\Lambda^T\eta\Lambda = \eta$ which is also a kind of orthogonality condition but it is not O(4) because $\mathbb{M}^4$ is not $\mathbb{R}^4$. It is called O(1,3) because one coordinate is different than the others, and the dimension is called $1 + 3$ to distinguish it from 4.

The orthogonal groups O(3) and O(1,3) are too wide, because one would like to restrict them to the transformations that are continuously connected to the identity. This allows to build up any transformation by starting with the identity and applying many tiny transformations. This property of transformations has two advantages. On one hand it allows to do calculus with the transformations and eventually leads to Lie algebra structures, and on the other hand these types of transformations give rise to conserved quantities to be studied later.

Rotations or Lorentz transformations with only non-zero elements in the diagonal and where these diagonal elements are only $+1$ or -1 can either be ordinary rotations respectively Lorentz transformations if it has an even number of -1 or can be a parity transformation if it has an odd number of -1. If only one diagonal element is -1, then the corresponding coordinate is reflected, and that is a discrete transformation which is not continuously connected to the identity but satisfies the orthogonality condition.

To eliminate parity transformations the condition is $\det(R) = 1$ for rotations R and $\det(\Lambda) = 1$ for Lorentz transformations Λ. The determinant of a rotation R in any dimension can be ± 1 because $\det\big(R^T R\big) = \det\big(R^T\big)\det(R) = \det(R)^2$ and $\det(I) = 1$. Similarly, the determinant of a Lorentz transformation Λ can be ± 1 because $\det\big(\Lambda^T \eta \Lambda\big) = \det\big(\Lambda^T\big)\det(\eta)\det(\Lambda) = -\det(\Lambda)^2$ and $\det(\eta) = -1$. If the group elements with determinate -1 are eliminated the orthogonal group becomes a *special orthogonal group* which means that O(3) becomes SO(3) and O(1,3) becomes SO(1,3).

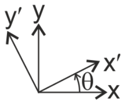
There is still one small issue with SO(1,3) in particular. One can prove that the upper left term of Λ which is Λ_{00} must satisfy $\Lambda_{00}^2 \geq 1$. This condition is satisfied by $\Lambda_{00} \geq 1$ and $\Lambda_{00} \leq -1$. Since $\Lambda_{00} = +1$ for the identity, one has to exclude transformations Λ with $\Lambda_{00} \leq -1$ which reverse time t. The resulting group of transformations is called the *proper orthochronous Lorentz group* denoted by SO(1,3)$^\uparrow$.

The complete symmetry group is ISO(1,3)$^\uparrow$ $= P^4 \ltimes$ SO(1,3)$^\uparrow$ with $4 + 3 + 3$ generators and is called *Poincaré group.* The group P^4 of translations contains the translations in space as well as in time. (The Poincaré group is a semidirect product indicated by $\ltimes$ because P^4 is a normal group while SO(1,3)$^\uparrow$ is not.)

The algebra of SO(1,3)$^\uparrow$ is $[R, R] = R$, $[B, B] = R$, $[R, B] = B$ where R is a rotation and B a boost. (The commutator $[A, B]$ is defined as $[A, B] = AB - BA$ and is therefore zero if A and B commute such that $AB = BA$.) This means that rotations build a subgroup of SO(1,3)$^\uparrow$, but boost do not.

Any two inertial observers in Special Relativity can be related by one of these transformations. Thus, Special Relativity is defined by the invariance of physical laws under the Lorentz transformations. To ensure that SO(1,3)$^\uparrow$ is a good symmetry of Special Relativity, one must always work with objects that transform in a well-defined way, and this also makes it easier. Any three-dimensional vector or scalar must be promoted to a four-dimensional vector or scalar with respect to SO(1,3)$^\uparrow$. Examples are $\vec{p} \to P^\mu$ and $t \to \tau$.

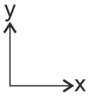
2.4 Spacetime Diagrams



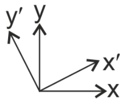
When one considers coordinate transformations, it is often useful to draw both the old and the new axes together to visualize what has changed. The matrix multiplication

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x\,\cos(\theta) + y\,\sin(\theta) \\ -x\,\sin(\theta) + y\,\cos(\theta) \end{pmatrix}$$

for a rotation in $\mathbb{R}^2$ results in the two equations $x' = x\,\cos(\theta) + y\,\sin(\theta)$ and $y' = -x\,\sin(\theta) + y\,\cos(\theta)$.



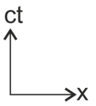
One can start instead from the old coordinate system and figure out where the new axes are. The x'-axis is where $y' = 0 = -x\,\sin(\theta) + y\,\cos(\theta) \Rightarrow y = x\,\tan(\theta)$. Similarly the y'-axis is where $x' = 0 = x\,\cos(\theta) + y\,\sin(\theta) \Rightarrow x = -y\,\cot(\theta)$. (Note that the slopes multiply to -1 as they should for orthogonal lines.)



The above equations build a system of two equations with the two unknowns x and y

$$\begin{vmatrix} -x\,\sin(\theta) + y\,\cos(\theta) & = 0 \\ x\,\cos(\theta) + y\,\sin(\theta) & = 0 \end{vmatrix}$$

and which is linear. However, one is not interested in the solution (the origin of the coordinate system) but the lines they represent which are the x'-axis and the y'-axis.

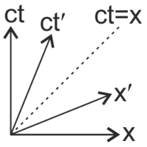


For Lorentz boosts along the x-axis (or a rotation in the tx-plane) an analog procedure is used. The matrix multiplication

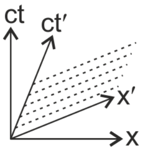
$$\begin{pmatrix} ct \\ x \end{pmatrix} \to \begin{pmatrix} ct' \\ x' \end{pmatrix} = \begin{pmatrix} \cosh(\phi) & -\sinh(\phi) \\ -\sinh(\phi) & \cosh(\phi) \end{pmatrix} \begin{pmatrix} ct \\ x \end{pmatrix} = \begin{pmatrix} ct\,\cosh(\phi) - x\,\sinh(\phi) \\ -ct\,\sinh(\phi) + x\,\cosh(\phi) \end{pmatrix}$$

for the rotation in $\mathbb{M}^2$ leads to $ct' = ct\,\cosh(\phi) - x\,\sinh(\phi)$ and $x' = -ct\,\sinh(\phi) + x\,\cosh(\phi)$ which are also two linear equations.

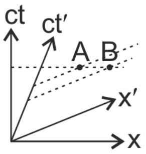
Similarly to the rotation in $\mathbb{R}^2$, the ct'-axis is where $x' = 0 = -ct\,\sinh(\phi) + x\,\cosh(\phi)$, and the x'-axis is where $ct' = 0 = ct\,\cosh(\phi) - x\,\sinh(\phi)$. This leads to $ct = x\,\coth(\phi)$ for the ct'-axis and $ct = x\,\tanh(\phi)$ for the x'-axis. The two axes ct' and x' are actually orthogonal to each other but drawing them in $\mathbb{R}^2$ instead of $\mathbb{M}^2$ cannot show this. This is sometimes called the scissoring effect of a boost.



One important feature of this is how to identify lines of constant time or position. Time t' is constant on lines parallel to the x'-axis, and similarly, position x' is constant on lines parallel to the ct'-axis. This means mathematically that $ct = x\,\tanh(\phi) + \text{constant}$ for constant time t' because $ct' = \text{constant} = ct\,\cosh(\phi) - x\,\sinh(\phi)$. Thus, lines of constant time and position in the coordinate system ct' and x' are not perpendicular to each other similar to the axes.

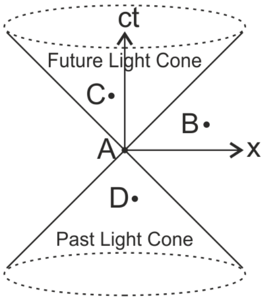


If two events A and B are simultaneous in the frame S with the coordinates ct and x, they lie on a line parallel to the x-axis. In the frame S' with the coordinates ct' and x' they do however not lie on the same line parallel to the x'-axis. This is the loss of absolute time in Special Relativity as compared to Galilean relativity. Whether two events happen at the same time depends therefore on the frame with the observer. In one frame A may precede B and in another frame B may precede A.



2.5 Causal Structure

In physics the notion of cause and effect are crucial. In Galilean relativity causality is encoded in the statement that causes must precede effects and that can be done consistently because two simultaneous events in absolute time are always simultaneous. However, when it is no longer clear what the temporal relations are between two events then causality has to be carefully reevaluated. In Special Relativity the causal structure has the form of a cone and is called the *light cone.*



Each *event* in spacetime has a past and a future light cone associated with it where an event is just a point in spacetime. The cone is the set of points with $x^2 + y^2 + z^2 = c^2t^2$ whose intersection with the plane of ct and x is the two lines at ±45°. The cones for event A are shown in the figure. Event A can only influence future events with less than the speed of light such that the surfaces of the cones are connected to A by signals moving at the speed of light c. This means that A can only influence events in the future light cone, and that it can only have been influenced by events in the past light cone. The causal structure makes therefore that A could cause C, and D could cause A and C, but A could not cause B.

The light cones replace the causal structure of Galilean relativity with absolute time. If the event A is the birth of a person, then this person can only influence (and experience) events in his or her future light cone and have been influenced by events in his or her past light cone. Whatever has happened in the universe outside this past light cone will be a secret forever for this person.

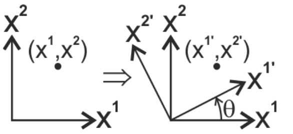
3 The Mathematics of Tensors

3.1 Index Notation

So far matrices have been used to represent vectors, transformations and the metric. One advantage of matrices is that one can use them to represent things that do not commute, but they have at least three disadvantages. Because they do not commute, one has to be careful about order when writing expressions with matrices. They can become very big such that writing them out explicitly may be a problem. The most important disadvantage is however that there are objects and operations one cannot represent by matrices and matrix multiplication. Index notation[3] will remove these disadvantages and keep all the advantages of matrices such as the possibility to represent non-commuting objects.

[3] Computers use internally index notation when they process matrices.

The example of a rotation about θ in $\mathbb{R}^2$ is used to introduce index notation. The rotation is



$$\begin{pmatrix} x^1 \\ x^2 \end{pmatrix} \to \begin{pmatrix} x^{1'} \\ x^{2'} \end{pmatrix} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \end{pmatrix}$$

written using matrix multiplication. When using the notation

$$\begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} = \begin{pmatrix} \Lambda^{1'}{}_1 & \Lambda^{1'}{}_2 \\ \Lambda^{2'}{}_1 & \Lambda^{2'}{}_2 \end{pmatrix}$$

instead, one can write

$$x^i \to x^{i'} = \Lambda^{i'}{}_j \, x^j$$

using the Einstein summation convention which demands that repeated indices are summed.

Several important points to note:

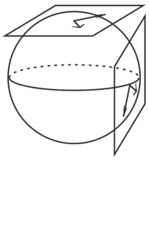
1. The number of dimensions does not matter because $x^i \to x^{i'} = \Lambda^{i'}{}_j \, x^j$ does not specify the dimension which has to be known from the context.
2. One can also write $x^i \to x^{i'} = x^j \, \Lambda^{i'}{}_j$ because order does not matter while it does matter for matrices.
3. One knows immediately how to evaluate something like $M_{ijk} \, N^{ijk}$ given the elements of M_{ijk} and N^{ijk} even though one cannot represent this in terms of matrices.
4. Repeated indices always come in upper (also called contravariant) and lower (also called covariant) index pairs as, for example, in $\Lambda^{i'}{}_j \, x^j$.
5. In the four dimensions of spacetime greek indices such as μ, ν and λ are used. They take values 0, 1, 2, 3 corresponding to (ct, x, y, z). When talking about space alone, latin indices such as i and j are used.
6. In some cases objects will come with one primed and one unprimed index as in $T^{\mu'}_{\nu} X^{\nu}$ indicating a transformation from the unprimed to the primed coordinate system. Any object other than a coordinate transformation should only have either unprimed or primed indices.
7. In cases where objects in index notation can be represented by matrices and one wants to represent them by matrices, one has to be careful with the order. The rule to go from index notation to matrix expressions is that the repeated sum indices must be immediately adjacent. Thus, $X^{\nu} \, T^{\mu'}_{\nu}$, for example, must first be reordered to $T^{\mu'}_{\nu} X^{\nu}$.

3.2 Vectors and Tangent Spaces

Vectors are elements of some vector space with the corresponding axioms. Here a vector is always either a vector in spacetime such as X^{μ} or a vector in space alone such as $\Delta\vec{x}$ or $\vec{E}$. They correspond to physical quantities with a magnitude and a direction. They have three important features:

1. The vector itself is invariant under coordinate changes. (The vector has different components in different coordinate systems, but the actual physical quantity does not change.)
2. When one chooses coordinates, the vectors have components which satisfy a specific transformation law. The vector is not just the components but the components together with the basis vectors, and this combination is invariant. (The mathematical description changes but not the physical quantity.)
3. Despite that vectors usually have been introduced in terms of entities connecting two positions in space, vectors do not in general exist within a space or spacetime. This is because the usual rules for vector manipulation in linear algebra requires these objects to exist in flat spaces but spaces and spacetimes can be curved.

The last point leads to the question how one works with intrinsically flat objects in general curved spaces. The sphere S^2 embedded in $\mathbb{R}^3$ is a curved two-dimensional space. At each point in the space a *tangent space* is defined as the set of tangent vectors to all curves passing through that point. Vectors at each point in the space live in these tangent spaces. In flat spaces all the tangent spaces are parallel, but in curved spaces they are not.



This has two important consequences:

a) One cannot freely move vectors around the space since in general the tangent spaces change.
b) Comparing vectors defined at two different points in a space will be tricky.

To uncover how vector components transform under coordinate changes, one can start with the particularly simple vector ds. The components of this vector are just coordinate differentials dX^μ. Because coordinates transform as $X^\mu \to X^{\mu'} = \Lambda^{\mu'}{}_\nu X^\nu$ and the components of ds are little segments of coordinates they must transform as $dX^\mu \to dX^{\mu'} = \Lambda^{\mu'}{}_\nu \, dX^\nu$. Thus, one knows how one particular kind of vectors transforms, but one would like a rule for the transformation of the components of an arbitrary vector.

A vector is invariant under coordinate transformation and only its components change. In the (orthonormal) basis of $\hat{e}_{(\mu)}$ the infinitesimal displacement is $ds = dX^\mu \, \hat{e}_{(\mu)}$. If ds is invariant and the components transform as $dX^\mu \to dX^{\mu'} = \Lambda^{\mu'}{}_\nu \, dX^\nu$, one can conclude

$$\begin{aligned} ds = dX^\mu \, \hat{e}_{(\mu)} \to ds' = dX^{\mu'} \, \hat{e}_{(\mu')} = ds& \\ = \Lambda^{\mu'}{}_\mu \, dX^\mu \, \hat{e}_{(\mu')} = \Lambda^{\mu'}{}_\mu \, dX^\mu \, \Lambda^\lambda{}_{\mu'} \, \hat{e}_{(\lambda)} = \Lambda^\lambda{}_{\mu'} \, \Lambda^{\mu'}{}_\mu \, dX^\mu \, \hat{e}_{(\lambda)} = dX^\mu \, \hat{e}_{(\mu)}& \end{aligned}$$

guessing $\Lambda^\lambda{}_{\mu'} \, \hat{e}_{(\lambda)}$ as the transformation of the basis vectors based on indices. This indicates that

$$\Lambda^\lambda{}_{\mu'} \, \Lambda^{\mu'}{}_\nu = \delta^\lambda{}_\nu$$

must be satisfied, and this obviously means $\Lambda^{-1} \, \Lambda = I$. (Note that $\Lambda^\mu{}_{\mu'}$ is the inverse and $\Lambda_\mu{}^{\mu'}$ the transpose of $\Lambda^{\mu'}{}_\mu$. Because $\Lambda^T \eta \, \Lambda = \eta$, one can conclude $\Lambda^T \neq \Lambda^{-1}$.)

Given the transformation properties of the basis vectors and the fact that vectors are invariant, the general transformation laws are

$$V^\mu \to V^{\mu'} = \Lambda^{\mu'}{}_\nu \, V^\nu \qquad \hat{e}_{(\mu)} \to \hat{e}_{(\mu')} = \Lambda^\lambda{}_{\mu'} \, \hat{e}_{(\lambda)} \qquad \Lambda^\lambda{}_{\mu'} \Lambda^{\mu'}{}_\nu = \delta^\lambda{}_\nu \tag{3.1}$$

for vector components V^μ and basis vectors $\hat{e}_{(\mu)}$. To figure out whether an object is really a vector, one can check whether it has the correct transformation properties.

3.3 Dual Vectors and Cotangent Spaces

One can define dual vectors by three conditions:

1. Dual vectors are straight directed objects defined at a point in space and live in a *cotangent space*.
2. Dual vectors are invariant but given a coordinate system they can be expressed in terms of components and dual basis vectors which do transform.
3. Dual vectors linearly absorb vectors and produce scalars.

To illustrate the last point, consider a dual vector ω and the linear combination $a\,V + b\,W$ of two vectors. The fact that the dual vector linearly absorbs this vector means $\omega(aV+bW) = a\,\omega(V)+b\,\omega(W)$, and that the result is a scalar. A *scalar* is an invariant whose explicit coordinate representation is also invariant.

As vectors are specified in a basis $\hat{e}_{(\mu)}$ and with components $V = V^\mu$ as $V^\mu \, \hat{e}_{(\mu)}$, dual vectors are similarly specified in a basis $\hat{\theta}^{(\mu)}$ and with components ω_μ as $\omega = \omega_\mu \, \hat{\theta}^{(\mu)}$. To define what it means that a dual vector consumes a vector it is because of the linearity sufficient to define it as

$$\hat{\theta}^{(\mu)} \, \hat{e}_{(\nu)} = \delta^\mu{}_\nu \tag{3.2}$$

for the basis vectors and the dual basis vectors. The transformation laws for vectors and dual vectors are

$$\begin{aligned} \hat{e}_{(\mu)} \to \hat{e}_{(\mu')} &= \Lambda^\mu{}_{\mu'} \, \hat{e}_{(\mu)} \qquad & V^\mu \to V^{\mu'} &= \Lambda^{\mu'}{}_\mu \, V^\mu \\ \hat{\theta}^{(\mu)} \to \hat{\theta}^{(\mu')} &= \Lambda^{\mu'}{}_\mu \, \hat{\theta}^{(\mu)} \qquad & \omega_\mu \to \omega_{\mu'} &= \Lambda^\mu{}_{\mu'} \, \omega_\mu \end{aligned} \tag{3.3}$$

because $\delta^\mu{}_\nu$ is a scalar and therefore invariant and the transformation properties of the basis vectors $\hat{e}_{(\nu)}$ are known.

The meaning of $\omega(V)$ becomes clear

$$\omega(V) = \omega_\mu\, \hat{\theta}^{(\mu)}\, V^\nu\, \hat{e}_{(\nu)} = \omega_\mu\, V^\nu\, \hat{\theta}^{(\mu)}\, \hat{e}_{(\nu)} = \omega_\mu\, V^\nu\, \delta^\mu{}_\nu = \omega_\mu\, V^\mu = \omega_0\, V^0 + \omega_1\, V^1 + \omega_2\, V^2 + \omega_3\, V^3 \in \mathbb{R}$$

because of the linearity. This looks like a dot product between two vectors but is actually a combination of a vector and a dual vector. The dot product in $\mathbb{R}^n$ is $V \cdot W = \delta_{ij}\, V^i\, W^j$ where δ_{ij} is the metric on $\mathbb{R}^n$. For Special Reality the metric $\eta_{\mu\nu}$ is used such that $\eta_{\mu\nu} V^\mu W^\nu = V_\mu W^\mu$. The rules are

$$V^\mu \to V_\mu = \eta_{\mu\nu}\, V^\nu \qquad V_\mu \to V^\mu = \eta^{\mu\nu}\, V_\nu \qquad \eta^{\mu\nu} = (\eta_{\mu\nu})^{-1} \tag{3.4}$$

for turning vectors in dual vectors and vice versa in Special Relativity. Thus the metric raises and lowers indices. Note that in Special Relativity $\eta^{\mu\nu} = \eta_{\mu\nu}$ but this is not true in general. (In the following the basis vectors $\hat{e}_{(\mu)}$ and the dual basis vectors $\hat{\theta}^{(\mu)}$ are dropped assuming that (3.2) is satisfied, and only components will be used.)

3.4 Tensors

The following two properties describe *tensors*:

1. Tensors represent physical quantities that are invariant but when given as explicit coordinate representation will typically have components that transform.
2. Tensors exist in flat tangent or cotangent spaces or tensor products of these at each point.

One can label tensors with (tangent,cotangent) classifying scalars as $(0,0)$, vectors as $(1,0)$, dual vectors as $(0,1)$, the metric $\eta_{\mu\nu}$ as $(0,2)$, the inverse metric $\eta^{\mu\nu}$ as $(2,0)$ and so on. In general one could have tensors of type (m,n) with m upper and n lower indices. The tensor $T^\kappa{}_{\lambda\mu\nu}$, for example, is of type $(1,3)$ and is actually fully specified $T = T^\kappa{}_{\lambda\mu\nu}\, \hat{e}_{(\kappa)} \otimes \hat{\theta}^{(\lambda)} \otimes \hat{\theta}^{(\mu)} \otimes \hat{\theta}^{(\nu)}$.

In a formal way tensors can be defined as multi-linear maps from the space of vectors and dual vectors into the real numbers[4] or, mathematically, $T(V, ..., V, \omega, ..., \omega) = \text{scalar}$. This definition assumes a "well-fed" tensor such as $T^\kappa{}_{\lambda\mu\nu}\, V^\lambda\, V^\mu\, V^\nu\, \omega_\kappa \in \mathbb{R}$ with no "free" indices in contrast to a "starving" tensor such as $T^\kappa_{\lambda\mu\nu}\, V^\lambda\, V^\mu = F^\kappa{}_\nu$ or an "over-fed" tensor such as $T^\kappa{}_{\lambda\mu\nu}\, V^\lambda\, V^\mu\, V^\nu\, V^\xi\, \omega_\kappa = G^\xi$.

Another definition of a tensor states that a tensor is something that transforms like a tensor which is kind of tautological but is still a good definition. The coordinate transformation of tensor components is determined by their index structure. Scalars c, vectors V, dual vectors ω and higher-order tensors T transform analogously to (3.3) as

$$\begin{array}{ccc} c \to c' = c & V^\mu \to V^{\mu'} = \Lambda^{\mu'}{}_\mu V^\mu & \omega_\mu \to \omega_{\mu'} = \Lambda^\mu{}_{\mu'} \omega_\mu \\ T^{\mu\nu} \to T^{\mu'\nu'} = \Lambda^{\mu'}{}_\mu \Lambda^{\nu'}{}_\nu T^{\mu\nu} & T^\mu{}_\nu \to T^{\mu'}{}_{\nu'} = \Lambda^{\mu'}{}_\mu \Lambda^\nu{}_{\nu'} T^\mu{}_\nu & T_{\mu\nu} \to T_{\mu'\nu'} = \Lambda^\mu{}_{\mu'} \Lambda^\nu{}_{\nu'} T_{\mu\nu} \end{array}$$

and so on. (Note that in $\Lambda^{\mu'}{}_\mu \Lambda^{\nu'}{}_\nu T^{\mu\nu}$ the two transformations $\Lambda^{\mu'}{}_\mu$ and $\Lambda^{\nu'}{}_\nu$ are the same transformation. It would not make sense if one is a rotation and the other a boost, for example.)

A very special tensor is the metric $\eta_{\mu\nu}$. It provides the notion of distance because of $\Delta s^2 = \eta_{\mu\nu}\, \Delta X^\mu\, \Delta X^\nu$, but it also takes vector indices to their corresponding dual vector indices as shown in (3.4). Also other tensors may give scalars as in $T_{\mu\nu}\, \Delta X^\mu\, \Delta X^\nu = c$ or dual vectors as in $T_{\mu\nu}\, V^\nu = W_\mu$, but it is not the distance nor the corresponding dual vector. In this sense the metric $\eta_{\mu\nu}$ and the inverse metric $\eta^{\mu\nu}$ are special. Some other equations which are only true for the metric are

$$\eta_{\lambda\mu}\, \eta^{\mu\nu} = \delta_\lambda{}^\nu \qquad \eta^{\kappa\nu}\, \eta^{\lambda\mu}\, \eta_{\mu\nu} = \eta^{\kappa\nu}\, \eta_{\nu\mu}\, \eta^{\mu\lambda} = \eta^{\kappa\nu}\, \delta_\nu{}^\lambda = \eta^{\kappa\lambda} \qquad \eta_{\mu\nu}\eta^{\mu\nu} = \eta_{\nu\mu}\eta^{\mu\nu} = \delta_\nu{}^\nu = 4$$

because the metric is a tensor and one can use it as any other tensor of the same order. The fact that the metric is symmetric makes it easier to come clear with the indices.

Coordinates X^μ are not vectors (because only ΔX^μ is a vector) but they transform like vectors. The question is how derivatives $\frac{\partial}{\partial X^\mu}$ of coordinates transform. Because $\frac{\partial}{\partial X^\mu}(X^\nu) = \delta_\nu{}^\mu$ is a number, $\frac{\partial}{\partial X^\mu}$ must transform oppositely of X^μ and $\frac{\partial}{\partial X^\mu}$ usually written as ∂_μ transforms as a dual vector.

[4] In Special and General Relativity numbers are always real numbers. In other fields of physics there can also be tensors based on complex numbers.

3.5 Tensor Equations

The physical world has no predetermined absolute coordinate system (inertial or otherwise) and truly fundamental laws of physics should be built out of quantities which are at least invariant under transformations connecting inertial frames. This means that they should be built out of tensors.

Scalar equations are fine but limited. One would like to write equations like $A^{\mu\nu} = B^{\mu}\, C^{\nu}{}_{\lambda}\, D^{\lambda} + E^{\mu\nu\kappa}\, F_{\kappa}$. Such equations in components are not invariant because they transform when the coordinates change. However this is fine as long as the left-hand side of the equation transforms exactly the same way as the right-hand side because one can insert all the basis vectors and dual basis vectors and then both sides are invariant.

Tensors can also contain things like $\partial_\mu A_\nu$ but index symmetries should also agree on both sides of the equation. Given $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ then $F_{\nu\mu}$ must be $-F_{\mu\nu}$. As an example illustrating the use of tensors, Maxwell's equations

$$\vec{\nabla} \times \vec{B} - \partial_t \vec{E} = \vec{j} \qquad \vec{\nabla} \cdot \vec{E} = \rho \qquad \vec{\nabla} \times \vec{E} + \partial_t \vec{B} = 0 \qquad \vec{\nabla} \cdot \vec{B} = 0$$

can be used. Each vector product contains three equations such that these are eight equation in total. While it is obvious that these equations are invariant under rotations, it is not clear that they are invariant under boosts.

If one introduces

$$J^\mu = \begin{pmatrix} \rho \\ j^1 \\ j^2 \\ j^3 \end{pmatrix} \qquad F_{\mu\nu} = \begin{pmatrix} 0 & -E^1 & -E^2 & -E^3 \\ E^1 & 0 & B^3 & -B^2 \\ E^2 & -B^3 & 0 & B^1 \\ E^3 & B^2 & -B^1 & 0 \end{pmatrix}$$

as a vector J^μ and a $(0,2)$ field strength tensor $F_{\mu\nu}$ which is antisymmetric as $F_{\mu\nu} = -F_{\nu\mu}$, then one can express Maxwell's equations as

$$\partial_\mu F^{\mu\nu} = J^\nu \qquad \partial_{[\mu} F_{\nu\lambda]} = 0$$

in terms of tensors which are also eight equations in total. One knows therefore that they are invariant under all Lorentz transformations including boosts. They predict c and, as a consequence, c is invariant under boosts.

The brackets in the second equation demand that one completely antisymmetrize over the exchange of indices. Antisymmetrizing $T_{\mu\nu}$ means $T_{[\mu\nu]} = \frac{1}{2}(T_{\mu\nu} - T_{\nu\mu})$. Because there are only the four possibilities 012, 013, 023, 123 for $\mu\nu\lambda$, the first and the second equation represent four equations each. One of the four equations encoded in $\partial_{[\mu} F_{\nu\lambda]} = 0$ is $\frac{1}{6}(\partial_0 F_{12} + \partial_1 F_{20} + \partial_2 F_{01} - \partial_0 F_{21} - \partial_1 F_{02} - \partial_2 F_{10}) = 0$.

4 Applications to Physics

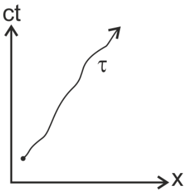
4.1 Relativistic Kinematics and Dynamics

In three dimensions with $i \in \{1,2,3\}$ for Newtonian mechanics

$$x_i(t) \to v_i(t) = \frac{dx_i}{dt} \to a_i(t) = \frac{dv_i}{dt} \qquad \sum F_i(t) = m\, a_i(t) = \frac{dp_i}{dt}$$

describe the kinematics on the left side and the dynamics on the right side where $p_i = mv_i$. Time t is a universal, invariant, monotonically increasing parameter which can parametrize motion.

In four-dimensional spacetime with $\mu \in \{0,1,2,3\}$ for Special Relativity $x_i \to X^\mu$ and $v_i \to U^\mu$. One problem is the question what U^0 is, and a second problem is that $U^\mu = dX^\mu/dt$ is not a vector nor any other kind of tensor because dX^μ is a vector in spacetime but dt is a component of a vector. In Special Relativity t is no longer universal and invariant as it was in Newtonian mechanics. Thus another parameter is needed which is universal, invariant and monotonically increasing.



The length of the worldline $s = \int \sqrt{ds^2}$ looks like a candidate for a universal, invariant and monotonically increasing quantity, but because $ds^2 = -c^2dt^2+dx^2+dy^2+dz^2 < 0$ for $v < c$ (and especially in a rest frame with $v = 0$), instead of s the proper time $\tau = \int \sqrt{|ds^2|}$ is used. In the rest frame with $ds^2 = -c^2dt^2$ the quantity $\tau = c\int dt$ is the "rest time". The classification for infinitesimal displacements where $ds^2 < 0$ is called *timelike* for matter, $ds^2 = 0$ is called *lightlike* for light (massless particles), and $ds^2 > 0$ is called *spacelike* for tachyonic motions can be extended to all vectors.

In Special and General Relativity often units are used such that the speed of light is one and distances are measured in light-seconds. One can always restore factors of c by making the units consistent. From now on the convention $c = 1$ is also used.

Defining 4-velocity U^μ and 4-momentum P^μ using the mass m as

$$U^\mu = \frac{dX^\mu}{d\tau} \qquad\qquad P^\mu = m\,U^\mu \tag{4.1}$$

gives the result

$$U^\mu U_\mu = \eta_{\mu\nu}U^\mu U^\nu = \eta_{\mu\nu}\frac{dX^\mu}{d\tau}\frac{dX^\nu}{d\tau} = \frac{\eta_{\mu\nu}dX^\mu dX^\nu}{d\tau^2} = \frac{\eta_{\mu\nu}dX^\mu dX^\nu}{-ds^2} = \frac{\eta_{\mu\nu}dX^\mu dX^\nu}{-\eta_{\mu\nu}dX^\mu dX^\nu} = -1$$

which seems strange because one expects $U^\mu U_\mu$ to be the speed squared similar to $\vec{v}\cdot\vec{v} = v^2$ in Newtonian mechanics. However comparing it with

$$\vec{v} = \left(\frac{dx}{ds}, \frac{dy}{ds}, \frac{dz}{ds}\right) \qquad\qquad \vec{v}\cdot\vec{v} = \frac{dx^2}{ds^2} + \frac{dy^2}{ds^2} + \frac{dz^2}{ds^2} = 1$$

where ds is the path length shows that this is just a consequence of how the paths are parametrized.

The components of the 4-velocity U^μ are

$$U^0(\tau) = \frac{dt}{d\tau} = \frac{\gamma dt_{\text{rest}}}{d\tau} = \gamma\frac{d\tau}{d\tau} = \gamma \qquad U^i(\tau) = \frac{dx^i}{d\tau} = \frac{dx^i}{dt}\frac{dt}{d\tau} = \gamma v^i \qquad \gamma = \frac{1}{\sqrt{1-v^2}}$$

in the frame S with coordinates (t, x, y, z). Note that the velocity $\vec{v}$ is the velocity of the particle and the v in γ connects the frame S to the frame S_{rest}, but they are luckily the same velocity. In the rest frame is $U^0 = 1$ and $U^i = 0$, and in general one can write

$$U^\mu = \begin{pmatrix}\gamma \\ \gamma\vec{v}\end{pmatrix} \qquad\qquad P^\mu = \begin{pmatrix}m\gamma \\ m\gamma\vec{v}\end{pmatrix}$$

for the 4-velocity U^μ and the 4-momentum P^μ.

For $v \ll 1$ (where 1 means c) the approximation $\gamma \approx 1 + \frac{1}{2}v^2 + ...$ gives $m\gamma \approx m + \frac{1}{2}mv^2 + ... = E$ and $m\gamma v^i \approx mv^i + ... = p^i$ for the 4-momentum because $\frac{1}{2}mv^2$ is the non-relativistic kinetic energy[5] and mv^i the non-relativistic momentum. Therefore $P^0 = E$ is the relativistic energy with the term m, $\vec{p}$ with the three components P^i is the relativistic momentum, and one gets

$$P^\mu = mU^\mu = \begin{pmatrix}E \\ \vec{p}\end{pmatrix} \qquad P_\mu P^\mu = m^2U_\mu U^\mu = -m^2 = -E^2 + p^2 \qquad E^2 = p^2 + m^2 \tag{4.2}$$

where the equation $E^2 = p^2 + m^2$ (or $E^2 = |\vec{p}^2|c^2 + m^2c^4$ with $c \neq 1$) is called the *mass-shell condition*.

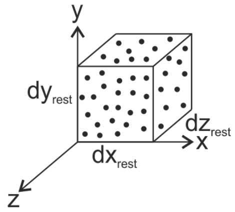
Even though the parametrization does not work when $m^2 = 0$, this result does. One can now relate the distinction between timelike, lightlike and spacelike relation between two events to the 4-momentum. Then timelike means $P_\mu P^\mu < 0$ or $m^2 > 0$, lightlike means $P_\mu P^\mu = 0$ or $m^2 = 0$ and spacelike (or tachyonic) means $P_\mu P^\mu > 0$ or $m^2 < 0$.

One could try to relativize forces for the dynamics, but the primary concern here is the gravitational force which will play out a bit differently.

[5] The famous formula $E = mc^2$ states that mc^2 is the energy of a particle at rest. This term also exists in the non-relativistic approximation for $v \ll c$ and is therefore not a consequence of Special Relativity.

4.2 Densities and the Energy-Momentum Tensor

Given the energy-momentum vector P^μ the question is why one also needs the energy-momentum tensor $T^{\mu\nu}$. In General Relativity the sources of gravity are usually large, so instead of thinking of one particle at a time, a large number of them should be considered. The source of the gravitation and therefore of the curvature is energy and momentum where energy includes mass.



Because densities may vary throughout space and time one works with infinitesimal quantities. The *particle number density* where dN is the number of particles in the small volume $dV = dx\,dy\,dz$ can be written as $dn = \frac{dN}{dV}$ in form of the infinitesimal particle number density. One can define dn_{rest} where the volume dV_{rest} is at rest. The number of particles dN is a scalar, and the question is whether dn_{rest} is a tensor. Boosting the infinitesimal volume along x by $-v$ results in

$$dn \to dn' = \frac{dN}{\frac{1}{\gamma}\,dx\,dy\,dz} = \gamma\frac{dN}{dV} = \gamma dn > dn$$

and shows that it is not a tensor. Because $dn' = \gamma dn_{\text{rest}}$ is similar to $dt' = \gamma dt_{\text{rest}}$ one might guess that dn is the time-component of a 4-vector. If U^μ is the 4-velocity of the infinitesimal volume then

$$dN^\mu = dn_{\text{rest}}U^\mu = \begin{pmatrix} dn_{\text{rest}}\gamma \\ dn_{\text{rest}}\gamma\vec{v}\end{pmatrix} = \begin{pmatrix} dn \\ dn\,\vec{v}\end{pmatrix} \Rightarrow dN^\mu_{\text{rest}} = \begin{pmatrix} dn_{\text{rest}} \\ \vec{0}\end{pmatrix}$$

then this is a vector because $dn_{\text{rest}} = \frac{dN}{dV_{\text{rest}}}$ is constant and U^μ is a vector. (Note that the infinitesimal volume is at rest when the average velocity of all the particles moving in different directions is zero.) To summarize, one had to introduce vector dN^μ in order to make a density when starting with a scalar dN.

There is a deeper explanation for why creating a tensor density seems to raise the tensor nature of what one starts with. In the above case the scalar as a $(0,0)$ tensor was raised to a vector as a $(1,0)$ tensor. To specify a two-dimensional area in three dimensions a size and a direction (a normal dual vector) is needed. A two-dimensional surface in three dimensions is defined by $f(x,y,z) = 0$. A sphere S^2_R centered around the origin is, for example, specified by $f(x,y,z) = x^2 + y^2 + z^2 - R^2 = 0$. The quantity normal to the surface is $(\partial_x f, \partial_y f, \partial_z f)$ and this is a dual vector.

To specify a three-dimensional volume in four dimensions one needs also a size and a direction where the size is dV and the direction is specified by the dual vector n_μ. In $n_\mu\,dV$ the dual vector n_μ is orthogonal to the three-dimensional volume dV. (It does not matter whether this is in a four-dimensional space or in three-plus-one-dimensional spacetime.)

Given a density $\frac{dN}{dV}$ using $dN = \frac{dN}{dV}\,dV$ is not well-suited in spacetime. With the density dN^μ as a 4-vector and the volume $n_\mu\,dV$ the expression $dN = dN^\mu\,n_\mu\,dV$ is a scalar.

The goal is to create a density from the 4-momentum P^μ to allow distribution of energy and momentum. More specific, the goal is to find $dP^\mu = (?)\,n_\nu\,dV$ and this needs a $(2,0)$ tensor $T^{\mu\nu}$ which is called the *energy-momentum-tensor*. This tensor plays an important role in General Relativity because it is going to provide the source that creates curvature and gravitational interaction. The energy-momentum tensor is symmetric such that $T^{\mu\nu} = T^{\nu\mu}$ and that it only has ten independent components. It can be represented by a 4×4 matrix and is a $(2,0)$ tensor.

If one considers $n_\mu = (1,0,0,0)$ and $dV = dx\,dy\,dz$ then $dP^\mu = T^{\mu\nu}\,n_\nu\,dV = T^{\mu 0}\,n_0\,dV = T^{\mu 0}\,dV$. Because P^0 is the relativistic energy E and P^i the i-th component of the relativistic momentum $\vec{p}$ this gives

$$dP^0 = T^{00}\,dV \Rightarrow T^{00} = \frac{dE}{dV} = \rho \qquad\qquad dP^i = T^{i0}\,dV \Rightarrow T^{i0} = \frac{dp^i}{dV} = \pi^i$$

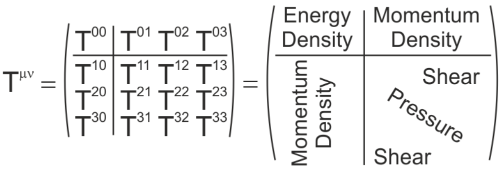
which are the *energy density* ρ and the *momentum density* π^i.

Similarly, if one considers $n_\mu = (0,1,0,0)$ (choosing x as spatial direction) and $dV = dt\,dy\,dz = dt\,dA_{yz}$ then $dP^\mu = T^{\mu\nu}\,n_\nu dV = T^{\mu 1}\,n_1 dV = T^{\mu 1}\,dt\,dA_{yz}$. Using the fact again that P^i is the i-th component of the relativistic momentum $\vec{p}$ leads to

$$T^{11} = \frac{dp^1}{dt}\frac{1}{dA_{yz}} = F^1_{\text{net}}\frac{1}{dA_{yz}} \qquad\qquad T^{21} = \frac{dp^2}{dt}\frac{1}{dA_{yz}} = F^2_{\text{net}}\frac{1}{dA_{yz}}$$

where T^{11} is the pressure on dA_{yz} and T^{21}, T^{31} are the shear on dA_{yz}. (The derivative of the momentum is a force, the component of a force perpendicular to an area and divided by the size of an area is a pressure, and a component of a force parallel to an area is a shear.)

To summarize, the energy-momentum tensor is a $(2,0)$ tensor where T^{00} is the energy density, the three components $T^{0i} = T^{i0}$ represent the momentum density, the spacial diagonal components T^{ii} correspond to the pressure, and the remaining components specify the shear. In the following mostly only the diagonal terms of the energy-momentum tensor will be needed to be non-zero.



The components of the energy-momentum tensor can be specified as

$$T^{\mu\nu} = \frac{dP^\mu}{n_\nu\, dV} = \frac{dP^\mu\, U^\nu}{dV_{\text{rest}}} \tag{4.3}$$

because the unclear mathematical meaning of a term where a vector is divided by a dual vector coming from $dP^\mu = T^{\mu\nu}\, n_\nu\, dV$ can be resolved by recalling $dN^\mu = dn_{\text{rest}} U^\mu$ and concluding

$$dN^\mu = dn_{\text{rest}}\, U^\mu = dN \frac{U^\mu}{dV_{\text{rest}}} \qquad \text{and} \qquad dN^\mu = \frac{dN}{n_\mu\, dV} \qquad \Rightarrow \qquad \frac{1}{n_\mu\, dV} = \frac{U^\mu}{dV_{\text{rest}}}$$

which is a useful result also in other situations. This allows to define the system (the volume) in the rest frame giving some size with the 4-velocity relative to the rest frame.

As a first example this is applied to dust which is defined as a collection of particles at rest with respect to each other. In the rest frame the volume sits still and in the volume the particles sit still. The question is what is the energy-momentum tensor in an arbitrary frame. The infinitesimal 4-momentum is $dP^\mu = dN\, m\, U^\mu$ where U^μ is the same velocity for all particles and $P^\mu = m\, U^\mu$ is the momentum of the individual dust particle. Using (4.3) gives

$$T^{\mu\nu} = \frac{dP^\mu\, U^\nu}{dV_{\text{rest}}} = \frac{dN\, m\, U^\mu\, U^\nu}{dV_{\text{rest}}} = dn_{\text{rest}}\, m\, U^\mu\, U^\nu = \rho_{\text{rest}}\, U^\mu\, U^\nu$$

where ρ_{rest} is the rest energy density because at rest m (actually mc^2) is all energy there is and dn_{rest} is the particle density. In the rest frame one gets

$$U^\mu = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \qquad\qquad T^{\mu\nu}_{\text{rest}} = \begin{pmatrix} \rho_{\text{rest}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

for the energy-momentum tensor of dust. This is not surprising and one could have guessed this result because the particles have no momentum, and the shear and pressure are also obviously zero.

As a second example a perfect fluid is chosen which is defined as a collection of particles with random velocities in an overall rest frame. Typically one can ignore viscosity (shear) so that $T^{ij} = 0$ for $i \neq j$ and assume isotropy so that $T^{11} = T^{22} = T^{33} = p$ as part of the definition of a perfect fluid. Also the momentum density is zero because there is no net flow. Thus one can guess

$$T^{\mu\nu}_{\text{rest}} = \begin{pmatrix} \rho & 0 & 0 & 0 \\ 0 & p & 0 & 0 \\ 0 & 0 & p & 0 \\ 0 & 0 & 0 & p \end{pmatrix}$$

for the energy-momentum tensor in the rest frame with only an energy density ρ and pressure p. One can write this as $T^{\mu\nu}_{\text{rest}} = (\rho + p)\, U^\mu_{\text{rest}}\, U^\nu_{\text{rest}} + p\, \eta^{\mu\nu}$ giving

$$T^{\mu\nu}_{\text{rest}} = \begin{pmatrix} \rho & 0 & 0 & 0 \\ 0 & p & 0 & 0 \\ 0 & 0 & p & 0 \\ 0 & 0 & 0 & p \end{pmatrix} = \begin{pmatrix} \rho + p & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} -p & 0 & 0 & 0 \\ 0 & p & 0 & 0 \\ 0 & 0 & p & 0 \\ 0 & 0 & 0 & p \end{pmatrix}$$

which is useful because it is the tensor expression $T^{\mu\nu} = (\rho + p)\, U^\mu\, U^\nu + p\, \eta^{\mu\nu}$ constructed in the rest frame but true in any frame.

This is the advantage of working with tensors. One can find an equation in a simple setting (often the rest frame) and the equation is true in any reference frame if it is a tensor expression. If pressure p is set to zero, the tensor expression found for the perfect fluid becomes $T^{\mu\nu} = \rho\, U^\mu\, U^\nu$ and this is what has been found for dust in the previous example. The equation

$$T^{\mu\nu} = (\rho + p)\, U^\mu\, U^\nu + p\, \eta^{\mu\nu} \tag{4.4}$$

is true for any perfect fluid in any frame. In General Relativity the energy-momentum tensor $T^{\mu\nu}$ will play the role of sources for curvature in a manner analogous to how the charge-current density J^μ does for electromagnetism, and one can consider all sources as perfect fluids. The only thing left to be specified is how ρ and p are related.

The energy-momentum tensor $T^{\mu\nu}$ is an important component in Einstein's equations, and the metric $g_{\mu\nu}$ is a function to be determined by solving these equations. There are a few things to keep in mind:

	Dimension	Metric	Inverse Metric	Vectors and Dual Vectors	Path of a Particle
Galilean Space	3	$g_{ij} = \delta_{ij}$	$\delta_{ij} = \delta^{ij}$	$v^i = v_i$	$x^i(t)$
Minkowski Spacetime	$3+1$	$g_{\mu\nu} = \eta_{\mu\nu}$	$\eta_{\mu\nu} = \eta^{\mu\nu}$	$V^\mu \neq V_\mu$	$X^\mu(\tau)$
Curved Spacetime	$3+1$	$g_{\mu\nu}(X^\lambda)$	$g_{\mu\nu} \neq g^{\mu\nu}$	$V^\mu \neq V_\mu$	$X^\mu(\tau)$

Note that the metric is also in flat space not independent of the location and becomes $g_{\mu\nu}(X^\lambda)$ as it is in curved space as soon as one chooses coordinate systems such as polar coordinates other than Cartesian coordinates.

4.3 Relativistic Newtonian Gravity

A question which comes up here is whether curved spacetime is really needed. It may be possible to build a relativistic version of Newton's gravity. Newton's law of gravity is

$$\vec{F}_{G_{12}} = \frac{G\, m_1\, m_2}{|\vec{r}_1(t) - \vec{r}_2(t)|^3}\, (\vec{r}_1(t) - \vec{r}_2(t))$$

and uses the same t for two distant points $\vec{r}_1$ and $\vec{r}_2$ which implies that the gravitational influence is communicated with infinite speed. If one tries to fix this with the gravitational influence communicated by c, orbits become unstable. The relativistic gravity fails therefore.

This may surprise because a very similar theory that does work with finite speed is the Coulomb interaction

$$\vec{F}_{C_{12}} = \frac{k\, q_1\, q_2}{|\vec{r}_1(t) - \vec{r}_2(t)|^3}\, (\vec{r}_1(t) - \vec{r}_2(t))$$

because the full treatment of electrodynamics with magnetic fields, the Liénard–Wiechert potential and so on makes it completely consistent.

One might try a similar path for gravity which would lead to gravito-magnetic effects, but there are two big problems. It is still based on mass m and misses the observed gravitational effects on particles with $m = 0$, and it does not even get the predictions right when $m \neq 0$. (However, it serves as a good approximation to General Relativity in the weak-field limit.) Despite all these attempts one needs a completely different starting point than Newtonian gravity.

4.4 Equivalence Principles

From considerations of Newtonian gravity $\vec{F}_G = m_G \vec{g}$ and Newtonian mechanics $\sum \vec{F} = m_I \vec{a}$ where $\vec{g}$ is the gravitational force and $\vec{a}$ the acceleration, one observes the so-called *weak equivalence principle* $m_G = m_I$. In words, the inertial mass which is the resistance to acceleration independent of the kind of forces acting and the gravitational mass which plays a role in one specific force are perfectly equal according to all kinds of experiments. If the gravitational force is the only force acting then $\vec{g} = \vec{a}$.

The weak equivalence principle $\vec{g} = \vec{a}$ means that for massive objects and external gravity, a uniform external gravitational field is indistinguishable from a uniform acceleration as illustrated in figure 4.1 (a) with a little person throwing a ball in a lab which is considered small compared to the earth[6]. One can extend the weak equivalence principle to the *Einstein equivalence principle* stating that for any object and any force (except internal gravitation) a uniform external gravitational field is indistinguishable from uniform acceleration. This shows that gravitation does not only act on matter with $m \neq 0$ but also on light with $m = 0$ as shown in the figure (b) where the little person uses a flashlight. (Note that an observer in an inertial frame sees the light beam as a straight line.)

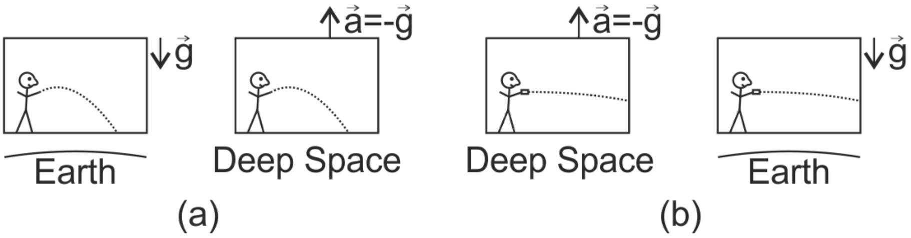


Figure 4.1: The effect of the equivalence principles

It is obvious that a theory with the Einstein equivalence principle is going beyond the Newtonian law of gravity. It is sufficient to motivate the basics of General Relativity, but one can go further. The *strong equivalence principle* states that for any object and any force a uniform external gravitational field is indistinguishable from a uniform acceleration. The restriction to external gravitational fields can be removed, and this singles out General Relativity over alternative theories such as the Brans–Dicke theory. This is the strongest version of the equivalence principle because all restrictions have been removed. The difference between it and Einstein's version only shows up when very large objects are involved.

Newtonian mechanics and Special Relativity share the concept of inertial frames which play a crucial role. In Newtonian mechanics the inertial frames are those in which one can use Newton's second law, and in Special Relativity the framework is based on the tenet that physics including c is unchanged when viewed from different inertial frames. However the question is how one can define respectively find inertial frames. Newton's first law states that in an inertial frame an object experiencing no net force will move with a constant velocity.

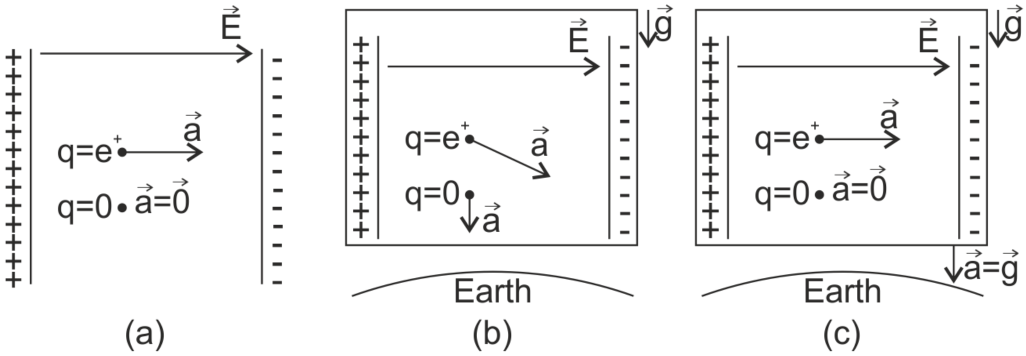


Figure 4.2: Inertial frames in the presence of gravity

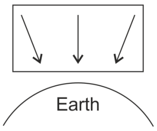
Aside from trying to balance multiple forces, the easiest way to establish an inertial frame is to find an object which is unaffected by the forces present. Then this object defines an inertial frame in which one can use Newton's second law to describe the motion of objects that do feel forces. This works well for electromagnetism as shown in figure 4.2 (a) where a charged particle with $q = e^+$ feels the uniform

[6] The gravitational field of the earth is not really uniform because the earth is a sphere but if the lab is small compared to the earth it looks like being uniform.

electric field as the force $\vec{F} = q\vec{E}$ while an uncharged particle with $q = 0$ is not affected by the electric field and can be used to establish an inertial frame. However the problem with gravity is that the Einstein equivalence principle states that there are no objects which are unaffected by gravity as shown in figure (b). Consequently no object can define an inertial frame. Gravity is in this sense very special because it is the only force everything feels and everything feels the same way.

Einstein realized that a reference frame which is freely falling under the influence of external gravity is the best scenario for establishing inertial frames as illustrated in figure (c). By going to a freely-falling frame in the presence of gravity one actually mimics the behavior one expects to see in deep space without gravity. The key to this working is the universality of how gravity influences objects. Everything experiences $\vec{a} = \vec{g}$ regardless of mass. This type of universal influence is one more reason to suspect that gravity is less like the other forces and instead is tied to a universal feature like spacetime.

In all the equivalence principles it has been made clear that the gravitational field is uniform since the canceling accelerations are uniform. If the lab is too big one starts seeing nonuniform tidal effects. Restriction to small regions of space and time is needed such that the final refinement of the Einstein equivalence principle can be phrased: Experiments performed in a small, freely-falling lab over a short time give results that are indistinguishable to those in an inertial frame in empty space.



5 Geometry of Curved Space

5.1 Manifolds

Special Relativity states that the laws of physics are invariant under transformations connecting inertial frames, and spacetime is isotropic in space and homogenous in spacetime. General Relativity in contrast states that the laws of physics are invariant under diffeomorphisms of spacetime and the connection (gauge-field) facilitating this invariance should be rendered dynamical by the introduction of an invariant field strength tensor. Similar to any gauge theory there is on one side a symmetry which is here the invariance under diffeomorphism in spacetime and there is on the other side a gauge field which should be able to propagate. The symmetry describes how things behave on curved space and respond to gravity, and the gauge field describes how curvature arises dynamically from sources.

The task of describing the symmetry is similar to what has been done in Special Relativity with identifying the transformations and expressing how tensors are transformed, but the coordinate systems and the metric are no longer globally defined. Since each region can be made flat exclusively of the others, there is no preferred coordinate system. (Coordinates are really not physical, thus much of this can be considered in a coordinate independent form.) The fact that each region can be made flat is exactly the Einstein equivalence principle because in a small region in space and time a freely-falling lab experiences what looks like an inertial frame, a piece of spacetime without gravity.

This has two important implications. Firstly one can use Special Relativity in a freely-falling frame since this correctly describes physics in deep space without gravity. Secondly when looking for what type of spacetimes are allowed in General Relativity, they must be locally flat, and this leads to manifolds.

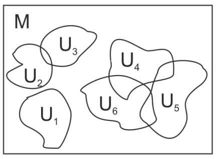
The following nomenclature will be used:

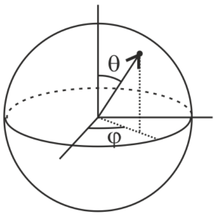
$$A: \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad B: \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad C: \begin{pmatrix} a & ... & ... & ... \\ ... & b & ... & ... \\ ... & ... & c & ... \\ ... & ... & ... & ... \end{pmatrix} \quad D: \begin{pmatrix} -a & ... & ... & ... \\ ... & b & ... & ... \\ ... & ... & c & ... \\ ... & ... & ... & ... \end{pmatrix}$$

For the metric $g_{\mu\nu}$ case A specifies an Euclidean space, case B a Minkowski space, case C a Riemannian space which has an Euclidean signature, and case D a pseudo-Riemannian space which has a Lorentzian signature. (The numbers a, b, c, ... are assumed to be positive.) The metric of the Euclidean space and of the Minkowski space assume this simple form only in in Cartesian coordinates.

A C^p n-dimensional *manifold* is defined as a set M with a maximal atlas. The set M here is a collection of points in space or spacetime. (Nothing is said yet about coordinates.)

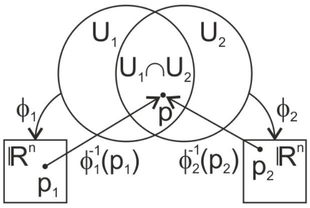
The set M which is assumed to be n-dimensional can be broken up into pieces called charts where all the charts cover all of M. A *chart* is defined as a subset U of M with a one-to-one map $\phi : U \to \mathbb{R}^n$ such that the image of ϕ is open in $\mathbb{R}^n$. (A one-to-one map in contrast to a one-to-many or a many-to-one map is also called injective. If $a' = \phi(a)$ and $b' = \phi(b)$ then two conditions are fulfilled: Firstly if $a \neq b$ then $a' \neq b'$, and secondly if $a' \neq b'$ then $a \neq b$. Open means that the boundary is not included.) These maps just mean the ability to choose coordinates represented as n-tuples in $\mathbb{R}^n$ for a patch U.





The sphere S^2 embedded in $\mathbb{R}^3$ as an example can be specified as $x^2 + y^2 + z^2 = R^2$ but to identify every point on the surface two values are sufficient. One can use the polar angle θ and the azimuthal angle φ, and therefore (θ, φ) are coordinates for a curved space taking values in $\mathbb{R}^2$. Thus all that is required of charts is that one can setup coordinates as in this example. The mapping is from a possibly curved space to a flat (Euclidean) space $\mathbb{R}^n$. (Note that in this example the condition of injectivity is not satisfied. The coordinate φ is undefined for $\theta = 0$.)

An *atlas* on M is a collection of charts $\{(U_\alpha, \phi_\alpha)\}$ such that the union of all charts U_α is equal to M and the charts are sewn together with C^p transition functions.



Supposing the two charts U_1 and U_2 overlap, and a point p is in the overlapping part $U_1 \cap U_2$. With $\phi_1(p) = p_1$ and $\phi_2(p) = p_2$ and the fact that both maps are injective, $\phi_1^{-1}(p_1) = \phi_2^{-1}(p_2) = p$. Because both points p_1 and p_2 are in $\mathbb{R}^n$, the function $\phi_1 \bullet \phi_2^{-1}$ defined as

$$p_1 = \phi_1(p) = \phi_1 \bullet \phi_2^{-1}(p_2)$$

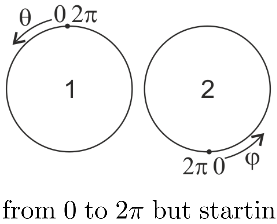
for all $p \in U_1 \cap U_2$ is a map $\mathbb{R}^n \to \mathbb{R}^n$. With $p_1 = (x_1, y_1, ...)$ and $p_2 = (x_2, y_2, ...)$ the *transition function* $\phi_{12} = \phi_1 \bullet \phi_2^{-1}$ gives

$$x_1 = f_x(x_2, y_2, ...) \qquad\qquad y_1 = f_y(x_2, y_2, ...) \qquad\qquad ...$$

and these are just coordinate changes. (Note that this only makes sense if p_2 satisfies $p_2 = \phi_2(p)$ for a $p \in U_1 \cap U_2$.)

Because the transition functions are functions $\mathbb{R}^n \to \mathbb{R}^n$ calculus and especially derivatives can be used. If the p-th derivative of all transition functions $\phi_{\alpha\beta}$ exists and is continuous, the $\phi_{\alpha\beta}$ is C^p. The function f with $f(x) = x$ for $x \geq 0$ and f with $f(x) = -x$ for $x < 0$ is C^0 because the first derivative has a discontinuity at $x = 0$. The sine and cosine functions on the other hand are examples of C^∞ functions. If $\phi_{\alpha\beta}$ is C^∞ and invertible the it is called a *diffeomorphism*. The coordinate transformations used in the following will almost always be diffeomorphisms.

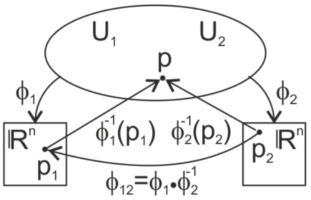
Assuming that there is a set M which has two different atlases satisfying all conditions then these are (if one leaves out the word “maximal” in the definition of a manifold) two different manifolds. This is not nice because two manifolds in this situations should be the same independent of the atlas chosen as long all conditions are satisfied. Thus instead of a single atlas one uses an equivalence class of atlases. Every single atlas one can imagine to cover a manifold is in this equivalence class. This basically means that one can take one manifold and work with different atlases which are all equivalently good representations of the manifold. One important consequence of the maximality of the atlas is that the C^p condition must hold on all atlases even when two charts completely overlap.



In the following all these definitions will be applied to the example of the manifold S^1 to prove that the simple circle is a manifold. To show this, one has to construct one representative element from the maximal atlas. The circle S^1 is one-dimensional but it will turn out that one needs at least two charts. The first chart chosen uses θ to parametrize the circle from 0 to 2π starting from top, and the second chart chosen uses φ to parametrize the circle also from 0 to 2π but starting from the bottom such that $\theta, \varphi \in (0, 2\pi]$. To map S^1 to $\mathbb{R}^1$ an open subset of $\mathbb{R}^1$ is needed, but $(0, 2\pi]$ is closed on one side and contains 2π. Restricting θ and φ both to the open interval $U_1 = U_2 = (\frac{\pi}{2} - \varepsilon, \frac{3\pi}{2} + \varepsilon)$ and defining $\phi_1(\theta) = x_1$ and $\phi_2(\varphi) = x_2$ delivers the two charts building the atlas. If ε is just a small positive number, the two charts overlap at $\pi \pm \varepsilon$ and $3\pi \pm \varepsilon$ and cover all of S^1. The transition function $\phi_{12} = \phi_1 \bullet \phi_2^{-1}$ is $x_1 = \phi_{12}(x_2) = x_2 + \pi$ and belongs to C^∞.

5.2 General Coordinate Transformations

One cannot always cover M with one chart and can therefore not always have a global coordinate system on M. Only if one can cover M with one chart then one can set up a global coordinate system. Otherwise one uses coordinate transformations to piece together patches. In the case that M can be covered with one chart, one can use two charts covering all of M but equipped with different maps.



One can write a coordinate change as $X^{\mu'}(X^\mu)$ where the new coordinates $X^{\mu'}$ are a function of the old coordinates X^μ. The mapping ϕ_{12} changes the coordinates of $\mathbb{R}^n$ on the right side of the figure into the coordinates of $\mathbb{R}^n$ on the left side. Thus one can ignore that there is a set M inbetween and just study the change of coordinates in $\mathbb{R}^n$. The question is how the derivatives of coordinate changes behave.

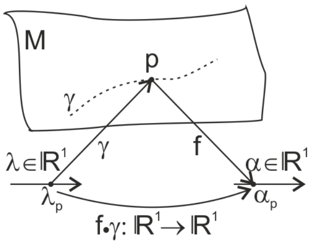
The chain rule states that $\frac{d}{dx}g(f(x)) = \frac{d}{dx}(g \bullet f)(x) = \frac{dg}{df}\frac{df}{dx}$. With two variables

$$f_1(x,y) \;\; f_2(x,y) \;\; g_1(f_1,f_2) \;\; g_2(f_1,f_2) \Rightarrow \quad \begin{array}{ll} \frac{\partial g_1}{\partial x} = \frac{\partial g_1}{\partial f_1}\frac{\partial f_1}{\partial x} + \frac{\partial g_1}{\partial f_2}\frac{\partial f_2}{\partial x} & \frac{\partial g_1}{\partial y} = \frac{\partial g_1}{\partial f_1}\frac{\partial f_1}{\partial y} + \frac{\partial g_1}{\partial f_2}\frac{\partial f_2}{\partial y} \\ \frac{\partial g_2}{\partial x} = \frac{\partial g_2}{\partial f_1}\frac{\partial f_1}{\partial x} + \frac{\partial g_2}{\partial f_2}\frac{\partial f_2}{\partial x} & \frac{\partial g_2}{\partial y} = \frac{\partial g_2}{\partial f_1}\frac{\partial f_1}{\partial y} + \frac{\partial g_2}{\partial f_2}\frac{\partial f_2}{\partial y} \end{array}$$

and this can be written with index notation as

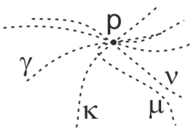
$$\frac{\partial}{X^\mu} = \frac{\partial X^{\mu'}}{\partial X^\mu}\frac{\partial}{\partial X^{\mu'}} \qquad\qquad \frac{\partial}{X^{\mu'}} = \frac{\partial X^\mu}{\partial X^{\mu'}}\frac{\partial}{\partial X^\mu} \tag{5.1}$$

for the transformation laws for partial derivatives.

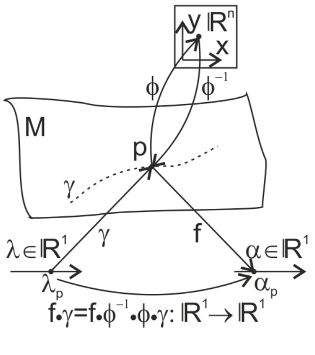


Vectors and dual vectors exist in tangent and cotangent spaces, respectively. Before they are described in terms of coordinates, vectors can be defined without introducing coordinates. Given a manifold M, a point $p \in M$ and a curve γ through p, this curve is assumed to have been parametrized through $\gamma : \mathbb{R}^1 \to M, \lambda \to \gamma(\lambda)$ which only gives points on γ such as $\gamma(\lambda_p) = p$. Now an additional C^∞ map $f : M \to \mathbb{R}^1, p \to f(p) = \alpha_p$ which is defined for all $p \in M$ is introduced. This gives a function $f \bullet \gamma : \mathbb{R}^1 \to \mathbb{R}^1, \lambda_p \to \alpha_p$ with no reference to M.

Defining $\frac{df}{d\lambda}$ as the change in α when λ changes gives $\frac{df}{d\lambda} = \frac{d}{d\lambda}(f \bullet \gamma)$. The parameter λ is associated with the specific chosen curve. The derivative $\frac{df}{d\lambda}$ is called the directional derivative of f along λ. This derivative shows how the function $f : M \to \mathbb{R}^1$ varies as one moves along a given curve. Using a different curve will give a different directional derivative.



If one now considers all curves through the point p parametrized with different parameters $\kappa, \lambda, \mu, \nu, ...$ then the set of directional derivatives for these curves independent of a selected function f forms a vector space $\{\frac{d}{d\kappa}, \frac{d}{d\lambda}, \frac{d}{d\mu}, \frac{d}{d\nu}, ...\}$. (All the axioms for a vector space are satisfied.) This construct establishes a vector space of tangents to M at p and can therefore act as the *tangent space*.



The tangent space has been defined independent of coordinates. In order to combine it with charts and introduce a coordinate representation of vectors, one can ask how components would transform under coordinate transformations. A chart with $p \in U$ and ϕ allows to write $f \bullet \gamma$ in an environment around p as $f \bullet \phi^{-1} \bullet \phi \bullet \gamma$ because U is open and ϕ is one-to-one. This function can be split into two parts where one is $f \bullet \phi^{-1} : \mathbb{R}^n \to \mathbb{R}^1$ and the other is $\phi \bullet \gamma : \mathbb{R}^1 \to \mathbb{R}^n$ such that one can label $\phi \bullet \gamma$ as $(\phi \bullet \gamma)^\mu$. Using the chain rule gives

$$\frac{df}{d\lambda} = \frac{d}{d\lambda}\left((f \bullet \phi^{-1}) \bullet (\phi \bullet \gamma)\right) = \frac{\partial(f \bullet \phi^{-1})}{\partial(\phi \bullet \gamma)^\mu}\frac{d(\phi \bullet \gamma)^\mu}{d\lambda}$$

which can be written as $\dfrac{df}{d\lambda} = \dfrac{\partial f}{\partial X^\mu}\dfrac{dX^\mu}{d\lambda}$ with defining $X^\mu = (\phi \bullet \gamma)^\mu$ and $f(X^\mu) = (f \bullet \phi^{-1})$.

This leads to a so-called coordinate-adapted basis

$$\frac{d}{d\lambda} = \frac{dX^\mu}{d\lambda}\partial_\mu \tag{5.2}$$

because function f is arbitrary. In this formula $\frac{d}{d\lambda}$ is any directional derivative corresponding to a tangent vector, $\frac{dX^\mu}{d\lambda}$ are the components, and ∂_μ are the basis vectors.

Since any vector V must exist in the tangent space, it must be expressible as $V = V^\mu \partial_\mu$. Using (5.1) in

$$V^\mu \partial_\mu = V^{\mu'} \partial_{\mu'} = V^{\mu'} \frac{\partial X^\mu}{\partial X^{\mu'}} \partial_\mu = V^\nu \frac{\partial X^{\mu'}}{\partial X^\nu} \frac{\partial X^\mu}{\partial X^{\mu'}} \partial_\mu$$

since vectors are invariant, one gets

$$V^\mu \to V^{\mu'} = \frac{\partial X^{\mu'}}{\partial X^\mu} V^\mu \qquad\qquad \partial_\mu \to \partial_{\mu'} = \frac{\partial X^\mu}{\partial X^{\mu'}} \partial_\mu \tag{5.3}$$

as the transformation laws for vector components and basis vectors.

The dual vector $\omega_\mu \, \hat{\theta}^{(\mu)}$ consumes vectors such that $\hat{\theta}^{(\mu)} \, \hat{e}_{(\nu)} = \delta^\mu_\nu$. Similar to $\hat{e}_{(\mu)} = \partial_\mu$ one introduces $dX^\mu = \hat{\theta}^{(\mu)}$. Then the transformation laws are

$$\omega_\mu \to \omega_{\mu'} = \frac{\partial X^\mu}{\partial X^{\mu'}} \omega_\mu \qquad\qquad dX^\mu \to dX^{\mu'} = \frac{\partial X^{\mu'}}{\partial X^\mu} dX^\mu \tag{5.4}$$

for dual vector components and dual basis vectors because $dX^\mu \partial_\mu = dX^{\mu'} \partial_{\mu'}$.

For general tensors the transformation law for tensor components depend similarly to Special Relativity on the upper and lower indices. They can be deduced from (5.3) for vector components corresponding to upper indices and from (5.4) for dual vector components corresponding to lower indices. The example

$$T^{\kappa\lambda}{}_{\mu\nu} \to T^{\kappa'\lambda'}{}_{\mu'\nu'} = \frac{\partial X^{\kappa'}}{\partial X^\kappa} \frac{\partial X^{\lambda'}}{\partial X^\lambda} \frac{\partial X^\mu}{\partial X^{\mu'}} \frac{\partial X^\nu}{\partial X^{\nu'}} T^{\kappa\lambda}{}_{\mu\nu}$$

illustrates this for a (2,2) tensor.

Comparing the transformation of a vector in Special Relativity on the left side and in General Relativity on the right side

$$V^\mu \to V^{\mu'} = \Lambda^{\mu'}{}_\mu V^\mu \qquad\qquad V^\mu \to V^{\mu'} = \frac{\partial X^{\mu'}}{\partial X^\mu} V^\mu$$

shows that $\Lambda^{\mu'}{}_\mu$ is a constant or global transformation on (t, x, y, z) while $\frac{\partial X^{\mu'}}{\partial X^\mu}$ allows local or coordinate dependent transformations of (t, x, y, z). (Working in inertial frames with rectangular coordinates in flat spacetime makes these two transformations the same. But with the approach in General Relativity one can work in any frame with any kind of coordinates also in curved spacetime.)

The fact that the transformation in Special Relativity is global and the transformation in General Relativity is local causes a problem similar to the problem in gauge theories when going from a global symmetry to a local symmetry. Unlike the constant transformation in Special Relativity the variable transformation in General Relativity can in general not be moved past derivatives. Considering the derivative of a dual vector as a simple example for a tensor

$$\partial_\mu T_\nu \to \partial_{\mu'} T_{\nu'} = \frac{\partial X^\mu}{\partial X^{\mu'}} \partial_\mu \left(\frac{\partial X^\nu}{\partial X^{\nu'}} T_\nu \right) = \frac{\partial X^\mu}{\partial X^{\mu'}} \frac{\partial X^\nu}{\partial X^{\nu'}} \partial_\mu T_\nu + T_\nu \frac{\partial X^\mu}{\partial X^{\mu'}} \partial_\nu \left(\frac{\partial X^\nu}{\partial X^{\nu'}} \right)$$

the first term is all one should get if $\partial_\mu T_\nu$ is supposed to transformed as a tensor, but the second term is only zero if $\frac{\partial X^\nu}{\partial X^{\nu'}}$ is constant. Thus, the derivative of a tensor is in general not itself a tensor. However for physics one needs tensor equations and derivatives. The solution is the same as in the Standard Model of Particle Physics where localizing the symmetry of a gauge theory requires the redefinition of the derivative (and a gauge field) to the so-called covariant derivative which produces tensors from tensors.

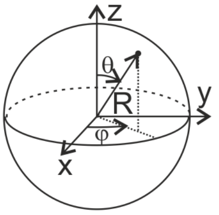
5.3 Local Inertial Coordinates

A single space can admit many different metrics coming from different coordinate choices. For example, $\mathbb{R}^3$ can have the have the two metrics on the left side and in the middle

$$g_{ij} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad g_{ij} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2\,\sin(\theta)^2 \end{pmatrix} \qquad g_{ij} = \begin{pmatrix} R^2 & 0 \\ 0 & R^2\,\sin(\theta)^2 \end{pmatrix}$$

as shown in Appendix A. The left metric which satisfies $g_{ij} = g^{ij}$ corresponds to Cartesian coordinates and the metric in the middle with $g_{ij} \neq g^{ij}$ corresponds to spherical polar coordinates. The look very different. On the other hand the metric in the middle and the metric on the right side look very similar aside from the fact that one corresponds to a three-dimensional space and the other to a two-dimensional space, but the metric in the middle belongs to the flat space $\mathbb{R}^3$ and the right metric to the sphere S^2 which is curved. Just looking at the metric does therefore not make it easy to see whether a space is flat or curved, and in General Relativity the difference whether spacetime is curved and flat is caused by gravity or the absence of it.

Another problem is that Einstein's equivalence principle states that experiments performed in a small, freely-falling lab over a short time interval give results that are indistinguishable to those in an inertial frame in empty space. This means that by choosing the right coordinates the metric can be brought to the form $g_{\mu\nu} = \eta_{\mu\nu}$ of Special Relativity with $\partial_\lambda g_{\mu\nu} \approx 0$ near the point in question and in a small enough environment around this point such that spacetime looks flat in this region. The coordinates that do this are called *local inertial coordinates*.



As an example for local inertial coordinates the sphere S^2 with coordinates (θ, φ) and radius R is used. It can be represented as

$$x = R\,\sin(\theta)\,\cos(\varphi) \qquad y = R\,\sin(\theta)\,\sin(\varphi) \qquad z = R\,\cos(\theta)$$

in Cartesian coordinates, and

$$g_{ij} = \begin{pmatrix} R^2 & 0 \\ 0 & R^2\,\sin(\theta)^2 \end{pmatrix} \qquad ds^2 = R^2\,d\theta^2 + R^2\,\sin(\theta)^2\,d\varphi^2$$

are the corresponding metric and line element.

At the north pole with $\theta \approx 0$ the metric is degenerate because $g_{11} = R^2$ and the other three components of the metric are zero. Because $\sin(\theta) \approx \theta$ and $z \approx R$ ($\approx$ constant) near the north pole, one can choose better suited coordinates

$$x = R\,\theta\,\cos(\varphi) \qquad y = R\,\theta\,\sin(\varphi)$$

for a small environment around the north pole. Inverting gives

$$\theta = \frac{\sqrt{x^2+y^2}}{R} \qquad \varphi = \tan^{-1}\left(\frac{y}{x}\right)$$
$$d\theta = \frac{1}{R\sqrt{x^2+y^2}}(x\,dx + y\,dy) \qquad d\varphi = \frac{1}{x^2+y^2}(x\,dy - y\,dx)$$

such that

$$\begin{aligned} ds^2_{xy} = &\left(\frac{x^2}{x^2+y^2} + R^2\,\sin\left(\frac{\sqrt{x^2+y^2}}{R}\right)\frac{y^2}{(x^2+y^2)^2}\right)dx^2 \\ &+ \left(\frac{y^2}{x^2+y^2} + R^2\,\sin\left(\frac{\sqrt{x^2+y^2}}{R}\right)\frac{x^2}{(x^2+y^2)^2}\right)dy^2 \\ &+ 2\left(\frac{xy}{x^2+y^2} - R^2\,\sin\left(\frac{\sqrt{x^2+y^2}}{R}\right)\frac{xy}{(x^2+y^2)^2}\right)dx\,dy \end{aligned}$$

and with

$$\sin\left(\frac{\sqrt{x^2+y^2}}{R}\right) = \frac{\sqrt{x^2+y^2}}{R} - \frac{1}{6}\left(\frac{\sqrt{x^2+y^2}}{R}\right)^3 + ...$$

one gets

$$ds^2_{xy} = \left(1 - \frac{2y^2}{3R^2} + ...\right) dx^2 + \left(1 - \frac{2x^2}{3R^2} + ...\right) dy^2 + \left(\frac{4xy}{3R^2} + ...\right) dx\, dy$$

for the line element expressed in $\{x, y\}$ coordinates. The metric is therefore

$$g_{ij} \approx \begin{pmatrix} 1 - \dfrac{2y^2}{3R^2} & \dfrac{2xy}{3R^2} \\ \dfrac{2xy}{3R^2} & 1 - \dfrac{2x^2}{3R^2} \end{pmatrix}$$

with $g_{ij} = I$ for $x = y = 0$ and $\partial_k g_{ij}|_{x=y=0} = 0$. The second derivatives $\frac{\partial^2}{\partial x^2}$, $\frac{\partial^2}{\partial y^2}$ and $\frac{\partial^2}{\partial x \partial y}$, however, will in general not vanish even for $x = y = 0$. These will turn out to be the quantities from which one can build a good measure of curvature.

In Special Relativity when determining a tensor equation in the rest frame, the equation is true in any frame. Similarly in General Relativity when a problem is solved in local inertial coordinates and expressed in terms of tensors, the solution is true in any coordinates. But one has to be careful because $\partial_\mu T_\nu$ is not a tensor.

5.4 Covariant Derivative

As shown above and repeated here the derivative of a tensor may not be a tensor, but derivatives are important in physics because physics without the possibility of change is boring. The translation laws are

$$T_\mu{}^\nu \to T_{\mu'}{}^{\nu'} = \frac{\partial X^\mu}{\partial X^{\mu'}} \frac{\partial X^{\nu'}}{\partial X^\nu} T_\mu{}^\nu \qquad\qquad \partial_\mu \to \partial_{\mu'} = \frac{\partial X^\mu}{\partial X^{\mu'}} \partial_\mu$$

for a (1,1) tensor as an example on the left and for the derivative on the right. The derivative is

$$\partial_\mu T^\nu \to \partial_{\mu'} T^{\nu'} = \frac{\partial X^\mu}{\partial X^{\mu'}} \partial_\mu \left(\frac{\partial X^{\nu'}}{\partial X^\nu} T^\nu\right) = \frac{\partial X^\mu}{\partial X^{\mu'}} \frac{\partial X^{\nu'}}{\partial X^\nu} \partial_\mu T^\nu + \frac{\partial X^\mu}{\partial X^{\mu'}} T^\nu \partial_\mu \left(\frac{\partial X^{\nu'}}{\partial X^\nu}\right)$$

for a (1,0) tensor where the first term in the final sum is tensorial but the second term is not. Thus the result is not a tensor if the second term does not vanish because the transformation is constant. The second term is the derivative of the transformation. The derivative of a scalar C

$$\partial_\mu C \to \partial_{\mu'} C = \frac{\partial X^\mu}{\partial X^{\mu'}} \partial_\mu C$$

is a tensor because C does not change. For all other tensors a new derivative is needed[7].

The new derivative called *covariant derivative* has the form $\nabla_\mu = \partial_\mu + \Gamma_{\mu\bullet}^{\bullet}$ where the Γ is the new part which is called a *connection*. This new derivative ∇_μ should be linear, have some kind of Leibniz property, make $\nabla_\mu V^\nu$ a tensor, commute with contractions, reduce to ∂_μ when acting on scalars, and be compatible with the metric. Finally, the connection should be torsionfree.

In a first step some of the properties of ∇ when applied to two vectors S and T are explored. Linearity means $\nabla(S + T) = \nabla S + \nabla T$ and the Leibniz property means $\nabla(S \otimes T) = (\nabla S) \otimes T + S \otimes (\nabla T)$. Both

[7] "In coordinate systems other than Cartesian ones the coordinate axes themselves move around when one moves to a different point in the space, and this must be taken into account for a proper consideration of differentiation." (Citation from *Introduction to General Relativity* by Lewis Ryder, Cambridge University Press, 2009.)

conditions are satisfied by $\nabla_\mu V^\nu = \partial_\mu V^\nu + \Gamma^\nu_{\mu\lambda} V^\lambda$. This is the partial derivative of the vector plus a linear transformation of the vector. Taking, for example, the derivative with respect to time gives

$$\nabla_0 V^\nu = \partial_0 V^\nu + \Gamma^\nu_{0\lambda} V^\lambda$$

where one can think of $\Gamma^\nu_{0\lambda} V^\lambda$ as a matrix multiplying a vector, and this is what is meant by a linear transformation. Linearity is obviously satisfied, but also the Leibniz condition holds as

$$\begin{aligned}\nabla_\mu(S^\nu T^\lambda) &= \partial_\mu(S^\nu T^\lambda) + \Gamma^\nu_{\mu\alpha} S^\alpha T^\lambda + \Gamma^\lambda_{\mu\beta} S^\nu T^\beta \\ &= (\partial_\mu S^\nu) T^\lambda + S^\nu(\partial_\mu T^\lambda) + (\Gamma^\nu_{\mu\alpha} S^\alpha) T^\lambda + S^\nu(\Gamma^\lambda_{\mu\beta} T^\beta) = (\nabla S)_\mu{}^\nu T^\lambda + S^\nu (\nabla T)_\mu^\lambda\end{aligned}$$

shows. The required tensorial transformation is

$$\nabla_\mu V^\nu \to \nabla_{\mu'} V^{\nu'} = \frac{\partial X^\mu}{\partial X^{\mu'}} \frac{\partial X^{\nu'}}{\partial X^\nu} \nabla_\mu V^\nu$$

where ∇_μ contains $\Gamma^\nu_{\mu\lambda}$ and $\nabla_{\mu'}$ contains $\Gamma^{\nu'}_{\mu'\lambda'}$. Because the transformation laws for ∂_μ and V^ν are known, one can deduce

$$\Gamma^{\nu'}_{\mu'\lambda'} = \frac{\partial X^\mu}{\partial X^{\mu'}} \frac{\partial X^{\nu'}}{\partial X^\nu} \frac{\partial X^\lambda}{\partial X^{\lambda'}} \Gamma^\nu_{\mu\lambda} - \frac{\partial X^\mu}{\partial X^{\mu'}} \frac{\partial X^\lambda}{\partial X^{\lambda'}} \frac{\partial^2 X^{\nu'}}{\partial X^\mu \partial X^\lambda} \tag{5.5}$$

as the connection transformation law. The first term is tensorial and the second term is not as one could have expected because there must be such a term to compensate for the non-tensorial part in ∂_μ.

In a second step this is extended to a connection on tensors. To move from vectors to tensors one can consider $\nabla_\mu(T^\lambda{}_\lambda) = (\nabla T)_\mu{}^\lambda{}_\lambda$ where first contraction and then derivative should give the same result as derivative first and contraction afterwards. This can be written as

$$\nabla_\mu(T^\lambda{}_\lambda) = \nabla_\mu(T^\nu{}_\lambda\, \delta^\lambda_\nu) = (\nabla_\mu T^\nu{}_\lambda)\delta^\lambda_\nu + T^\nu{}_\lambda(\nabla_\mu \delta^\lambda_\nu) = (\nabla T)_\mu{}^\nu{}_\lambda \delta^\lambda_\nu + T^\nu{}_\lambda(\nabla_\mu \delta^\lambda_\nu) = (\nabla T)_\mu{}^\lambda{}_\lambda + T^\nu{}_\lambda(\nabla_\mu \delta^\lambda_\nu)$$

which implies $\nabla_\mu \delta^\lambda_\nu = 0$ if the covariant derivative commutes with contractions. Since $\nabla_\mu C = \partial_\mu C$ for a scalar C, one can get the covariant derivative of dual vectors with

$$\begin{aligned}\nabla_\mu(\omega_\lambda V^\lambda) &= (\nabla_\mu \omega_\lambda) V^\lambda + \omega_\lambda(\nabla_\mu V^\lambda) = (\partial_\mu \omega_\lambda + \tilde{\Gamma}^\sigma_{\mu\lambda} \omega_\sigma) V^\lambda + \omega_\lambda(\partial_\mu V^\lambda + \Gamma^\lambda_{\mu\nu} V^\nu) \\ &= (\partial_\mu \omega_\lambda) V^\lambda + \omega_\lambda(\partial_\mu V^\lambda) + \tilde{\Gamma}^\sigma_{\mu\lambda} \omega_\sigma V^\lambda + \Gamma^\lambda_{\mu\nu} \omega_\lambda V^\nu) = \partial_\mu(\omega_\lambda V^\lambda) + \tilde{\Gamma}^\sigma_{\mu\lambda} \omega_\sigma V^\lambda + \Gamma^\lambda_{\mu\nu} \omega_\lambda V^\nu\end{aligned}$$

which only gives $\nabla_\mu(\omega_\lambda V^\lambda) = \partial_\mu(\omega_\lambda V^\lambda)$ if $\tilde{\Gamma}^\sigma_{\mu\lambda} \omega_\sigma V^\lambda + \Gamma^\lambda_{\mu\nu} \omega_\lambda V^\nu = 0$ or, after replacing λ with σ and ν with λ, if $\tilde{\Gamma}^\sigma_{\mu\lambda} \omega_\sigma V^\lambda = -\Gamma^\sigma_{\mu\lambda} \omega_\sigma V^\lambda$. The covariant derivative is therefore

$$\nabla_\mu V^\nu = \partial_\mu V^\nu + \Gamma^\nu_{\mu\lambda} V^\lambda \qquad \nabla_\mu \omega_\nu = \partial_\mu \omega_\nu - \Gamma^\lambda_{\mu\nu} \omega_\lambda \qquad \nabla_\mu T^\alpha{}_\beta = \partial_\mu T^\alpha{}_\beta + \Gamma^\alpha_{\mu\lambda} T^\lambda{}_\beta - \Gamma^\kappa_{\mu\beta} T^\alpha{}_\kappa \tag{5.6}$$

for vectors, dual vectors and (1,1) tensors. This shows that this is a derivative which has the important property that it results in a tensor when applied to a tensor.

In a third step it is made sure that the connection is torsionfree and compatible with the metric because there are still several possibilities for the $\Gamma^\lambda_{\mu\nu}$ remaining. Suppose there are two different connections $\Gamma^\lambda_{\mu\nu}$ and $\tilde{\Gamma}^\lambda_{\mu\nu}$ for two different covariant derivatives ∇ and $\tilde{\nabla}$ then

$$\nabla_\mu V^\lambda - \tilde{\nabla}_\mu V^\lambda = \partial_\mu V^\lambda + \Gamma^\lambda_{\mu\nu} V^\nu - \partial_\mu V^\lambda - \tilde{\Gamma}^\lambda_{\mu\nu} V^\nu = (\Gamma^\lambda_{\mu\nu} - \tilde{\Gamma}^\lambda_{\mu\nu}) V^\nu$$

which shows that $\Gamma^\lambda_{\mu\nu} - \tilde{\Gamma}^\lambda_{\mu\nu}$ must be a tensor because V^ν and $\nabla_\mu V^\lambda - \tilde{\nabla}_\mu V^\lambda$ are tensors. The difference of two connections is therefore a tensor. So one can start with a connection $\Gamma^\lambda_{\mu\nu}$ and form the *torsion tensor* $T^\lambda_{\mu\nu}$ defined as

$$T^\lambda_{\mu\nu} = \Gamma^\lambda_{\mu\nu} - \Gamma^\lambda_{\nu\mu} = 2\Gamma^\lambda_{[\mu\nu]}$$

which is a tensor (as its name suggest). In general given a connection $\Gamma^\lambda_{\mu\nu}$ one can break it up in a antisymmetric and a symmetric piece

$$\Gamma^\lambda_{\mu\nu} = \frac{1}{2}\left(\Gamma^\lambda_{\mu\nu} - \Gamma^\lambda_{\nu\mu}\right) + \frac{1}{2}\left(\Gamma^\lambda_{\mu\nu} + \Gamma^\lambda_{\nu\mu}\right) = \frac{1}{2}\Gamma^\lambda_{[\mu\nu]} + \frac{1}{2}\Gamma^\lambda_{(\mu\nu)}$$

to get a torsionfree version $\Gamma^{\lambda}_{(\mu\nu)}$ from any connection. (The antisymmetric piece is tensorial, but the symmetric piece is not. Thus one can eliminate the antisymmetric piece and only use the symmetric non-tensorial piece.) Finally using metric compatibility which means $\nabla_\mu\, g_{\kappa\lambda} = 0$ to make the metric not constant but covariantly constant together with the torsionfree property leads to the so-called Christoffel connection[8] with the Christoffel symbols $\Gamma^{\lambda}_{\mu\nu}$ which will be used here. With $\Gamma^{\lambda}_{\mu\nu} = \Gamma^{\lambda}_{\nu\mu}$, $g_{\mu\nu} = g_{\nu\mu}$ and

$$\begin{aligned}
&(A)\,\kappa\mu\nu : \nabla_\kappa\, g_{\mu\nu} = \partial_\kappa\, g_{\mu\nu} - \Gamma^{\lambda}_{\kappa\mu}\, g_{\lambda\nu} - \Gamma^{\lambda}_{\kappa\nu}\, g_{\mu\lambda} = 0\\
&(B)\,\mu\nu\kappa : \nabla_\mu\, g_{\nu\kappa} = \partial_\mu\, g_{\nu\kappa} - \Gamma^{\lambda}_{\mu\nu}\, g_{\lambda\kappa} - \Gamma^{\lambda}_{\mu\kappa}\, g_{\nu\lambda} = \partial_\mu\, g_{\nu\kappa} - \Gamma^{\lambda}_{\mu\nu}\, g_{\lambda\kappa} - \Gamma^{\lambda}_{\kappa\mu}\, g_{\lambda\nu} = 0\\
&(C)\,\nu\kappa\mu : \nabla_\nu\, g_{\kappa\mu} = \partial_\nu\, g_{\kappa\mu} - \Gamma^{\lambda}_{\nu\kappa}\, g_{\lambda\mu} - \Gamma^{\lambda}_{\nu\mu}\, g_{\kappa\lambda} = \partial_\nu\, g_{\kappa\mu} - \Gamma^{\lambda}_{\kappa\nu}\, g_{\mu\lambda} - \Gamma^{\lambda}_{\mu\nu}\, g_{\lambda\kappa} = 0
\end{aligned}$$

and $(A) - (B) - (C) = \partial_\kappa\, g_{\mu\nu} - \partial_\mu\, g_{\nu\kappa} - \partial_\nu\, g_{\kappa\mu} + 2\,\Gamma^{\lambda}_{\mu\nu}\, g_{\lambda\kappa}$ one gets

$$\Gamma^{\lambda}_{\mu\nu} = \frac{1}{2} g^{\lambda\kappa}\left(\partial_\mu\, g_{\nu\kappa} + \partial_\nu\, g_{\kappa\mu} - \partial_\kappa\, g_{\mu\nu}\right) \tag{5.7}$$

which defines the Christoffel connection in terms of the metric. The object $\Gamma^{\lambda}_{\mu\nu}$ can be calculated given explicit coordinates X^μ, but it is obviously not a tensor and does therefore not transform the way tensors do. The transformation law for them is given by (5.5).

5.5 Interpretation of the Covariant Derivative

As an rather easy example $\mathbb{R}^2$ with coordinates (r, θ) and

$$ds^2 = dr^2 + r^2\, d\theta^2 \qquad g_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix} \qquad g^{ij} = \begin{pmatrix} 1 & 0 \\ 0 & r^{-2} \end{pmatrix}$$

is used. The Christoffel symbols are $\Gamma^{r}_{\theta\theta} = -r$, $\Gamma^{\theta}_{r\theta} = \Gamma^{\theta}_{\theta r} = \frac{1}{r}$ and all others are zero. Calculating $\vec{\nabla} \cdot \vec{v}$ using the apparatus of the Christoffel connection gives

$$\begin{aligned}
\nabla_\mu V^\mu = \delta^\mu_\nu\, \nabla_\mu\, V^\nu &= \delta^\mu_\nu(\partial_\mu\, V^\nu + \Gamma^{\nu}_{\mu\lambda}\, V^\lambda) = \partial_\mu\, V^\mu + \Gamma^{\mu}_{\mu\lambda}\, V^\lambda\\
&= \partial_r\, v^r + \partial_\theta\, v^\theta + \Gamma^{r}_{rr}\, v^r + \Gamma^{r}_{r\theta}\, v^\theta + \Gamma^{\theta}_{\theta r}\, v^r + \Gamma^{\theta}_{\theta\theta}\, v^\theta = \partial_r\, v^r + \partial_\theta\, v^\theta + \frac{1}{r} v^r
\end{aligned}$$

compared with the expression $\vec{\nabla} \cdot \vec{v} = \partial_r\, v^r + \frac{1}{r}\partial_\theta\, v^\theta + \frac{1}{r} v^r$ from electromagnetism textbooks shows a disagreement in the middle term. The difference comes from the fact that the expression from the textbooks use an orthonormal basis with $\hat{e}_{(r)} \cdot \hat{e}_{(r)} = 1$ and $\hat{e}_{(\theta)} \cdot \hat{e}_{(\theta)} = 1$ leading to the metric I while the basis used in this example satisfies $\hat{e}_{(r)} \cdot \hat{e}_{(r)} = 1$ but $\hat{e}_{(\theta)} \cdot \hat{e}_{(\theta)} = r^2$ leading to the metric shown above.

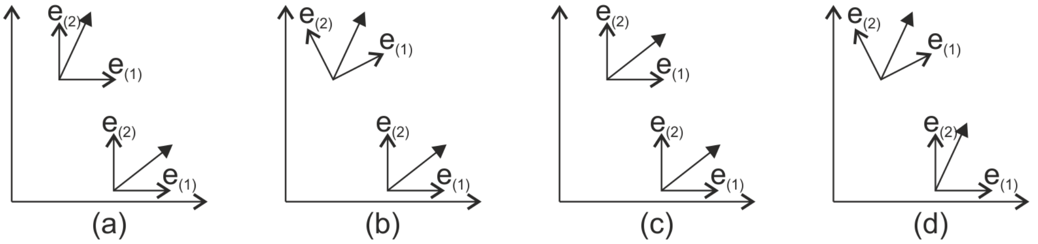


Figure 5.1: The movement of a vector in space

The covariant derivative $\nabla_\mu\, V^\nu = \partial_\mu\, V^\nu + \Gamma^{\nu}_{\mu\lambda}\, V^\lambda$ describes how the vector V^ν changes as it moves around in space. This movement can happen in two ways as shown in figure 5.1. The quantity $\partial_\mu\, V^\nu$ tells whether the vector V^ν changes relative to the basis while the quantity $\Gamma^{\nu}_{\mu\lambda}\, V^\lambda$ tells how the basis vectors change as the vector V^ν moves in space. If one compares the vector in (a) moving from the right to the left position and in (b) also moving from the right to the left position and ignores the basis vectors, then the vector has changed exactly the same way. In (a) however $\partial_\mu\, V^\nu \neq 0$ and $\Gamma^{\nu}_{\mu\lambda}\, V^\lambda = 0$ because

[8] Other theories of gravity use other connections. If gravity and spin are combined, for example, the condition that the connection is free of torsion has to be relaxed.

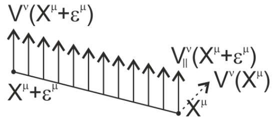
the basis has not changed, and in (b) $\partial_\mu V^\nu = 0$ and $\Gamma^\nu_{\mu\lambda} V^\lambda \neq 0$ because the components of the vector have not changed with respect to the basis. In both cases $\nabla_\mu V^\nu \neq 0$ because the vector has changed. If $\nabla_\mu V^\nu = 0$ then the vector V^ν is *covariantly constant*, and this can also happen in two ways. In (c) neither the components nor the basis change such that $\partial_\mu V^\nu = 0$ and $\Gamma^\nu_{\mu\lambda} V^\lambda = 0$ with the obvious consequence $\nabla_\mu V^\nu = 0$ while in (d) $\partial_\mu V^\nu \neq 0$ and $\Gamma^\nu_{\mu\lambda} V^\lambda \neq 0$ but the changes cancel each other such that the result is $\nabla_\mu V^\nu = 0$.

5.6 Parallel Transport

The covariant derivative can also tell how to move vectors in space from one tangent space to another tangent space. When considering

$$\partial_\mu V^\nu|_{X^\mu} = \lim_{\varepsilon^\mu \to 0} \frac{V^\nu(X^\mu + \varepsilon^\mu) - V^\mu(X^\mu)}{\varepsilon^\mu} \qquad \nabla_\mu V^\nu|_{X^\mu} = \lim_{\varepsilon^\mu \to 0} \frac{V^\nu_{||}(X^\mu + \varepsilon^\mu) - V^\mu(X^\mu)}{\varepsilon^\mu}$$

the left definition does not make sense because the two vectors $V^\nu(X^\mu + \varepsilon^\mu)$ and $V^\mu(X^\mu)$ exist in different tangent spaces. The vector $V^\nu_{||}(X^\mu + \varepsilon^\mu)$ is not the vector $V^\nu(X^\mu + \varepsilon^\mu)$ itself in its tangent space but is moved back to the tangent space of $V^\nu(X^\mu)$ using parallel transport before subtracting.



Parallel transport keeps the vector parallel to itself during the move in a very sensible way. Since $V^\nu_{||}(X^\mu + \varepsilon^\mu)$ and $V^\nu(X^\mu)$ are in the same tangent space they can be subtracted. This sensible way to move vectors from one tangent space to another is achieved by the covariant derivative.

A formal definition of *parallel transport* using the covariant derivative ∇_μ moves the vector V^ν along a curve $X^\mu(\lambda)$ using

$$\frac{d}{d\lambda} = \frac{dX^\mu}{d\lambda}\partial_\mu \qquad \frac{D}{d\lambda} = \frac{dX^\mu}{d\lambda}\nabla_\mu = \frac{d}{d\lambda} + \Gamma^\bullet_{\mu\bullet}\frac{dX^\mu}{d\lambda} \qquad \frac{DV^\nu}{d\lambda} = \frac{dX^\mu}{d\lambda}\nabla_\mu V^\nu = 0$$

where the left definition in flat space with Cartesian coordinates is replaced by the middle definition using the covariant derivative such that the right equation is satisfied. The right equation written as

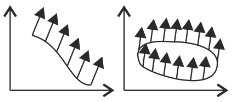
$$\frac{dV^\nu}{d\lambda} + \Gamma^\nu_{\mu\kappa}\frac{dX^\mu}{d\lambda} V^\kappa = 0 \tag{5.8}$$

defines therefore the parallel transport of the vector V^ν along the curve $X^\mu(\lambda)$. To use equation (5.8) one can specify the vector one wants to start with and then solve this differential equation. This will tell the components of the vector as it moves along the curve. That is the unique parallel transport of that vector along that path.

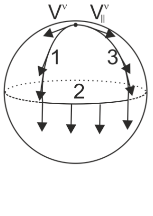
One can extend this definition to arbitrary tensors

$$\frac{D}{d\lambda}T^\alpha_{\ \beta} = \frac{dX^\mu}{d\lambda}\nabla_\mu T^\alpha_{\ \beta} = \frac{dX^\mu}{d\lambda}\left(\partial_\mu T^\alpha_{\ \beta} + \Gamma^\alpha_{\mu\rho}T^\rho_{\ \beta} - \Gamma^\sigma_{\mu\beta}T^\alpha_{\ \sigma}\right) = 0$$

such that given a path $X^\mu(\lambda)$ and the value of $T^\alpha_{\ \beta}$ at $X^\mu(\lambda_0)$ one can solve this differential equation for $T^\alpha_{||\beta}$ at any other point in the path.



Parallel transport is important for three reasons. Firstly it is part of constructing a consistent derivative ∇_μ which includes also the change of the basis vectors. Secondly it helps detect curvature. Thirdly it helps identify geodesic paths which encode the response of particles to the curvature of space.



Moving a vector along a path in flat space $\mathbb{R}^2$ does not change its direction. Thus after moving it around on a closed path it still points in the same direction and remains the same vector. This is different when moving the vector on a curved space such as S^2. Moving a vector from the north pole along a meridian (1) to the equator, along the equator (2) for a while and finally along another meridian (3) back to the north pole results in another vector pointing in a different direction. If the vector is initially tangent to the path, it will remain parallel to itself as best as it can but it has to exist in the tangent space at each point it is moved to.

5.7 Geodesics

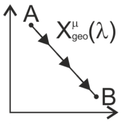
There are two different covariant derivatives

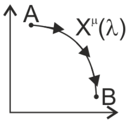
$$\nabla_\mu V^\nu = \lim_{\Delta X^\mu \to 0} \frac{V^\nu_{||}(X^\mu + \Delta X^\mu) - V^\mu(X^\mu)}{\Delta X^\mu}$$

$$\frac{DV^\nu}{d\lambda} = \lim_{\Delta\lambda \to 0} \frac{V^\nu_{||}(X^\mu(\lambda + \Delta\lambda)) - V^\mu(X^\mu(\lambda))}{\Delta X^\mu} = \frac{dX^\mu}{d\lambda} \nabla_\mu$$

which both use parallel transport but are based on different paths. In the derivative $\nabla_\mu V^\nu$ a value for the coordinate μ is selected and the path is a shift along X^μ. In $\frac{DV^\nu}{d\lambda}$ which is called *directional covariant derivative* the shift is along the curve $X^\mu(\lambda)$ which can be arbitrary and may not be aligned along a coordinate axis. The expression $(dX^\mu/d\lambda)\,\nabla_\mu$ states that to find out how V^ν varies with λ one can first figure out how V^ν varies with the coordinates and then figure out how the coordinates vary with λ.

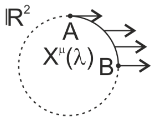
The first half of electromagnetism is given by Maxwell's equations which specify the sources and the second half is given by the Lorentz force and Newton's second law which specify what a charged particle does when put into an electromagnetic configuration. In General Relativity there is an analogous split into Einstein's equations telling how sources create curvature and the geodesic equation showing how particles move through spacetime.

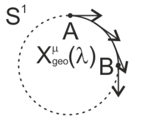




There are two ways to define a geodesic path $X^\mu_{\text{geo}}(\lambda)$ both with strengths and weaknesses. Curves $X^\mu(\lambda)$ which extremize the distance between two points are geodesics, and curves $X^\mu(\lambda)$ which parallel-transport their own tangent vectors are geodesics.

In $\mathbb{R}^2$ the shortest path between two points A and B is the straight line, and it is obvious that the tangent vectors which are parallel to the straight line are parallel-transported on this line. A path on the other hand which is not a straight line is therefore nongeodesic, and the different tangent vectors are not parallel to each other.





Comparing a circle in the two different spaces $\mathbb{R}^2$ and S^1 shows that parallel transport can mean two different things on two paths looking the same. In $\mathbb{R}^2$ what starts at A as a tangent vector is at B no longer a tangent vector. The circle is therefore in $\mathbb{R}^2$ not the shortest path from A to B. In S^1 the tangent vector at A parallel-transported to B is still a tangent vector. This path from A to B is an extremal path in S^1 and therefore a geodesic. Tangent vectors are constrained to exist in the tangent space. In S^1 the tangent space at one point is a line touching the circle in this point.

To formalize the definition of geodesics by parallel transport one should recall that $\frac{dX^\mu}{d\lambda}$ are the components of the tangent vectors for the curve $X^\mu(\lambda)$. For a geodesic $X^\mu_{\text{geo}}(\lambda)$ these components should be covariantly constant along the curve or, in other words, should be parallel to each other. The path $X^\mu(\lambda)$ is therefore a geodesic if

$$\frac{D}{d\lambda}\frac{dX^\mu}{d\lambda} = \frac{dX^\nu}{d\lambda}\nabla_\nu \frac{dX^\mu}{d\lambda} = \frac{dX^\nu}{d\lambda}\left(\partial_\nu \frac{dX^\mu}{d\lambda} + \Gamma^\mu_{\nu\kappa}\frac{dX^\kappa}{d\lambda}\right) = 0$$

is satisfied which can be written as

$$\frac{d^2X^\mu}{d\lambda^2} + \Gamma^\mu_{\nu\kappa}\frac{dX^\nu}{d\lambda}\frac{dX^\kappa}{d\lambda} = 0 \tag{5.9}$$

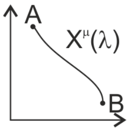
and is called the *geodesic equation.*

This is a second order differential equation which therefore needs two boundary conditions before it can be solved. This lead to two possibilities:

1. One can generate the geodesic “launched” from initial position $X^\mu(\lambda_0)$ and “velocity” $\frac{dX^\mu}{d\lambda}|_{\lambda_0}$.
2. One can give an initial and final position and find the extremal path between them.

Since $\Gamma^\mu_{\nu\kappa}$ depends on $g_{\mu\nu}$, the explicit form will vary for different geometries.

As an intuitive example, $\Gamma = 0$ in $\mathbb{R}^3$ with Cartesian coordinates such that $\frac{d^2X^\mu}{d\lambda^2} = 0$ has the solution $X^\mu(\lambda) = \lambda\varepsilon^\mu + X^\mu_0$ where ε^μ and X^μ_0 are constants from boundary conditions. This is a straight line.



Another example is $\mathbb{R}^2$ with polar coordinates (r, θ) which shows the extremization more explicitly. The path between A and B is $X^\mu(\lambda) = (r(\lambda), \theta(\lambda))$ and

$$ds^2 = dr^2 + r^2\,d\theta^2 \qquad \Gamma^r_{\theta\theta} = -r \qquad \Gamma^\theta_{\theta r} = \Gamma^\theta_{r\theta} = \frac{1}{r}$$

are the line element and the relevant Christoffel symbols for polar coordinates in $\mathbb{R}^2$. The parameter λ is chosen to be the path length s such that $X^\mu(s) = (r(s), \theta(s))$. The total path length from A to B is

$$s_{AB} = \int_A^B ds = \int_A^B \sqrt{dr^2 + r^2 d\theta^2} = \int_A^B \sqrt{\frac{dr^2}{ds^2} + r^2\frac{d\theta^2}{ds^2}}\,ds = \int_A^B \sqrt{v_r^2 + r^2 v_\theta^2}\,ds$$

with $v_r = \frac{dr}{ds}$ and $v_\theta = \frac{d\theta}{ds}$ because extremizing this is akin to extremizing an action $S = \int L(x^i, v^i)dt$ of a Lagrangian in classical mechanics. The solution is therefore the Euler-Lagrange equation

$$\frac{d}{ds}\left(\frac{\partial L}{\partial v^i}\right) - \frac{\partial L}{\partial x^i} = 0 \qquad L = \sqrt{v_r^2 + r^2 v_\theta^2} \qquad \begin{aligned} \frac{d}{ds}\left(\frac{\partial L}{\partial v_r}\right) - \frac{\partial L}{\partial r} &= \frac{d^2 r}{ds^2} - r\left(\frac{d\theta}{ds}\right)^2 = 0 \\ \frac{d}{ds}\left(\frac{\partial L}{\partial v_\theta}\right) - \frac{\partial L}{\partial \theta} &= \frac{d^2\theta}{ds^2} + \frac{2}{r}\frac{dr}{ds}\frac{d\theta}{ds} = 0 \end{aligned}$$

where t has been replaced by s. If one compares this with the result of the geodesic equation (5.9)

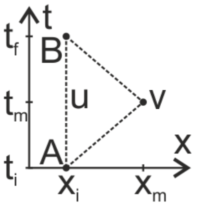
$$\frac{d^2 X^\mu}{ds^2} + \Gamma^\mu_{\nu\kappa}\frac{dX^\nu}{ds}\frac{dX^\kappa}{ds} = 0 \qquad \begin{aligned} \frac{d^2 r}{ds^2} + \Gamma^r_{\theta\theta}\left(\frac{d\theta}{ds}\right)^2 &= \frac{d^2 r}{ds^2} - r\left(\frac{d\theta}{ds}\right)^2 = 0 \\ \frac{d^2\theta}{ds^2} + \Gamma^\theta_{\theta r}\frac{d\theta}{ds}\frac{dr}{ds} + \Gamma^\theta_{r\theta}\frac{dr}{ds}\frac{d\theta}{ds} &= \frac{d^2\theta}{ds^2} + \frac{2}{r}\frac{dr}{ds}\frac{d\theta}{ds} = 0 \end{aligned}$$

where λ has been replaced by s, then both ways give the same result. (Note that $L = 0$ because of the above integrals from A to B which are $\int ds = \int L\,ds$.)

If one is working in a Lorentzian signature spacetime such as Minkowski spacetime then

$$\tau_{AB} = \int_A^B \sqrt{-ds^2}$$

and timelike geodesics which correspond to massive particles actually maximize the spacetime length. (A massive object can be at rest where $ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2$ becomes $ds^2 = -c^2dt^2$.)



To demonstrate this in an example one can consider a geodesic path u with constant velocity or at rest in $\mathbb{M}^2$ with $ds^2 = -dt^2 + dx^2$ (setting $c = 1$) and an accelerated non-geodesic path v. Both start at t_i in point A with x_i and end at t_f in point B also with x_i, but u did not move and v moved to a point with x_m. For u one gets

$$\tau_{AB} = \int_A^B dt = t_f - t_i$$

and the time from A to B is just $t_f - t_i$. For v one gets

$$\tau_{AB} = \int_A^B \sqrt{dt^2 - dx^2} = \int_A^B \sqrt{1 - v_x^2}\,dt = \int_{t_i}^{t_m} \sqrt{1 - v_x^2}\,dt + \int_{t_m}^{t_f} \sqrt{1 - (-v_x)^2}\,dt = \int_{t_i}^{t_f} \sqrt{1 - v_x^2}\,dt$$

and the time from A to B is $< t_f - t_i$. If v_x is constant, the value $\sqrt{1 - v_x^2}$ is smaller than one. This calculation is however also valid for velocities which are not constant. Thus the path directly from A to B is therefore longer than the indirect path from A to B, and there does not exist a shortest path because one can make it shorter and shorter by going further and further away. As a consequence in this space the only extremal path is the longest path.

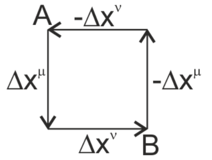
By the way, these two trajectories are representative of those in the twin paradox. The twin who remains on earth corresponds to u and is older than the one that travels with a spaceship away and then comes back corresponding to v.

5.8 Curvature

One would like to have a coordinate invariant way of specifying that a space is flat:

i) The metric even in flat space depending on the chosen coordinates does not easily show whether a space is flat or not.
ii) The Christoffel symbols in flat space with Cartesian coordinates vanish, but the condition that the Christoffel symbols must be zero in flat space does not work because they are not tensors.
iii) One can always choose locally inertial coordinates such that the metric becomes the identity and the Christoffel symbols vanish even in curved space.

Thus, these three criteria are not useful to distinguish flat and curved spaces.



As shown above, parallel-transporting a vector on S^2 from the north pole to the equator, further on the equator for a while and back to the north pole brings the vector back changed. (This is not true for all closed path because parallel-transporting a vector on the equator alone does not change it, but there is always a closed path where the vector changes.) Choosing an infinitesimal closed path from A to B and back to A as in the figure, one can take a vector and parallel-transport it from A to B along one path and back to A along the other path. Calling the vector V_A^λ this vector at A before it gets parallel-transported and $\tilde{V}_A^\lambda$ after it has been parallel-transported, then

$$\tilde{V}_A^\lambda = (-\Delta\hat{X}^\nu)\,(-\Delta\hat{X}^\mu)\,\Delta\hat{X}^\nu\,\Delta\hat{X}^\mu\,V_A^\lambda$$

where the $\Delta\hat{X}^\mu$ and so on are some operators performing the parallel-transport. If $\tilde{V}_A^\lambda - V_A^\lambda = \delta V_A^\lambda$ is zero, then the space is flat. (Instead of the infinitesimal square one can use another closed path such as a triangle on a sphere S^2, for example, which goes from the north pole to the equator, along the equator for a while and back to the north pole.) The question remains what $\Delta\hat{X}^\mu$ is.

To find out whether it is ∇_μ one can compute the commutator of two covariant derivatives as they act on vectors. Using the fact that $\nabla_\nu V^\lambda$ is a (1,1) tensor, the commutator is

$$\begin{aligned}
[\nabla_\mu,\nabla_\nu]V^\lambda &= \nabla_\mu\nabla_\nu V^\lambda - \nabla_\nu\nabla_\mu V^\lambda \\
&= \big(\partial_\mu(\nabla_\nu V^\lambda) - \Gamma^\kappa_{\mu\nu}\nabla_\kappa V^\lambda + \Gamma^\lambda_{\mu\kappa}\nabla_\nu V^\kappa\big) - \big(\partial_\nu(\nabla_\mu V^\lambda) - \Gamma^\kappa_{\nu\mu}\nabla_\kappa V^\lambda + \Gamma^\lambda_{\nu\kappa}\nabla_\mu V^\kappa\big) \\
&= \big(\partial_\mu(\partial_\nu V^\lambda + \Gamma^\lambda_{\nu\rho}V^\rho) - \Gamma^\kappa_{\mu\nu}\nabla_\kappa V^\lambda + \Gamma^\lambda_{\mu\kappa}\nabla_\nu V^\kappa\big) \\
&\quad - \big(\partial_\nu(\partial_\mu V^\lambda + \Gamma^\lambda_{\mu\rho}V^\rho) - \Gamma^\kappa_{\nu\mu}\nabla_\kappa V^\lambda + \Gamma^\lambda_{\nu\kappa}\nabla_\mu V^\kappa\big) \\
&= \big(\partial_\mu\partial_\nu V^\lambda + (\partial_\mu\Gamma^\lambda_{\nu\rho})V^\rho + \Gamma^\lambda_{\nu\rho}\partial_\mu V^\rho - \Gamma^\kappa_{\mu\nu}\partial_\kappa V^\lambda - \Gamma^\kappa_{\mu\nu}\Gamma^\lambda_{\kappa\rho}V^\rho + \Gamma^\lambda_{\mu\kappa}\partial_\nu V^\kappa + \Gamma^\lambda_{\mu\kappa}\Gamma^\kappa_{\nu\rho}V^\rho\big) \\
&\quad - \big(\partial_\nu\partial_\mu V^\lambda + (\partial_\nu\Gamma^\lambda_{\mu\rho})V^\rho + \Gamma^\lambda_{\mu\rho}\partial_\nu V^\rho - \Gamma^\kappa_{\nu\mu}\partial_\kappa V^\lambda - \Gamma^\kappa_{\nu\mu}\Gamma^\lambda_{\kappa\rho}V^\rho + \Gamma^\lambda_{\nu\kappa}\partial_\mu V^\kappa + \Gamma^\lambda_{\nu\kappa}\Gamma^\kappa_{\mu\rho}V^\rho\big) \\
&= \big((\partial_\mu\Gamma^\lambda_{\nu\rho})V^\rho + \Gamma^\lambda_{\mu\kappa}\Gamma^\kappa_{\nu\rho}V^\rho\big) - \big((\partial_\nu\Gamma^\lambda_{\mu\rho})V^\rho + \Gamma^\lambda_{\nu\kappa}\Gamma^\kappa_{\mu\rho}V^\rho\big) \\
&= \big(\partial_\mu\Gamma^\lambda_{\nu\rho} - \partial_\nu\Gamma^\lambda_{\mu\rho} + \Gamma^\lambda_{\mu\kappa}\Gamma^\kappa_{\nu\rho} - \Gamma^\lambda_{\nu\kappa}\Gamma^\kappa_{\mu\rho}\big)\,V^\rho
\end{aligned}$$

after removing canceling terms such $\partial_\mu\partial_\nu V^\lambda - \partial_\nu\partial_\mu V^\lambda$. The commutator $[\nabla_\mu,\nabla_\nu]$ parallel-transports a vector on two different paths from one point to another, and because $\Gamma \sim \partial g$ this is $[\nabla_\mu,\nabla_\nu] \sim \partial^2 g + (\partial g)^2$. It does not vanish even in a local inertial frame and is therefore a good measure for curvature.

Because $[\nabla_\mu,\nabla_\nu]$ is a tensor, also $\partial_\mu\Gamma^\lambda_{\nu\rho} - \partial_\nu\Gamma^\lambda_{\mu\rho} + \Gamma^\lambda_{\mu\kappa}\Gamma^\kappa_{\nu\rho} - \Gamma^\lambda_{\nu\kappa}\Gamma^\kappa_{\mu\rho}$ is a tensor. Therefore the tensor

$$[\nabla_\mu,\nabla_\nu] = R^\lambda{}_{\rho\mu\nu} = \partial_\mu\Gamma^\lambda_{\nu\rho} - \partial_\nu\Gamma^\lambda_{\mu\rho} + \Gamma^\lambda_{\mu\kappa}\Gamma^\kappa_{\nu\rho} - \Gamma^\lambda_{\nu\kappa}\Gamma^\kappa_{\mu\rho} \tag{5.10}$$

which is called the *Riemann curvature tensor* measures the curvature. It has $4^4 = 256$ independent components in four dimensions. With $R_{\kappa\rho\mu\nu} = g_{\kappa\lambda}R^\lambda{}_{\rho\mu\nu}$ one can show that

$R_{\kappa\rho\mu\nu} = -R_{\kappa\rho\nu\mu}$ (antisymmetric in the last two indices because of the definition of the commutator)
$R_{\kappa\rho\mu\nu} = -R_{\rho\kappa\mu\nu}$ (antisymmetric in the first two indices)
$R_{\kappa\rho\mu\nu} = R_{\mu\nu\kappa\rho}$ (symmetric under exchange of the first two and the last two indices)
$R_{\kappa\rho\mu\nu} + R_{\kappa\nu\rho\mu} + R_{\kappa\mu\nu\rho} = 0$ (cyclic reordering of the last three indices)

such that the number of independent components reduce to 20. Because $R^\lambda{}_{\rho\mu\nu}$ is a tensor, it vanishes in all coordinate systems if it vanishes in one of them.

A useful fact in the following is that $T^{(\mu\nu)}W_{[\mu\nu]} = 0$ as one can easily check in two dimensions with $T^{(ij)}W_{[ij]} = T^{11}W_{11} + T^{12}W_{12} + T^{21}W_{21} + T^{22}W_{22} = T^{11}0 + T^{12}W_{12} - T^{12}W_{12} + T^{22}0$.

Space is flat if $R^{\lambda}{}_{\rho\mu\nu} = 0$, and this property is coordinate invariant because it is a tensor, but there are other degrees of flatness which can be specified with quantities built from the Riemann curvature tensor. All contractions of the Riemann curvature tensor are:

i) $R^{\lambda}{}_{\lambda\mu\nu} = g^{\lambda\kappa}\, R_{\kappa\lambda\mu\nu} = g^{(\lambda\kappa)}\, R_{[\kappa\lambda]\mu\nu} = 0$
ii) $R^{\lambda}{}_{\rho\lambda\nu} = g^{\lambda\kappa}\, R_{\kappa\rho\lambda\nu} = R_{\rho\nu}$ called *Ricci tensor*
iii) $R^{\lambda}{}_{\rho\mu\lambda} = -R^{\lambda}{}_{\rho\lambda\mu} = -R_{\rho\nu}$

The Ricci tensor is symmetric such that $R_{\mu\nu} = R_{\nu\mu}$ because of $R_{[\kappa\lambda]\,[\mu\nu]}$. One can go a step further and contract the two indices of the Ricci tensor to get $R = R^{\nu}{}_{\nu} = g^{\nu\mu}R_{\mu\nu}$ called *Ricci scalar.*

One can ask what one gets if one takes the Riemann curvature tensor and subtracts all possible contractions of it. The remaining tensor is the *Weyl tensor* $C_{\rho\sigma\mu\nu}$. This tensor is not used in the following, but the idea is that the set $\{C_{\rho\sigma\mu\nu}, R_{\mu\nu}, R\}$ contains all information in the Riemann curvature tensor.

The degrees of flatness are:

$R^{\lambda}{}_{\rho\mu\nu} = 0$ means really flat (flat-flat) such as $\mathbb{R}^n$, $\mathbb{M}^n$, $\mathbb{T}^n$.
$R_{\mu\nu} = 0$ means Ricci-flat such as $\text{AdS}^5 \times \text{S}^5$.
$C_{\rho\sigma\mu\nu} = 0$ means conformally-flat such as all two-dimensional pseudo-Riemannian manifolds.
$R = 0$ does not mean much.

The space $\mathbb{T}^n$ is the n-dimensional torus. The space $\text{AdS}^5 \times \text{S}^5$ is ten-dimensional and S^5 is a five-dimensional sphere with radius R. A sphere has positive curvature, but the anti-de Sitter space AdS^5 has a negative radius and is a negatively curved space. The Riemann curvature tensor is not zero, but the Ricci tensor is because the R and $-R$ cancel. Conformally-flat means that it can be mapped to flat space with a conformal transformation. If a space is maximally symmetric, then $R = 0$ determines the curvature completely. Ricci-flat spaces play an important role in General Relativity.

6 Physics in Curved Spaces

6.1 Conserved Quantities and Spacetime Symmetries

Recalling an argument from non-relativistic mechanics allows to conclude from the Euler-Lagrange equations

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}^i}\right) - \frac{\partial L}{\partial x^i} = 0 \qquad p^i \equiv \frac{\partial L}{\partial \dot{x}^i} \qquad \Rightarrow \qquad \frac{dp^i}{dt} = \frac{\partial L}{\partial x^i}$$

(where p^i is the momentum conjugate to coordinate x^i) that p^i is constant if L does not depend on x^i. This is a simple version of a more powerful argument by Emmy Noether that also applies to field theories and internal as well as spacetime symmetries. In general, a continuous symmetry of an action gives rise to a conserved current. Symmetries and conserved quantities are incredibly useful tools for solving equations of motion. The Lagrangian formulation of General Relativity is not shown here, but if one restricts to spacetime symmetries (also called isometries) then one can take a slightly different approach.

In General Relativity are 4-momenta given by $P^\mu = mU^\mu = m\frac{dX^\mu}{d\tau}$ for massive particles. The geodesic equation can be written as

$$\frac{dX^\nu}{d\tau}\,\nabla_\nu\,\frac{dX^\mu}{d\tau} = P^\nu\,\nabla_\nu\,P^\mu = 0 \tag{6.1}$$

and further processed by multiplying with $g_{\kappa\mu}$ to

$$\begin{aligned}
0 &= P^\nu\,\nabla_\nu\,P_\kappa = P^\nu(\partial_\nu\,P_\kappa - \Gamma^\lambda_{\nu\kappa}\,P_\lambda) = P^\nu\,\partial_\nu\,P_\kappa - P^\nu\,\Gamma^\lambda_{\nu\kappa}\,P_\lambda \\
&= m\frac{dX^\nu}{d\tau}\,\partial_\nu\,P_\kappa - P^\nu\,\frac{1}{2}g^{\lambda\rho}\,(\partial_\nu g_{\kappa\rho} + \partial_\kappa g_{\rho\nu} - \partial_\rho g_{\nu\kappa})\,P_\lambda = m\frac{dP_\kappa}{d\tau} - \frac{1}{2}\,(\partial_\nu g_{\kappa\rho} + \partial_\kappa g_{\rho\nu} - \partial_\rho g_{\nu\kappa})\,P^\nu P^\rho \\
&= m\frac{dP_\kappa}{d\tau} - \frac{1}{2}\,(\partial_\nu\,g_{\rho\kappa} - \partial_\rho\,g_{\nu\kappa})\,P^\nu P^\rho - \frac{1}{2}\,(\partial_\kappa\,g_{\rho\nu})\,P^\nu P^\rho = m\frac{dP_\kappa}{d\tau} - \frac{1}{2}\,(\partial_\kappa\,g_{\rho\nu})\,P^\nu P^\rho
\end{aligned}$$

using metric compatibility $\nabla_\nu\, g_{\kappa\mu}$ as well as the fact that $P^\nu P^\rho$ is symmetric and $\partial_\nu\, g_{\rho\kappa} - \partial_\rho\, g_{\nu\kappa}$ is antisymmetric under $\nu \leftrightarrow \rho$. Hence one finds that $m\frac{dP_\kappa}{d\tau} = 0$ if $\partial_\kappa\, g_{\rho\nu} = 0$ or, in other words, if $g_{\rho\nu}$ is independent of X^κ. This is a conservation result similar to p^i is conserved if L is independent of x^i.

As nice as this result seems, it is not particularly useful because it is a coordinate-dependent statement. If one uses Cartesian coordinates in $\mathbb{R}^2$, for example, then the metric is the identity and therefore independent of x and y, but in polar coordinates the metric depends on r. That is weird because it is the same space. One would like to have a coordinate-independent way of identifying symmetries and conserved quantities.

Thus the goal is to find a momentum component which is constant. If P^{σ^*} is this constant momentum component (one particular component and not a whole vector) such that $\frac{dP^{\sigma^*}}{d\tau} = 0$, then with $K^\mu\, P_\mu = P^{\sigma^*} = K_\mu\, P^\mu$ which is a scalar, the vector K^μ is defined and

$$\frac{dP^{\sigma^*}}{d\tau} = 0 = \frac{d(K_\mu\, P^\mu)}{d\tau} = \frac{dX^\nu}{d\tau}\,\partial_\nu(K_\mu\, P^\mu) = \frac{dX^\nu}{d\tau}\,\nabla_\nu(K_\mu\, P^\mu)$$
$$0 = m\left(\frac{dX^\nu}{d\tau}\,\nabla_\nu(K_\mu\, P^\mu)\right) = P^\nu\, P^\mu\,\nabla_\nu\, K_\mu + K_\mu\, P^\nu\,\nabla_\nu\, P^\mu = P^\nu\, P^\mu\,\nabla_\nu\, K_\mu$$

is satisfied where $P^\nu\,\nabla_\nu\, P^\mu = 0$ because of the geodesic equation in the form (6.1). One can rewrite this equation further as

$$0 = P^\nu\, P^\mu\,\nabla_\nu\, K_\mu = P^\nu\, P^\mu\,\nabla_{(\nu}\, K_{\mu)} + P^\nu\, P^\mu\,\nabla_{[\nu}\, K_{\mu]} = P^\nu\, P^\mu\,\nabla_{(\nu}\, K_{\mu)}$$

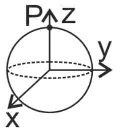
because one can split any matrix always into a symmetric and an antisymmetric part, and the antisymmetric part is zero because it itself is the product of the symmetric part $P^\nu\, P^\mu$ and the antisymmetric part $\nabla_{[\nu}\, K_{\mu]}$. This shows that K_μ is a symmetry of the geometry and $K_\mu\, P^\mu$ is a conserved quantity if $\nabla_{(\nu}\, K_{\mu)} = 0$.

The vector K^μ is called *Killing vector*, and the equation $\nabla_{(\nu}\, K_{\mu)} = 0$ which is a tensor equation (and therefore independent of the coordinates) is called the *Killing equation.* Killing vectors are in one-to-one correspondence to conserved quantities. For every Killing vector there is a symmetry of the geometry. To find a Killing vector one solves the Killing equation which is a differential equation. To solve it explicitly one has to choose a particular coordinate system, but as soon as one has found a solution, one can transform it to any coordinate system one wants.

The question is how many solutions of the Killing equation should one expect. In general this is hard to know in advance, but for *maximally symmetric* spaces there is a simple answer. Any manifold looks locally like $\mathbb{R}^n$ or $\mathbb{M}^n$ with n translations and $\frac{1}{2}n(n-1)$ rotations for $\mathbb{R}^n$ or Lorentz transformation for $\mathbb{M}^n$ which together form the Euclidean group or the Poincaré group, respectively. Thus in total there are $n + \frac{1}{2}n(n-1) = \frac{1}{2}n(n+1)$ local symmetries (10 in four dimensions).

This may not be the total number of Killing vectors because the entire space may not be as symmetric as $\mathbb{R}^n$ or $\mathbb{M}^n$ are. If there are all Killing vectors which are maximally possible (which means that all of the local symmetries are also globally valid), then the space is maximally symmetric. The spaces $\mathbb{R}^n$ or $\mathbb{M}^n$ are maximally symmetric, but there are other spaces which are maximally symmetric but are neither $\mathbb{R}^n$ nor $\mathbb{M}^n$.

The sphere S^2 is maximally symmetric, because through its embedding into $\mathbb{R}^3$ it is known that it is symmetric under rotations in the planes xy, xz, yz, and this gives $3 = \frac{1}{2}\,2\,(2+1)$ symmetries. From the point of view of somebody at the north pole P, a rotation in the xy-plane is a rotation, a rotation in the xz-plane is a translation in the x-direction, and a rotation in the yz-plane is a translation in the y-direction. Thus in fact there are not three rotations, but there are $n = 2$ translations and there is $\frac{1}{2}n(n-1) = 1$ rotation.



Maximally symmetric spaces do not have to be flat, but their curvature takes a simple form. Due to the translations invariance, if one knows the Riemann curvature tensor $R^\lambda{}_{\rho\mu\nu}$ at any point, it must have the same value at any other point and is therefore constant. For maximally symmetric spaces there is a simple way to calculate the curvature as

$$R_{\lambda\rho\mu\nu} = \frac{R}{n(n-1)}\,(g_{\lambda\mu}\, g_{\rho\nu} - g_{\lambda\nu}\, g_{\rho\mu}) \tag{6.2}$$

which does not contain any derivatives. Note that this equation satisfies the symmetries and antisymmetries of $R_{\lambda\rho\mu\nu}$. It is antisymmetric for $\lambda \leftrightarrow \rho$ and for $\mu \leftrightarrow \nu$, and it is symmetric for $\lambda\rho \leftrightarrow \mu\nu$.

This is a catalog of maximally symmetric spaces:

	Euclidean	Lorentzian
$R = 0$	$\mathbb{R}^n$, $\mathbb{T}^n$ (Euclidean, tori)	$\mathbb{M}^n$, $\mathbb{M}^n \times \mathbb{T}^n$ (Minkowski, Minkowski×tori)
$R > 0$	S^n (spheres)	dS^n (de Sitter)
$R < 0$	H^n (hyperbolic)	AdS^n (anti-de Sitter)

6.2 The Einstein Field Equations

All concepts have been presented to allow stating the content of General Relativity which answers two questions. The first question is how spacetime gets curved, and the second question is how this affects the behavior of particles. The answer to the first question is given by Einstein's field equations. The answer to the second question is that the path of a particle is a geodesic if gravity is the only acting force.

Spacetime gets curved according to *Einstein's field equations* which are

$$R_{\mu\nu} - \frac{1}{2}\, g_{\mu\nu}\, R = 8\pi\, G\, T_{\mu\nu} \tag{6.3}$$

where $R_{\mu\nu}$ is the Ricci tensor, $g_{\mu\nu}$ is the metric, R is the Ricci scalar, G is the Newton constant, and $T_{\mu\nu}$ is the energy-momentum tensor. The left side $G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}\, g_{\mu\nu}\, R$ is called the *Einstein tensor*, and the energy-momentum tensor $T_{\mu\nu}$ contains all sources including mass, energy, pressure and electromagnetic fields. Because $R_{\mu\nu}$, $g_{\mu\nu}$ and $T_{\mu\nu}$ are symmetric under $\mu \leftrightarrow \nu$, there are ten independent equations and the metric has ten independent components in four dimensions.

The Einstein field equations are similar to the Maxwell equations. One provides sources with $T_{\mu\nu}$ and then one solves the differential equations for the geometry in this case. The goal is to find the metric in order to calculate the Christoffel symbols, and one uses the Christoffel symbols to calculate the Riemann curvature tensor, and one uses the Riemann curvature tensor to finally calculate the Ricci tensor and the Ricci scalar.

Ten unknown components of the metric $g_{\mu\nu}$ and ten equations in (6.3) look fine, but the Riemann curvature tensor $R^{\lambda}{}_{\rho\mu\nu}$ defined in (5.10) also satisfies a geometric condition called *Bianchi identity* which states

$$\nabla^{\mu}\, R_{\nu\mu} = \frac{1}{2}\nabla_{\nu}\, R \tag{6.4}$$

and represents four differential equation for $\nu \in \{0, 1, 2, 3\}$. This gives fourteen equations but only ten unknowns. The Bianchi identity tells that some of the equations in the Einstein field equations are not really independent of each other, and counting them as ten independent equations is incorrect. Taking the Einstein equations and the Bianchi identity into account leaves six independent equations.

This situation is similar to electromagnetism where the two inhomogeneous Maxwell equations with ρ and $\vec{j}$ given are dynamical and correspond to four equations with six unknowns while the two homogeneous equations are geometrical and also correspond to four equations with the same six unknowns. The inhomogeneous equations are Euler-Lagrange equations coming from an action principle and the homogeneous equations are similar to the Bianchi identity. These are eight equations for six unknowns, and this is usually resolved by using the homogeneous equations. The electric field can be created as the gradient of a scalar field with $\vec{E} = \vec{\nabla}\Phi$, and the magnetic field can be created by the curl of a vector field with $\vec{B} = \vec{\nabla} \times \vec{A}$. In the end there are three equations with the four unknowns Φ and $\vec{A}$ from the potentials. This is fine because the potentials are only defined up to a gauge invariance.

The Bianchi identity reduces the ten equations in Einstein's equations to six independent equations in General Relativity, but the metric $g_{\mu\nu}$ has still ten unknown functions. Thus there are four remaining degrees of freedom which represent the freedom to change the four coordinates. This is in analogy to electromagnetism the gauge freedom of General Relativity, the freedom to redefine the coordinates. Changing the coordinates does not change the geometry but only the description of the geometry. The metric in General Relativity is analogous to the potentials in electromagnetism. To summarize, there is one unique geometry for a given $T_{\mu\nu}$, and this geometry can be represented by a family of metrics $g_{\mu\nu}$ related to each other by coordinate transformations. This answers the first question how spacetime is curved.

The second question is how this affects particles, or how curved spacetime changes physics. The answer uses a *minimal coupling principle.* The recipe to figure out how the familiar laws of physics get modified by curved spacetime is:

1. Start with a law valid in an inertial frame in flat spacetime.
2. Write the law in terms of true four-dimensional tensors.
3. Assert that the tensor form is true in curved spacetime as well.

In practice this means that one starts with a Lorentz invariant theory in terms of tensors and replaces $\eta_{\mu\nu} \to g_{\mu\nu}$ and $\partial_\mu \to \nabla_\mu$ everywhere. These steps allow to generalize any flat theory in flat space to curved spacetime.

For example the tensorial Maxwell equations in flat spacetime

$$\partial_\mu F^{\mu\nu} = J^\nu \qquad \partial_{[\mu} F^{\nu\lambda]} = 0 \qquad \text{where } F_{\nu\lambda} = \eta_{\nu\alpha}\,\eta_{\lambda\beta}\,F^{\alpha\beta}$$

become the Maxwell equations in curved spacetime

$$\nabla_\mu F^{\mu\nu} = J^\nu \qquad \nabla_{[\mu} F^{\nu\lambda]} = 0 \qquad \text{where } F_{\nu\lambda} = g_{\nu\alpha}\,g_{\lambda\beta}\,F^{\alpha\beta}$$

in electromagnetism.

For gravity as a second example particles move along straight lines in flat spacetime without gravity as

$$X^\mu(\lambda) = a^\mu \lambda + X^\mu_0 \qquad \text{or} \qquad \frac{d^2X^\mu}{d\lambda^2} = 0 \qquad \text{or} \qquad \frac{dX^\nu}{d\lambda}\,\partial_\nu\,\frac{dX^\mu}{d\lambda} = 0 \left(\text{using } \frac{d}{d\lambda} = \frac{dX^\nu}{d\lambda}\,\partial_\nu\right)$$

and move in curved spacetime as

$$\frac{dX^\nu}{d\lambda}\,\nabla_\nu\,\frac{dX^\mu}{d\lambda} = 0 \qquad \text{or} \qquad \frac{dX^\nu}{d\lambda}\,(\partial_\nu + \Gamma^\mu_{\nu\lambda})\,\frac{dX^\mu}{d\lambda} = 0 \qquad \text{or} \qquad \frac{d^2X^\mu}{d\lambda^2} + \Gamma^\mu_{\nu\kappa}\,\frac{dX^\nu}{d\lambda}\,\frac{dX^\kappa}{d\lambda} = 0$$

which is the geodesic equation in the form (5.9) and (6.1). The geodesic equation follows therefore from the above minimal coupling principle.

6.3 The Newtonian Limit

General Relativity is supposed to generalize Newtonian gravity. This means that there should be some sense in which one can take some limits and should get back Newtonian gravity. The two starting points are Einstein's equations (6.3) and the geodesic equation (5.9). Einstein's equations should tell how a Newtonian gravitational field is created from mass. In Newtonian physics only mass give raise to gravitational forces. The gravitational field can be derived from a gravitational potential. The equation to come out should be $\nabla^2\Phi = 4\pi G\rho$ where Φ is the gravitational potential and ρ is the mass density. The equation coming from the geodesic equation should be $\vec{a} = -\vec{\nabla}\Phi$ because of Newton's second law $\vec{F} = m\vec{a}$ together with $m\vec{a} = m\vec{g}$ and $-\vec{\nabla}\Phi = \vec{g}$, and because gravity is the only acting force.

The limits are firstly small velocities $\frac{dX^i}{d\tau} \ll \frac{dX^0}{d\tau}$, secondly a weak gravitational field $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$ for a small perturbation $h_{\mu\nu}$ with $||h_{\mu\nu}|| \ll 1$, and thirdly a static gravitational field $\partial_0 g_{\mu\nu} = 0$. As a consequence of the second limit $g^{\mu\nu} = \eta^{\mu\nu} - \eta^{\mu\alpha}\eta^{\nu\beta}h_{\alpha\beta}$ follows to ensure $g_{\lambda\mu}g^{\mu\nu} = \delta_\lambda{}^\nu$ as shown in

$$\begin{aligned} g_{\lambda\mu}g^{\mu\nu} &= (\eta_{\lambda\mu} + h_{\lambda\mu})(\eta^{\mu\nu} - \eta^{\mu\alpha}\eta^{\nu\beta}h_{\alpha\beta}) = \eta_{\lambda\mu}\eta^{\mu\nu} - \eta_{\lambda\mu}\eta^{\mu\alpha}\eta^{\nu\beta}h_{\alpha\beta} + \eta^{\mu\nu}h_{\lambda\mu} + O(h^2) \\ &= \delta_\lambda{}^\nu - \delta_\lambda{}^\alpha\eta^{\nu\beta}h_{\alpha\beta} + \eta^{\mu\nu}h_{\lambda\mu} = \delta_\lambda{}^\nu - \eta^{\nu\beta}h_{\lambda\beta} + \eta^{\mu\nu}h_{\lambda\mu} = \delta_\lambda{}^\nu \end{aligned}$$

ignoring $O(h^2)$ and higher.

The geodesic equation for a massive particle is

$$\frac{d^2X^\mu}{d\tau^2} + \Gamma^\mu_{\rho\sigma}\,\frac{dX^\rho}{d\tau}\,\frac{dX^\sigma}{d\tau} = 0 = \frac{d^2X^\mu}{d\tau^2} + \Gamma^\mu_{00}\left(\frac{dX^0}{d\tau}\right)^2$$

with $\lambda = \tau$ (proper time) and using the first limit. But

$$\Gamma^{\mu}_{00} = \frac{1}{2} g^{\mu\lambda} \left(\partial_0 g_{\lambda 0} + \partial_0 g_{0\lambda} - \partial_\lambda g_{00}\right) = -\frac{1}{2} \left(\eta^{\mu\lambda} - \eta^{\mu\alpha}\eta^{\lambda\beta} h_{\alpha\beta}\right) \partial_\lambda h_{00} \approx -\frac{1}{2}\eta^{\mu\lambda}\partial_\lambda h_{00}$$

using the third and second limit and ignoring $h_{\alpha\beta}\partial_\lambda h_{00}$ because the metric changes only slowly. Thus

$$\frac{d^2 X^\mu}{d\tau^2} + \Gamma^{\mu}_{00}\left(\frac{dX^0}{d\tau}\right)^2 \approx \frac{d^2 X^\mu}{d\tau^2} - \frac{1}{2}\eta^{\mu\lambda}\partial_\lambda h_{00}\left(\frac{dX^0}{d\tau}\right)^2 = 0$$

represents four equations. For $\mu = 0$ and $X^0 = t$ this means

$$\frac{d^2 t}{d\tau^2} - \frac{1}{2}\eta^{0\lambda}\partial_\lambda h_{00}\left(\frac{dt}{d\tau}\right)^2 = \frac{d^2 t}{d\tau^2} - \frac{1}{2}\eta^{00}\partial_0 h_{00}\left(\frac{dt}{d\tau}\right)^2 = \frac{d^2 t}{d\tau^2} + \frac{1}{2}\partial_0 h_{00}\left(\frac{dt}{d\tau}\right)^2 = \frac{d^2 t}{d\tau^2} = 0$$

using $\eta^{00} = -1$ and the third limit. This means that one can take $t = \tau$ which is good because in non-relativistic physics one can adequately parametrize motion with time t. For $\mu = i$ this means

$$\frac{d^2 X^i}{dt^2} - \frac{1}{2}\eta^{i\lambda}\partial_\lambda h_{00}\left(\frac{dt}{dt}\right)^2 = a^i - \frac{1}{2}\partial_i h_{00} = 0 \qquad \Rightarrow \qquad a^i = -\partial_i\left(-\frac{1}{2}h_{00}\right)$$

and gives $\vec{a} = -\vec{\nabla}\Phi$ with $\Phi = -\frac{1}{2}h_{00}$.

Writing Einstein's equations in trace-reverse form by building the trace over both sides gives

$$g^{\mu\nu}\left(R_{\mu\nu} - \frac{1}{2}\, g_{\mu\nu}\, R\right) = g^{\mu\nu}\left(8\pi\, G\, T_{\mu\nu}\right)$$

and $R - 2R = 8\pi G T$ or $R = -8\pi G T$. It follows

$$R_{\mu\nu} - \frac{1}{2}\, g_{\mu\nu}\, (-8\pi G T) = 8\pi\, G\, T_{\mu\nu} \qquad \text{or} \qquad R_{\mu\nu} = 8\pi G\left(T_{\mu\nu} - \frac{1}{2}g_{\mu\nu}T\right) \tag{6.5}$$

which is called the trace-reversed form of Einstein's equations. In Newtonian physics the only source for gravity is mass such that the only non-zero term in $T_{\mu\nu}$ is $T_{00} = \rho$. It follows

$$T = g^{\mu\nu}T_{\mu\nu} = \left(\eta^{\mu\nu} - \eta^{\mu\alpha}\eta^{\nu\beta}h_{\alpha\beta}\right)T_{\mu\nu} = \left(\eta^{00} - \eta^{0\alpha}\eta^{0\beta}h_{\alpha\beta}\right)T_{00} = (-1 - h_{00})\rho$$

because only $\eta^{00} = -1$ in $\eta^{0\alpha}\eta^{0\beta}$ is not zero. The right-hand side of the trace-reverse Einstein equation becomes

$$\begin{aligned} 8\pi G\left(T_{\mu\nu} - \frac{1}{2}g_{\mu\nu}T\right) &= 8\pi G\left[T_{00} - \frac{1}{2}g_{00}(-1 - h_{00})\rho\right] = 8\pi G\left[\rho - \frac{1}{2}(-1 - h_{00})(-1 - h_{00})\rho\right] \\ &= 8\pi G\left[\rho - \frac{1}{2}\rho + O(h^2)\right] = 4\pi G\rho \end{aligned}$$

and the left-hand side becomes

$$\begin{aligned} R_{00} = R^{\lambda}{}_{0\lambda 0} &= \partial_\lambda\Gamma^{\lambda}_{00} - \partial_0\Gamma^{\lambda}_{\lambda 0} + \Gamma^{\lambda}_{\lambda\kappa}\Gamma^{\kappa}_{00} - \Gamma^{\lambda}_{0\kappa}\Gamma^{\kappa}_{\lambda 0} = -\partial_0\Gamma^{0}_{00} + \partial_i\Gamma^{i}_{00} = \partial_i\Gamma^{i}_{00} \\ &= \partial_i\left(\frac{1}{2}g^{i\lambda}\left(\partial_0 g_{\lambda 0} + \partial_0 g_{0\lambda} - \partial_\lambda g_{00}\right)\right) = \partial_i\left(\frac{1}{2}g^{ij}\left(-\partial_j g_{00}\right)\right) = -\frac{1}{2}\partial_i g^{ij}\partial_j g_{00} \\ &= -\frac{1}{2}\partial_i\partial^i(-1 + h_{00}) = -\frac{1}{2}\nabla^2(h_{00}) = \nabla^2\Phi \end{aligned}$$

using $\Phi = -\frac{1}{2}h_{00}$ with the same Φ as above.

This shows that $\nabla^2\Phi = 4\pi G\rho$. There are two points to take from this. One is that when doing General Relativity and looking for solutions at some point the substitution $\Phi = -\frac{1}{2}h_{00}$ is made. The other is that General Relativity does not generalize Special Relativity but Newtonian gravity, and is at the same time a generalization of any physics to curved spacetime. In contrast to Newtonian physics, General Relativity is applicable in case of velocities close to the speed of light as Special Relativity, but also in case of strong and fast changing gravitational fields.

6.4 Central Forces in the Schwarzschild Metric

The Minkowski spacetime has the line element $ds^2 = -dt^2 + dx^2 + dy^2 + dz^2$ in rectangular coordinates (t, x, y, z). This is the geometry of flat spacetime. The Schwarzschild metric assumes spherical symmetry, and this is the only assumption which is a safe assumption in astronomy. The *Schwarzschild metric*

$$\begin{aligned} g_{\mu\nu} &= \begin{pmatrix} -\left(1-\frac{2GM}{r}\right) & 0 & 0 & 0 \\ 0 & \left(1-\frac{2GM}{r}\right)^{-1} & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & r^2\sin(\theta)^2 \end{pmatrix} \\ g^{\mu\nu} &= \begin{pmatrix} -\left(1-\frac{2GM}{r}\right)^{-1} & 0 & 0 & 0 \\ 0 & \left(1-\frac{2GM}{r}\right) & 0 & 0 \\ 0 & 0 & \frac{1}{r^2} & 0 \\ 0 & 0 & 0 & \frac{1}{r^2\sin(\theta)^2} \end{pmatrix} \end{aligned} \tag{6.6}$$

with the corresponding line element

$$ds^2 = -\left(1-\frac{2GM}{r}\right)dt^2 + \left(1-\frac{2GM}{r}\right)^{-1} dr^2 + r^2 d\theta^2 + r^2\sin(\theta)^2 d\phi^2$$

is diagonal and uses coordinates $\{t, r, \theta, \phi\}$ which are called Schwarzschild coordinates. They are not just spherical polar coordinates extended to spacetime. Often $r^2 d\Omega^2$ is written for $r^2 d\theta^2 + r^2\sin(\theta)^2 d\phi^2$ to simplify the line element by combining the angular dependences.

The geometry of the Schwarzschild metric describes the exterior of spherical source such as stars and planets. It represents a non-trivial curved geometry such that one can explore the minimal coupling principle (with the geodesic) in order to find predictions of General Relativity that differ from the Newtonian theory. Thus one can test General Relativity.

Central forces in Newtonian mechanics with coordinates (r, θ, ϕ) have the form $\vec{F} = f(r)\hat{r}$, and the torque is $\vec{\tau} = \vec{r} \times \vec{F} = \vec{0}$. Because $\vec{\tau} = d\vec{L}/dt = 0$ the angular momentum $\vec{L}$ is constant. Since the direction of $\vec{L}$ is constant, the motion must lie in a plane such that the problem is two-dimensional. One usually sets $\theta = \frac{\pi}{2}$ leaving r and ϕ as coordinates. Also the magnitude $L = mr^2\dot{\phi}$ is constant, and this allows to replace $\dot{\phi}$ by $\frac{L}{mr^2}$ when it seems convenient.

It is assumed that there are only two objects in the system. One is the source with mass M and the other is the test object with mass m. The total energy of the system is constant and can be written as

$$\begin{aligned} E_{\text{tot}} &= \frac{1}{2}m\,\vec{v}\cdot\vec{v} + V(r) = \frac{1}{2}m\,\dot{r}^2 + \frac{1}{2}m\,r^2\dot{\phi}^2 + V(r) = \frac{1}{2}m\,\dot{r}^2 + \frac{1}{2}m\,r^2\left(\frac{L}{mr^2}\right)^2 + V(r) \\ &= \frac{1}{2}m\,\dot{r}^2 + \frac{1}{2}\frac{L^2}{mr^2} + V(r) = \frac{1}{2}m\,\dot{r}^2 + V_{\text{eff}}(r) \end{aligned}$$

which is a differential equation for a one-dimensional motion in r. In the following the test mass is set $m = 1$ as long as $m > 0$.

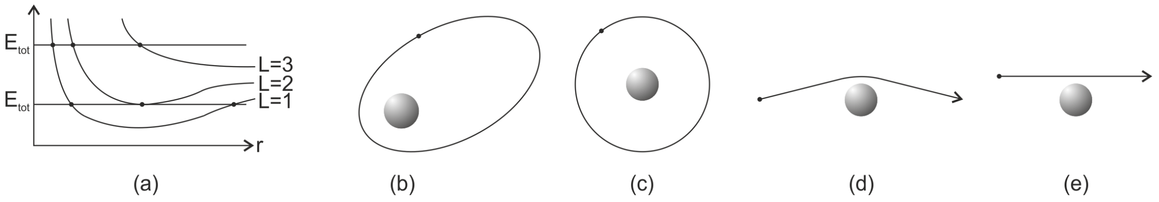


Figure 6.1: Newtonian central force

In figure 6.1 $V_{\text{eff}}(r) = \frac{L^2}{2r^2} - \frac{GM}{r}$ for the Newtonian gravitational potential is plotted for different values of L in (a). Depending on E_{tot} and L which are somewhat independent the behavior of the test mass is different. If the test mass has the lower E_{tot} in (a) then the test mass with $L = 1$ can move between two values of r and moves therefore on an ellipse as shown in (b) while it is stuck with one value of r with $L = 2$ and moves therefore on a circle as illustrated in (c). If the test mass comes in from infinity with the upper E_{tot} in (a) then it stays on the corresponding plotted curve for its value of L until it would need more energy than it has and therefore returns to infinity. This is scattering for all three values of L as drawn in (d). The circular and the elliptical orbit are stable in the sense that giving the test mass a little bit more energy keeps it on an elliptical orbit. The assumption was so far that $m > 0$. If $m = 0$ then this massless particle just passes the large mass M unaffected on a straight line as in figure (e).

The same central force can be calculated in the Schwarzschild metric (6.6). To proceed in a manner analogous to the Newtonian approach note that $K_t^\mu = (1, 0, 0, 0)$ and $K_\phi^\mu = (0, 0, 0, 1)$ are killing vectors because $g_{\mu\nu}$ is independent of t and ϕ. (Thus one does not have to solve the Killing equation but can simply use these two vectors.) Using $K_{t\mu} = (-(1 - \frac{2GM}{r}), 0, 0, 0)$ and $K_{\phi\mu} = (0, 0, 0, r^2)$ by setting $\theta = \frac{\pi}{2}$ in $r^2 \sin(\theta)^2$ the conserved "momenta" are

$$K_{t\mu}P^\mu = K_{t\mu}\frac{dX^\mu}{d\lambda} = -\left(1 - \frac{2GM}{r}\right)\frac{dt}{d\lambda} = E \qquad K_{\phi\mu}P^\mu = K_{\phi\mu}\frac{dX^\mu}{d\lambda} = r^2\frac{d\phi}{d\lambda} = L$$

since $m = 1$. The energy E and the angular momentum L are conserved quantities and therefore constant. The goal is now to get a differential equation in r such that it can be compared with the Newtonian case, but there is no $\frac{dr}{d\lambda}$ term.

However, there is another quantity ϵ with

$$\epsilon = -g_{\mu\nu}\frac{dX^\mu}{d\lambda}\frac{dX^\nu}{d\lambda} \qquad \Rightarrow \qquad \frac{D\epsilon}{d\lambda} = -g_{\mu\nu}\left[\left(\frac{D}{d\lambda}\frac{dX^\mu}{d\lambda}\right)\frac{dX^\nu}{d\lambda} + \frac{dX^\mu}{d\lambda}\left(\frac{D}{d\lambda}\frac{dX^\nu}{d\lambda}\right)\right]$$
$$= -g_{\mu\nu}\left[0\frac{dX^\nu}{d\lambda} + \frac{dX^\mu}{d\lambda}0\right] = 0$$

which is because of (5.9) a conserved quantity for geodesics. But ϵ can also be written as

$$\epsilon = -g_{\mu\nu}\frac{dX^\mu}{d\lambda}\frac{dX^\nu}{d\lambda} = \left(1 - \frac{2GM}{r}\right)\left(\frac{dt}{d\lambda}\right)^2 - \left(1 - \frac{2GM}{r}\right)^{-1}\left(\frac{dr}{d\lambda}\right)^2 - r^2\left(\frac{d\phi}{d\lambda}\right)^2$$
$$= \frac{E^2}{1 - \frac{2GM}{r}} - \frac{\left(\frac{dr}{d\lambda}\right)^2}{1 - \frac{2GM}{r}} - \frac{L^2}{r^2}$$

and be rearranged to

$$\frac{1}{2}E^2 = \frac{1}{2}\left(\frac{dr}{d\lambda}\right)^2 + \frac{1}{2}\epsilon - \frac{GM}{r}\epsilon + \frac{L^2}{2r^2} - \frac{GML^2}{r^3} = \frac{1}{2}\left(\frac{dr}{d\lambda}\right)^2 + V_{\text{eff}}(r)$$

with an effective potential similar to the Newtonian case. For $m > 0$ with $\lambda = \tau$ is $\epsilon = -U_\mu U^\mu = +1$, and for $m = 0$ where $ds^2 = 0$ is $\epsilon = 0$. The effective potential plotted in figure 6.2 is

$$V_{\text{eff}}(r) = \frac{L^2}{2r^2} - \frac{GML^2}{r^3} \qquad V_{\text{eff}}(r) = -\frac{GM}{r} + \frac{L^2}{2r^2} - \frac{GML^2}{r^3}$$

for massless particles (a) on the left side and for massive particles (b) on the right side.

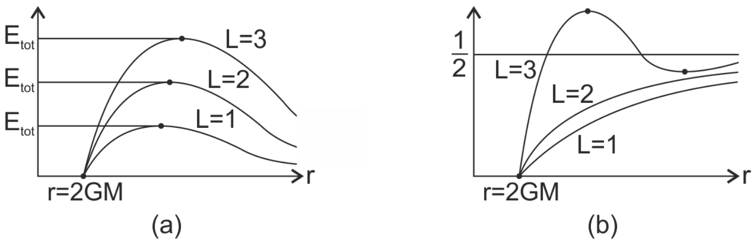


Figure 6.2: Central force in General Relativity

The difference to the Newtonian case is the term $-\frac{GML^2}{r^3}$ in both cases. For $m = 0$ figure (a) shows on one hand that gravity affects massless objects and on the other hand that the orbits of the massless objects are unstable. For $m > 0$ the curve for $L = 3$ in figure (b) shows an unstable circular orbit and a stable circular orbit. Thus if L is large enough then there is a stable/unstable pair of circular orbits, but if L is too small then no circular orbit exists. This fact in the form

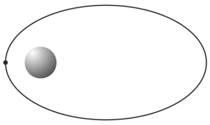
$$\left.\frac{V_{\text{eff}}(r)}{dr}\right|_{r_c} = 0 \qquad \Rightarrow \qquad r_{c\pm} = \frac{L^2 \pm \sqrt{L^4 - 12\,G^2M^2L^2}}{2GM}$$

is a minimum radius condition where r_{c+} is the stable and r_{c-} the unstable solution. The smallest possible radius is when $r_{\text{min}} = r_{c+} = r_{c-} = \frac{L^2}{2GM} = 6GM$, and below this value for r there are no stable circular orbits. So to test General Relativity one can just look for the instability of circular orbits with $r < r_{\text{min}}$ because Newtonian physics predicts stable circular orbits below this value. Some example values for $r_{\text{min}} = 6GM/c^2$ which is the value for $c \neq 1$ show

Massive Object	r_{min}	Radius R	Comment
Earth	$\approx 0.03\,\text{m}$	$\approx 6 \cdot 10^6\,\text{m}$	Inside the earth
Sun	$\approx 8850\,\text{m}$	$\approx 7 \cdot 10^8\,\text{m}$	Inside the sun
White Dwarf	$\approx 8850\,\text{m}$	$\approx 10^6\,\text{m}$	Inside the star
Neutron Star	$\approx 8850\,\text{m}$	$\approx 10^4\,\text{m}$	Inside the star
...			
Black Hole	$\approx 8850\,\text{m}$	$\approx 0\,\text{m}$	Outside the black hole

that this radius r_{min} is only for black holes not inside the object with mass M. The event horizon is at $r = 2GM$ and r_{min} is well outside despite that the radius of the black hole is approximately zero. Therefore this would be a test for General Relativity, but so far it is not possible to see whether a stable orbit is possible below this critical radius.

6.5 The Perihelion Shift



The perihelion shift provides a practical test for General Relativity based on existing astronomical data. The point of closest approach on an elliptical orbit is called perihelion. It is known that the perihelion of the planet mercury has moved a bit after every orbit.

In the Newtonian case, $E = \frac{1}{2}\left(\frac{dr}{dt}\right)^2 - \frac{GM}{r} + \frac{L^2}{2r^2}$ and $L = r^2\frac{d\phi}{dt}$ used in the form $dt \to \frac{r^2}{L}d\phi$ give a differential equation for r depending on ϕ

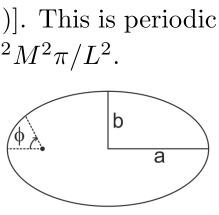
$$\left(\frac{dr}{d\phi}\right)^2 - \frac{2GM}{L^2}r^3 + r^2 = \frac{2E}{L^2}r^4 \qquad \Rightarrow \qquad r(\phi) = \frac{L^2}{GM(1 + e\cos(\phi))}$$

where $e = e(G, M, L)$ is a constant depending on G, M and L which are all constants. Because $\cos(\phi + 2\pi) = \cos(\phi)$ and therefore $r(\phi + 2\pi) = r(\phi)$ the perihelion does not change.

In General Relativity using $V_{\text{eff}}(r)$ and a similar rewriting one can show that to leading order

$$r(\phi) = \frac{L^2}{GM(1 + e\cos[(1-\alpha)\phi])}$$

where $\alpha = \frac{3G^2M^2}{L^2}$. Thus $r(\phi + 2\pi) \neq r(\phi)$ because $\cos[(1-\alpha)\phi] = \cos[(1-\alpha)(\phi + \frac{2\pi}{1-\alpha})]$. This is periodic in $\frac{2\pi}{1-\alpha} \approx 2\pi(1+\alpha) = 2\pi + 2\pi\alpha$. The angular shift after one orbit is $\Delta\pi = 2\pi\alpha = 6G^2M^2\pi/L^2$.



To get a numerical value L is needed which one can get by measuring the semi-major a and the semi-minor b axes of the orbit. The equation of an ellipse can be written as a function r depending on ϕ with

$$r(\phi) = \frac{(1-e^2)a}{1 + e\cos(\phi)} \qquad\qquad e = \sqrt{1 - \frac{b^2}{a^2}}$$

and comparing it with the Newtonian result (in leading order)

$$r(\phi) = \frac{L^2}{GM(1 + e\cos(\phi))} \qquad \Rightarrow \qquad \frac{L^2}{GM} = (1-e^2)a \qquad \Rightarrow \qquad \Delta\phi = \frac{6\pi GM}{(1-e^2)a}$$

allows to enter the data for the solar system. Note that $\Delta\phi$ is largest for the smallest a which is mercury. Using $a = 5.79 \cdot 10^{10}\,\mathrm{m}$, $e = 0.2056$ and $T = 88\,\mathrm{days}$ gives $\Delta\phi = 43$ arcseconds per century, but the observed value is $\Delta\phi = 5601$ arcseconds per century. However taking into account precession of equinoxes, gravitational pull of other planets and oblateness of the sun gives 5558 arcseconds per century alone from these effects. Before General Relativity and therefore just with Newtonian physics these two values have been found. Thus the missing piece for $5558 + 43 = 5601$ was the first experimental confirmation of General Relativity.

7 Solutions to the Einstein Field Equations and Black Holes

7.1 Exterior Schwarzschild Solution

In a first attempt at finding an exact solution to Einstein's equations a spherically symmetric solution is looked for because symmetries simplify mathematical problems such that they become easier than less symmetric solutions. It is obviously relevant for astrophysical applications because stars, planets and other objects in astrophysics are spherically symmetric as a good approximation such that one can use this symmetry for the choice of the coordinates.

Seeking solutions for Gauss' law $\vec{\nabla} \cdot \vec{E} = \frac{\rho}{\epsilon_0}$ in electromagnetism as a similar problem becomes also easier if one is interested in spherically symmetric solutions for the electric field. The goal is to find the electric field outside a time independent spherical source of charge such that $\rho = 0$.

Thus $\vec{E} = f(r)\,\hat{r}$ is the general form of a solution to $\vec{\nabla} \cdot \vec{E} = 0$ in $\{r, \theta, \phi\}$ coordinates. Gauss' law in spherical coordinates is

$$\vec{\nabla} \cdot \vec{E} = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 f(r)\right) = 0 \qquad \Rightarrow \qquad f(r) = \frac{k}{r^2} \qquad \Rightarrow \qquad \vec{E} = \frac{k}{r^2}\,\hat{r}$$

but that does not tell what k is. One integrates $\vec{\nabla} \cdot \vec{E} = \frac{\rho}{\epsilon_0}$ over a sphere and gets

$$\int \vec{\nabla} \cdot \vec{E}\, d^3x = \int \frac{\rho}{\epsilon_0}\, d^3x \qquad \int \frac{\rho}{\epsilon_0}\, d^3x = \frac{Q_{\text{enc}}}{\epsilon_0} \qquad \int \vec{\nabla} \cdot \vec{E}\, d^3x = 4\pi r^2 E \qquad 4\pi r^2 E = \frac{Q_{\text{enc}}}{\epsilon_0}$$

$$E = \frac{Q_{\text{enc}}}{4\pi\epsilon_0 r^2} \qquad k = \frac{Q_{\text{enc}}}{4\pi\epsilon_0} \qquad \Rightarrow \qquad \vec{E} = \frac{Q_{\text{enc}}}{4\pi\epsilon_0 r^2}\hat{r}$$

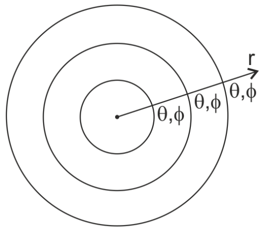
as the complete solution. This works when ρ is extended in some way (shell or a volume of charge) or pointlike as long as it is spherically symmetric and it is a solution outside. Of course if ρ is extended with radius R, one has to find also an interior solution for $r < R$.

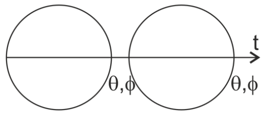
Solving this equation so easily was greatly facilitated by choosing the right coordinates. Of course one could have started with any coordinate system, but the process would have been much more complicated. Along the way one might have seen how various coordinate redefinitions could simplify the process and eventually get to spherical polar coordinates. Once one has a solution in any coordinate system, on can take it and transfer it to any coordinate system.

To find the Schwarzschild solution of the Einstein field equations it is assumed that one is outside of some mass distribution and seeks an exterior solution. Because the outside of this mass distribution is free of sources the energy-momentum tensor is zero. Using the trace-reversed version (6.5) of Einstein's equation with $T_{\mu\nu} = 0$ and therefore also $T = 0$ gives $R_{\mu\nu} = 0$ which is the vacuum form of Einstein's equation. To solve this gives ten independent functions $g_{\mu\nu}$ in coordinates $\{t, r, \theta, \phi\}$ which look like spherical polar coordinates. Spherical symmetry will say a lot about the ten functions to be determined.

One way to impose spherical symmetry is by imagining the spacetime in terms of an S^2-foliation. This means that one builds up the spacetime by stacking concentric two-spheres (like an onion) along a radial direction r and lined up along t. This has some consequences. If one looks at one of the two-spheres such that r and t are fixed then the metric should take the form $d\Omega^2 = d\theta^2 + \sin(\theta)^2\, d\phi^2$. It follows that $g_{\theta\theta} = 1$, $g_{\phi\phi} = \sin(\theta)$, $g_{\theta\phi} = 0$, and that the rest of the metric components do not depend on θ and ϕ. The two-spheres separate the coordinates t and r from the coordinates θ and ϕ.

Further if one aligns the two-sphere shells such that radial geodesics go through the same value of θ and ϕ on each slice then $g_{r\theta} = 0$ and $g_{r\phi} = 0$. Additionally, if one stacks the two-spheres along t such that motion purly along t keeps the values of θ and ϕ unchanged then $g_{t\theta} = 0$ and $g_{t\phi} = 0$.





So spherical symmetry eliminates five of the ten unknown functions, combines two into $d\Omega$ and tells that nothing else depends on θ and ϕ. With this knowledge the metric can be written in the form

$$ds^2 = -A(r,t)\,dt^2 + 2B(r,t)\,dr\,dt + C(r,t)\,dr^2 + D(r,t)\,r^2\,d\Omega^2$$

where the factor of two for $B(r,t)$ comes from the fact that this is the only off-diagonal element in the matrix and appears therefore twice. The goal now is to find the four functions $A(r,t)$, $B(r,t)$, $C(r,t)$ and $D(r,t)$.

One can now apply the gauge freedom and redefine $r \to r' = \sqrt{D(r,t)}r$ which can be inverted to find $r(r',t)$ as soon as $D(r,t)$ is known. Thus the line element is now

$$ds^2 = -A(r',t)\,dt^2 + 2B(r',t)\,dr'\,dt + C(r',t)\,dr'^2 + r'^2\,d\Omega^2$$

where this step is similar to choosing a gauge to simplify solving Maxwell's equation for ϕ and $\vec{A}$. Assuming this redefinition the primes will be dropped henceforth.

Exploiting this gauge freedom even further with $t \to t - f(r,t')$ and therefore $t = t' + f(r,t')$ changes

$$dt = dt' + \frac{\partial f}{\partial r}\,dr + \frac{\partial f}{\partial t'}\,dt' = \frac{\partial f}{\partial r}\,dr + \left(1 + \frac{\partial f}{\partial t'}\right)dt'$$

$$dt^2 = \left(\frac{\partial f}{\partial r}\right)^2 dr^2 + 2\,\frac{\partial f}{\partial r}\left(1 + \frac{\partial f}{\partial t'}\right)dr\,dt' + \left(1 + \frac{\partial f}{\partial t'}\right)^2 dt'^2$$

and gives

$$\begin{aligned} ds^2 = &-A(r,t')\left[\left(\frac{\partial f}{\partial r}\right)^2 dr^2 + \left(1 + \frac{\partial f}{\partial t'}\right)^2 dt'^2 + 2\,\frac{\partial f}{\partial r}\left(1 + \frac{\partial f}{\partial t'}\right)dr\,dt'\right] \\ &+ 2\,B(r,t')\left[\frac{\partial f}{\partial r}\,dr^2 + \left(1 + \frac{\partial f}{\partial t'}\right)dr\,dt'\right] + C(r,t')\,dr^2 + r^2\,d\Omega^2 \end{aligned}$$

putting it back into the metric. The function f can be chosen freely and it is selected such that the $drdt'$ cross-term disappears. This can be done with

$$2\left[-A(r,t')\,\frac{\partial f}{\partial r} + B(r,t')\right]\left(1 + \frac{\partial f}{\partial t'}\right)dr\,dt' = 0 \qquad \Rightarrow \qquad \frac{\partial f}{\partial r} = \frac{B(r,t')}{A(r,t')}$$

$$\Rightarrow \qquad f(r,t') = \int \frac{B(r,t')}{A(r,t')}\,dr + g(t')$$

and assuming that $t \to t'$ has been redefined using just such a function $f(r,t')$ then

$$ds^2 = -\left[A(r,t')\left(1 + \frac{\partial f}{\partial t'}\right)\right]dt'^2 + \left[-A(r,t')\left(\frac{B(r,t')}{A(r,t')}\right)^2 + 2\,B(r,t')\frac{B(r,t')}{A(r,t')} + C(r,t')\right]dr^2 + r^2\,d\Omega^2$$

becomes the new metric. Dropping the primes again the metric is essentially

$$ds^2 = g_{tt}(r,t)\,dt'^2 + g_{rr}(r,t)\,dr^2 + r^2\,d\Omega^2$$

with two unknown functions $g_{tt}(r,t)$ and $g_{rr}(r,t)$.

This metric is describing a spacetime and one can therefore assume that $A(r,t)$ and $C(r,t)$ are both positive and therefore $g_{tt}(r,t) < 0$. It also follows $g_{rr}(r,t) > 0$ without any assumption about $B(r,t)$ and one can write

$$g_{tt}(r,t) = -e^{2\alpha(r,t)} \qquad\qquad g_{rr}(r,t) = e^{2\beta(r,t)}$$

for the two functions to enforce these signs. Everything so far has only used spherical symmetry and gauge choices. Calculating the Ricci tensor using this metric for the Einstein vacuum equation $R_{\mu\nu} = 0$ with a computer tool results in a set of differential equations and shows that the diagonal elements of the Ricci tensor and R_{tr} are the only non-zero values. Further steps must now make sure that also these elements vanish.

In a first step one can show that $R_{tr} = \frac{2}{r}\frac{\partial\beta}{\partial t} = 0$ such that $\beta(r,t) = \beta(r)$. In a second step follows from $R_{\theta\theta} = e^{-2\beta}[r(\frac{\partial\beta}{\partial r} - \frac{\partial\alpha}{\partial r}) - 1] + 1 = 0$ and from the first step $\frac{\partial\beta}{\partial t} = 0$ that $-re^{2\beta}\,\partial_t\,\partial_r\,\alpha(r,t) = 0$ and therefore $\alpha(r,t) = f(r) + g(t)$.

The resulting metric and the redefinition $t \to t'$

$$ds^2 = -e^{2f(r)}e^{2g(t)}dt^2 + e^{2\beta(r)}dr^2 + r^2 d\Omega^2 \qquad t \to t' = \int e^{g(t)}dt \qquad dt' = e^{g(t)}dt$$

lead to the metric

$$ds^2 = -e^{2f(r)}dt'^2 + e^{2\beta(r)}dr^2 + r^2 d\Omega^2$$

which is diagonal and where none of the metric components depend on t'. The metric does therefore not depend on time such that $g_{\mu\nu}(r,t) = g_{\mu\nu}(r)$. Such a space is called stationary, and this implies the existence of a timelike Killing vector ∂_t. But this space is also invariant under $t \leftrightarrow -t$ due to the absence of any $dx'\,dt$ cross-terms. This condition makes the space static[9]. Thus the assumption of a spherically symmetric source-free solution implies a static geometry.

Returning to $R_{\mu\nu} = 0$ shows in the next two steps

$$\begin{aligned} R_{tt} &= e^{2(f-\beta)}\left[\frac{\partial^2 f}{\partial r^2} + \left(\frac{\partial f}{\partial r}\right)^2 - \frac{\partial f}{\partial r}\frac{\partial\beta}{\partial r} + \frac{2}{r}\frac{\partial f}{\partial r}\right] = 0 \quad &\Rightarrow \quad &\frac{\partial^2 f}{\partial r^2} + \left(\frac{\partial f}{\partial r}\right)^2 - \frac{\partial f}{\partial r}\frac{\partial\beta}{\partial r} + \frac{2}{r}\frac{\partial f}{\partial r} = 0 \\ R_{rr} &= &&-\frac{\partial^2 f}{\partial r^2} - \left(\frac{\partial f}{\partial r}\right)^2 + \frac{\partial f}{\partial r}\frac{\partial\beta}{\partial r} + \frac{2}{r}\frac{\partial\beta}{\partial r} = 0 \end{aligned}$$

and adding the two similar expressions gives $\frac{2}{r}(\frac{\partial f}{\partial r} + \frac{\partial\beta}{\partial r}) = 0$ or $f(r) = -\beta(r) + c$. The metric becomes

$$ds^2 = -e^{-2\beta(r)}e^{2c}dt^2 + e^{2\beta(r)}dr^2 + r^2 d\Omega^2 \qquad ds^2 = -e^{-2\beta(r)}dt'^2 + e^{2\beta(r)}dr^2 + r^2 d\Omega^2$$

after another redefinition $e^{2c}dt^2 \to dt'^2$.

The final step leads to

$$R_{\theta\theta} = e^{2f}\left(-2r\frac{\partial f}{\partial r} - 1\right) + 1 = 0 \qquad \Rightarrow \frac{\partial}{\partial r}\left(r\,e^{2f}\right) \qquad = 1 \qquad \Rightarrow e^{2f} = e^{-2\beta} = 1 + \frac{c}{r}$$

for the functions $f(r)$ and $\beta(r)$ and to

$$ds^2 = -\left(1 + \frac{c}{r}\right)dt^2 + \left(1 + \frac{c}{r}\right)^{-1}dr^2 + r^2\,d\Omega^2$$

as the metric.

It is known from the Newtonian limit that $g_{00} = -(1 + 2\phi) = -(1 - \frac{2GM}{r})$ because $\phi = -\frac{GM}{r}$ such that the Schwarzschild metric in Schwarzschild coordinates is

$$ds^2 = -\left(1 - \frac{2GM}{r}\right)dt^2 + \left(1 - \frac{2GM}{r}\right)^{-1}dr^2 + r^2 d\Omega^2$$

as used above in (6.6). This is the geometry outside of any spherically symmetric mass such as a planet, a star or a black hole. The quantity $R = 2GM$ is called the *Schwarzschild radius*. For $r \to \infty$ or $M \to 0$ the metric becomes $\mathbb{M}^4$ as expected.

[9] To understand the difference between static and stationary one can consider a planet. If it is sitting still it creates a static geometry, and if it is spinning it creates a stationary geometry since $t \leftrightarrow -t$ reverses the spin.

7.2 Interior Schwarzschild Solution

To find the solution in the interior of an extended object with spherical symmetry the similarity to electromagnetism is again explored.

The equation $\vec{\nabla} \cdot \vec{E} = 0$ has been solved for $\vec{E}_{\text{out}} = \frac{Q_{\text{tot}}}{4\pi\epsilon_0}$. Now a solution $\vec{E}_{\text{in}}$ of $\vec{\nabla} \cdot \vec{E} = \frac{\rho}{\epsilon_0}$ is needed. There are various models of charge distribution possible which lead to different solutions. Here the simple uniform distribution (a uniformly charged ball)

$$\rho(r) = \begin{cases} \rho_* & r \leq R \\ 0 & r > R \end{cases} \qquad \text{such that } Q_{\text{tot}} = \tfrac{4}{3}\pi R^3 \rho$$

is assumed where R is the radius of the spherically symmetric charged object.

To solve for $\vec{E}_{\text{in}}$ one integrates over a spherical volume centered at the origin with $r < R$

$$\int \vec{\nabla} \cdot \vec{E}\, d^3x = \int \frac{\rho_*}{\epsilon_0}\, d^3x \qquad \int \vec{E}_{\text{in}} \cdot d\vec{a} = \frac{4}{3}\pi r^3 \frac{\rho_*}{\epsilon_0} = \frac{Q_{\text{tot}}}{\epsilon_0} \frac{r^3}{R^3} \qquad 4\pi r^2 E = \frac{Q_{\text{tot}}}{\epsilon_0} \frac{r^3}{R^3}$$

and gets

$$\vec{E}_{\text{in}} = \frac{Q_{\text{tot}}}{4\pi\epsilon_0} \frac{r}{R^3}\, \hat{r}$$

which shows that the functional dependence on r is different for $\vec{E}_{\text{in}} \sim r$ and $\vec{E}_{\text{out}} \sim \frac{1}{r^2}$, and that the solutions agree on the boundary.

The equivalent task in General Relativity is to solve Einstein's equations $G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} = 8\pi G T_{\mu\nu}$ in a region where $T_{\mu\nu} \neq 0$ and find a solution $g_{\mu\nu}$ that matches the Schwarzschild solution at the boundary. This is considerably more complicated than $R_{\mu\nu} = 0$, and trace reversing does not help either. However using spherical symmetry allows to adopt many of the results from the exterior analysis.

One can start from the metric

$$ds^2 = -e^{2\alpha(r,t)}\, dt^2 + e^{2\beta(r,t)}\, dr^2 + r^2\, d\Omega^2$$

which was found above without using $R_{\mu\nu} = 0$. For the exterior solution dependence on time has been eliminated, but for the interior case this will not generally be the case. Here time independence will be assumed for the model of the interiors such that time independence can also be assumed for the solution. This must be checked for consistency at the end of the calculation. Again using a mathematical computer tool gives non-trivial expressions for the diagonal elements $G_{\mu\mu}$ and all the others vanish. (Note that a diagonal metric does not always yield a diagonal Einstein tensor.)

One has to put also information about the source into $T_{\mu\nu}$ which is here assumed to be a perfect fluid source $T_{\mu\nu} = (\rho + p) U_\mu U_\nu + p g_{\mu\nu}$. Later also an equation of state relating ρ and p is needed. The quantity U_μ is the overall fluid dual 4-velocity and $U_\mu U^\mu = -1 = g^{\mu\nu} U_\mu U_\nu$. Because coordinates are assumed such that the overall 3-velocity of the source is zero, one can further conclude $-1 = g^{00} U_0 U_0 = -e^{-2\alpha} U_0 U_0$ and therefore also $U_0 = e^\alpha$. The off-diagonal elements of $T_{\mu\nu}$ are zero, and the four diagonal elements of $T_{\mu\nu}$ are $T_{tt} = e^{2\alpha(r)}\, \rho(r)$, $T_{rr} = e^{2\beta(r)}\, p(r)$, $T_{\theta\theta} = r^2\, p(r)$, $T_{\phi\phi} = r^2\, \sin(\theta)^2\, p(r)$ where $\alpha(r)$, $\beta(r)$, $\rho(r)$, $p(r)$ are unknown functions.

Motivated by the Schwarzschild solution one can exchange the unknown $\beta(r)$ for $M(r)$

$$e^{2\beta(r)} = \left[1 - \frac{2\, G\, M(r)}{r}\right]^{-1} \qquad \Rightarrow \qquad M(r) = \frac{r}{2G} \left[1 - e^{-2\beta(r)}\right]$$

and analyze the $G_{tt} = 8\pi G T_{tt}$ component to find $r^{-2}[e^{2\alpha - 2\beta}(-1 + e^{2\beta} + 2r\beta')] = 8\pi G e^{\alpha\alpha} \rho$ which becomes $\frac{dM}{dr} = 4\pi r^2 \rho$ with $\beta \to M$. Integrating yields

$$M(r) = 4\pi \int_0^r \rho(r')\, r'^2\, dr' \qquad M(R) = M = 4\pi \int_0^R \rho(r')\, r'^2\, dr'$$

for $0 < r \leq R$ with the matching condition for $r = R$ where $M(R) = M$. Given a density ρ the function $M(r)$ and therefore also $\beta(r)$ can be calculated.

To find also $\alpha(r)$ one can consider $G_{rr} = 8\pi G T_{rr}$ and gets from $r^{-2}[1 - e^{\alpha\beta} + 2r\alpha'] = 8\pi G e^{2\beta} p$

$$\frac{d\alpha}{dr} = \alpha' = \frac{GM(r) + 4\pi G r^3 p}{r[r - 2GM(r)]}$$

also with $\beta \to M$. Given a function for the pressure p the function α can be calculated.

Instead of considering the rest of Einstein's equations one can instead appeal to energy-momentum conservation $\nabla_\mu T^{\mu\nu} = 0$ which implies $(\rho + p)\frac{d\alpha}{dr} = -\frac{dp}{dr}$ from the $\nu = r$ term. Combining this with the above equation for α' results in

$$\frac{dp}{dr} = -\frac{(\rho + p)\left[GM(r) + 4\pi G r^3 p\right]}{r[r - 2GM(r)]}$$

called the Tolman-Oppenheimer-Volkoff equation. The importance of this equation is that it provides an equation of state. If one starts with $\rho(r)$ one can firstly find $M(r)$ (and also $\beta(r)$), secondly $p(r)$ and finally $\alpha(r)$.

As an example in analogy to the example from electromagnetism a star with constant density

$$\rho(r) = \begin{cases} \rho_* & r \leq R \\ 0 & r > R \end{cases} \qquad \Rightarrow \qquad M(r) = \begin{cases} \frac{4}{3}\,\pi\, r^2\, \rho_* & r \leq R \\ \frac{4}{3}\,\pi\, R^2\, \rho_* = M & r > R \end{cases}$$

is used. This allows to calculate the pressure with

$$\frac{dp}{dr} = -\frac{(\rho_* + p)\left[G\,\frac{4}{3}\,\pi\, r^2\, \rho_* + 4\,\pi\, G\, r^3\, p\right]}{r[r - 2G\,\frac{4}{3}\,\pi\, r^2\, \rho_*]} \qquad \Rightarrow \qquad \rho_* \frac{R\sqrt{R - 2GM} - \sqrt{R^3 - 2GMr^2}}{\sqrt{R^3 - 2GMr^2} - 3R\sqrt{R - 2GM}}$$

where $r < R$ is assumed. Finally solving for $\alpha(r)$ gives

$$e^{\alpha(r)} = \frac{2}{3}\left(1 - \frac{2GM}{R}\right)^{\frac{1}{2}} - \frac{1}{2}\left(1 - \frac{2GMr^2}{R^3}\right)^{\frac{1}{2}}$$

where $r < R$ is also assumed. Not only $e^{2\beta(r)}$ but also $e^{2\alpha(r)} = \left(1 - \frac{2GM}{R}\right)$ matches the Schwarzschild solution at $r = R$. Analyzing the $p(r)$ expression shows:

- As to be expected, the pressure $p(r)$ increases as r decreases.
- The pressure $p(r)$ at $r = 0$ diverges as $M \to \frac{4}{9}\frac{R}{G}$. If the mass is less, the pressure does not diverge but at this value or higher the pressure gets infinite.
- Therefore if the mass with respect to the radius is $M > \frac{4}{9}\frac{R}{G}$ then this solution is inconsistent. The consistency failure comes from the static assumption, and the system must evolve with time above this value.

7.3 Stellar Collapse

In the time-independent example of a uniform density the pressure goes to infinity for $M > \frac{4}{9}\frac{R}{G}$. The energy density ρ is responsible for the gravitational attraction and tries to pull everything to the center while the pressure p pushing outward tries to balance it. At this limit the pressure can no longer keep up, and the system collapses. This means that the mass M stays the same but the radius R gets smaller and smaller such that $\frac{4}{9}\frac{R}{G}$ also gets smaller and smaller.

To make sure that this situation can really happen, a first question is how realistic is the assumption of a constant density ρ. Although this is actually a pretty good model for stellar objects, it can be shown that for more general densities ρ and spherical symmetry, the condition $M > \frac{4}{9}\frac{R}{G}$ still leads to a collapse (Buchdahl's theorem). The more interesting observation, however, is that for the sun $M_{\text{sun}} = 1.98 \cdot 10^{30}\,\text{kg}$ and $c^2\frac{4}{9}\frac{R_{\text{sun}}}{G} = 1.38 \cdot 10^{29}\,\text{kg}$ such that $M_{\text{sun}} \gg \frac{4}{9}\frac{R_{\text{sun}}}{G}$. The fact that the sun has not yet collapsed needs therefore an explanation. The analysis was based on a perfect fluid model where particles do not interact, but obviously stellar interiors are undergoing nuclear interactions among others which generate an outward pressure. Taking this into account, the gravitational pressure does not have to be as large in order to balance the gravitational attraction.

When the star's nuclear fuel burns out it will begin to collapse. At a certain point in the collapse the electron-degeneracy pressure due to the Pauli exclusion principle can become large enough to halt collapse and the star turns into a *white dwarf*. Chandrasekhar found that for $M > 1.4\,M_{\text{sun}}$ even the electron degeneracy pressure cannot halt collapse. Eventually electrons and protons fuse to create neutrons and the neutron degeneracy pressure can halt collapse leaving a *neutron star* of radius in the order of 10 km. However if the Oppenheimer-Volkoff limit of $3 \sim 4\,M_{\text{sun}}$ is exceeded, even the neutron degeneracy pressure cannot halt collapse and the result is a *black hole*.

7.4 Schwarzschild Black Holes

The Schwarzschild metric (6.6) is a spherically symmetric solution to $R_{\mu\nu} = 0$. Two interesting values for r are $r = 2GM$ and $r = 0$. For normal astrophysical objects like stars and planets, but also for any spherical symmetric object such as a basketball both values are inside of the object where an interior solution to Einstein's equations with $T_{\mu\nu} \neq 0$ are needed. However objects of mass $3 \sim 4\,M_{\text{sun}}$ will eventually collapse to form a black hole. In this case $r = 2GM$ is well outside of the interior, and in fact one can approach $r \to 0$ as close as one likes and is still outside of the source.

Black holes sometimes get a bad wrap, but it is important to recall that the Schwarzschild solution describes the exterior geometry of any spherically symmetric source. Black holes do not suck any harder than comparable mass stars as long as one stays outside. Of course falling into a black hole is problematic, but so does falling into a star. What makes black holes so interesting is that one does not hit the interior until $r = 0$, yet the geometry one encounters along the way does some really interesting things.

The escape velocity in Newtonian gravity can be calculated from the kinetic energy needed to escape to infinity from a gravitating body. With

$$E_{\text{tot}} = \frac{1}{2} m v_{\text{escape}}^2 - \frac{GMm}{r} = 0 \qquad\qquad v_{\text{escape}} = \sqrt{\frac{2GM}{R}}$$

one barely escapes with $v \to 0$ as $r \to \infty$. Note that if $R = 2GM$ then the escape velocity is one which means speed of light, so no massive object can completely escape to $r = \infty$. However this is not a black hole yet because of two reasons. On one side is the assumption that the object is launched and that is the only energy given, but there may be a thruster such that the object may still escape. On the other side even though one cannot escape, one can still move away from $r = 0$ while for a black hole there is only one direction towards $r = 0$ once one is inside of $R = 2GM$.

In General Relativity clearly something interesting happens to ds^2 when $r = 2GM$. It looks singular but the question is whether this means that something in the geometry is becoming singular at $r = 2GM$. The answer is no because the situation is like in the space $\mathbb{R}^2$ with polar coordinates where $g^{\theta\theta} = r^{-2}$ also looks singular but there is no singularity in $\mathbb{R}^2$. At $r = 0$ certainly something bad happens and General Relativity breaks down, but studying things for $r > 0$ makes black holes so interesting.

To systematically explore the range $0 < r < 2GM$ one should remember that the metric is coordinate dependent and that one should look for invariant statements about the geometry. Here the Ricci tensor and therefore also the Ricci scalar are zero such that one has to go back to the Riemann curvature tensor. If nothing strange happens to the curvature at $r = 2GM$ one can look for more appropriate coordinates. Similar to $r = 0$ in $\mathbb{R}^2$ with polar coordinates, $r = 2GM$ may turn out to be a *coordinate singularity* where the metric in this set of coordinates behaves strangely but where nothing singular happens with the curvature.

In the Schwarzschild metric one finds that $R^{\mu\nu\rho\sigma} R_{\mu\nu\rho\sigma} = \frac{48G^2M^2}{r^2}$ and all other invariants are finite. This value gets infinite for $r \to 0$ as expected, but it is finite for $r = 2GM$. Thus while $r = 0$ represents a true curvature singularity, $r = 2GM$ is only a coordinate singularity (albeit a very interesting one). Thus there may be better coordinates.

Going from the Schwarzschild coordinates $\{t, r, \theta, \phi\}$ to the so-called Eddington-Finkelstein coordinates $\{v, r, \theta, \phi\}$ where $v = t + r + 2GM \ln\left|\frac{r}{2GM} - 1\right|$ or $t = v - r - 2GM \ln\left|\frac{r}{2GM} - 1\right|$ changes the Schwarzschild metric (6.6) to

$$ds^2 = -\left(1 - \frac{2GM}{r}\right) dv^2 + 2\,dv\,dr + r^2\,d\Omega^2$$

such that $r = 2GM$ is no longer problematic while $r = 0$ still is, but this was to be expected since it is a true curvature singularity. The Schwarzschild coordinates are useful for $r > 2GM$ and for $0 < r < 2GM$ and the Eddington-Finkelstein coordinates are useful everywhere except at $r = 0$. In particular they are more reliable for describing things as they pass through $r = 2GM$.

One can use the Schwarzschild metric in Eddington-Finkelstein coordinates to figure out what happens at $r = 2GM$. To understand black hole geometries one can examine the causal structure using light cones. Light travels along paths with $ds^2 = 0$. Here one can concentrate on purely radial trajectories with $d\theta = d\phi = 0$. These trajectories are

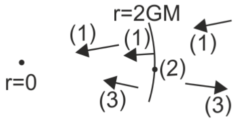
$$0 = -\left(1 - \frac{2GM}{r}\right) dt^2 + \left(1 - \frac{2GM}{r}\right)^{-1} dr^2 \qquad\qquad 0 = -\left(1 - \frac{2GM}{r}\right) dv^2 + 2\,dv\,dr$$

on the left side in Schwarzschild coordinates and on the right side in Eddington-Finkelstein coordinates. To get a sense of what is going on one can draw the behavior or the light cones in each case on a spacetime diagram.

In the case of the Schwarzschild coordinates these trajectories become

$$\frac{dr}{dt} = \pm\left(1 - \frac{2GM}{r}\right)^{-1} \rightarrow \begin{cases} \pm 1 & \text{as } r \rightarrow \infty \\ 0 & \text{as } r \rightarrow 2GM \end{cases}$$

and that shows far from $2GM$ a propagation $\frac{dr}{dt}$ at the speed of light which slows down the closer it is until it comes to rest at $2GM$.



In the case of the Eddington-Finkelstein coordinates there are several possible null trajectories:

(1) $dv = 0 \Rightarrow v = \text{constant} = t + r + 2GM \ln\left|\frac{r}{2GM} - 1\right|$: As t increases r must decrease so these trajectories are ingoing (for $r \lesseqgtr 2GM$).
(2) $dr = 0$ and $r = 2GM$: These trajectories are radially stationary and therefore neither ingoing nor outgoing.
(3) $\frac{dr}{dv} = \frac{1}{2}\left(1 - \frac{2GM}{r}\right)$:

$$\frac{dr}{dv} = \frac{1}{2}\left(1 - \frac{2GM}{r}\right) = \begin{cases} > 0 & r > 2GM \text{ (outgoing)} \\ < 0 & r < 2GM \text{ (ingoing)} \end{cases}$$

For $r < 2GM$ there are only ingoing trajectories, and at $r = 2GM$ there are stationary and ingoing, but no outgoing trajectories.

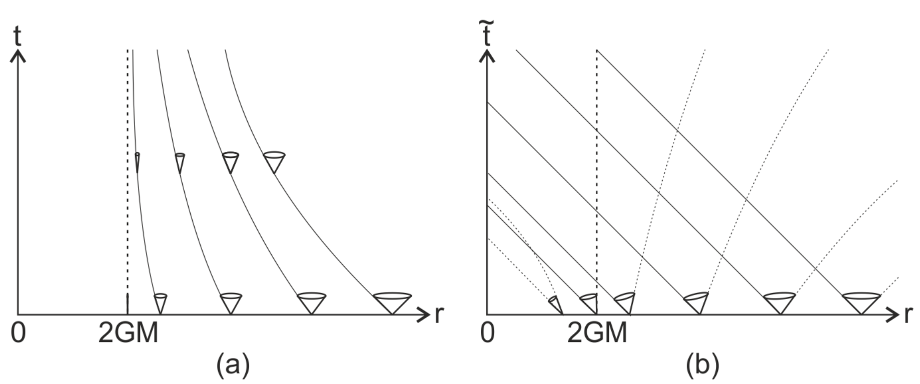
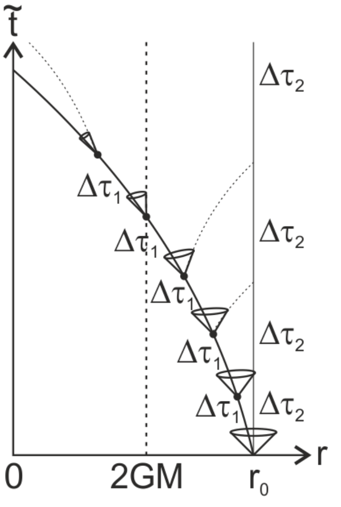


Figure 7.1: The light cones in the Schwarzschild metric

The spacetime diagrams in figure 7.1 show in (a) the light cones for the Schwarzschild coordinates and in (b) for the Eddington-Finkelstein coordinates but with a Schwarzschild compatible modified time coordinate $\tilde{t} = v - r$ for which $\tilde{t} \to t$ as $r \to \infty$. The light cones are for Schwarzschild coordinates at the usual 45° far from the spherically symmetric massive object as in flat space, but close-up when approaching $r = 2GM$. This closing up of the light cones which makes them collapse at $r = 2GM$ as shown in (a) is a visual signal of the problem with these coordinates. The light cones for the Eddington-Finkelstein coordinates modified with the $\tilde{t}$ time coordinate only tip more and more over when approaching $r = 0$ but pass $r = 2GM$ without problem as illustrated in (b). The left side keeps the 45° angle but the right side becomes 90° at $r = 2GM$ and is greater than 90° beyond this point. Thus even light can only go inward beyond $r = 2GM$ which is called the *event horizon.*

An observer stays at a fixed r assumed to be far away from the event horizon $r = 2GM$ and another observer follows a geodesic into the event horizon. The observer falling in sends at equal proper time intervals $\Delta\tau_1$ a light signal. The observer fixed at r_0 sees longer and longer intervals $\Delta\tau_2$ between these light signals. This illustrates that while the infalling observer reaches and passes the event horizon by his own internal clock, to the outside observer the infalling observer never seems to pass through the event horizon. The infalling observer emits light, and the outside observer sees this light more and more red-shifted as the infalling observer slows down and never passes the event horizon. Thus from the perspective of the outside observer the infalling observer red-shifts away.

This is a more complicated version of what is known from Special Relativity. If a moving observer sends out light signals at a rate of $\Delta\tau$, then to a "fixed" observer the signals arrive at a slower rate due to time-dilatation. Here though the effect is arising due to both the geometry and relative motion.



7.5 Maximally Extended Geometries and Wormholes

Consider the geometry defined by the line element with constant b

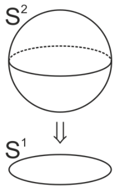
$$ds^2 = -dt^2 + dr^2 + (r^2 + b^2)\left(d\theta^2 + \sin(\theta)^2\, d\phi^2\right)$$

with $t \in (-\infty, \infty)$, $r \in [0, \infty)$, $\theta \in [0, \pi]$, $\phi \in [0, 2\pi)$. With $b = 0$ this is Minkowski space in spherical polar coordinates. This geometry has the properties:

- There is no t in the line element making it stationary, and there is no cross term $dt\, dx^i$ making it static.
- For $r \to \infty$ the geometry becomes Minkowski space $\mathbb{M}^4$ because b as a constant becomes more and more irrelevant.
- The geometry is S^2-foliated meaning that for fixed r and t one sweeps out an S^2 sphere by varying θ and ϕ.

Because the geometry is static, one can freeze t without loosing much and consider the remaining part $ds^2 = dr^2 + (r^2 + b^2)(d\theta^2 + \sin(\theta)^2\, d\phi^2)$ alone which gives a three-dimensional spatial geometry. The human being is severely handicapped in visualizing three-dimensional geometries. Living in a three-dimensional space makes it difficult to see other three-dimensional geometries.

The trick is that one collapses the three-dimensional geometry to two dimensions and embeds the two-dimensional geometry in the three-dimensional space $\mathbb{R}^3$. Doing so helps to visualize the curvature of the two-dimensional surface as bending in three dimensions. (Note that this is just a trick for visualizing curvature, but the curved space does not need an embedding in a flat space.) To go from three to two dimensions one can fix $\theta = \frac{\pi}{2}$ such that the line element becomes $ds^2 = dr^2 + (r^2 + b^2)\, d\phi^2$ and $\mathrm{S}^2 \Rightarrow \mathrm{S}^1$.



This two-dimensional metric is in a kind of polar type coordinates and looks circularly symmetric because it only depends on r. To embed it into three dimensions the simplest way is to extend the two-dimensional circular geometry into three dimensions using cylindrical coordinates $\{z, \rho, \psi\}$ in $\mathbb{R}^3$ with $z \in (-\infty, \infty)$, $\rho \in [0, \infty)$, $\psi \in [0, 2\pi)$. To do so the three functions $z(r, \phi)$, $\rho(r, \phi)$, $\psi(r, \phi)$ are needed.

By aligning the cylindrical coordinates conveniently, one can set $\psi(r,\phi)=\phi$. To figure out $z(r,\phi)$ and $\rho(r,\phi)$ the line element can be written as

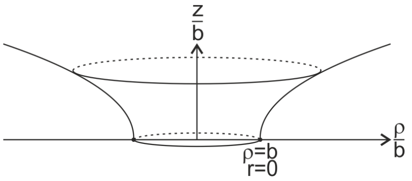
$$ds^2 = dz^2 + d\rho^2 + \rho^2\, d\psi^2 = \left(\frac{\partial z}{\partial r}\right)^2 dr^2 + \left(\frac{\partial \rho}{\partial r}\right)^2 dr^2 + \rho^2\, d\phi^2 = \left[\left(\frac{\partial z}{\partial r}\right)^2 + \left(\frac{\partial \rho}{\partial r}\right)^2\right] dr^2 + \rho^2\, d\phi^2$$

and comparison with $dr^2 + (r^2+b^2)\, d\phi^2$ shows that $(\frac{\partial z}{\partial r})^2 + (\frac{\partial \rho}{\partial r})^2 = 1$ and $\rho^2 = r^2 + b^2$. This finally leads to

$$\frac{d\rho}{dr} = \frac{r}{\sqrt{r^2+b^2}} \qquad \left(\frac{dz}{dr}\right)^2 + \frac{r^2}{r^2+b^2} = 1 \qquad \Rightarrow \qquad \begin{aligned} z(r) &= b\,\sinh^{-1}\left(\frac{r}{b}\right) \\ z(\rho) &= b\,\sinh^{-1}\left(\sqrt{\frac{\rho^2}{b^2}-1}\right) \end{aligned}$$

where $z(0)=0$ has been chosen. Note that for $r>0$ follows $z>0$ because $r = b\,\sinh\left(\frac{z}{b}\right)$.

This space is geodesically incomplete. This means if one picks a point and an initial direction, then the solutions to the geodesic equation for this geometry in some cases will terminate after a finite path length. These cases include any path that moves towards $r=0$. (Note that this does not happen in flat space in polar coordinates since paths can move through $r=0$ continuously.



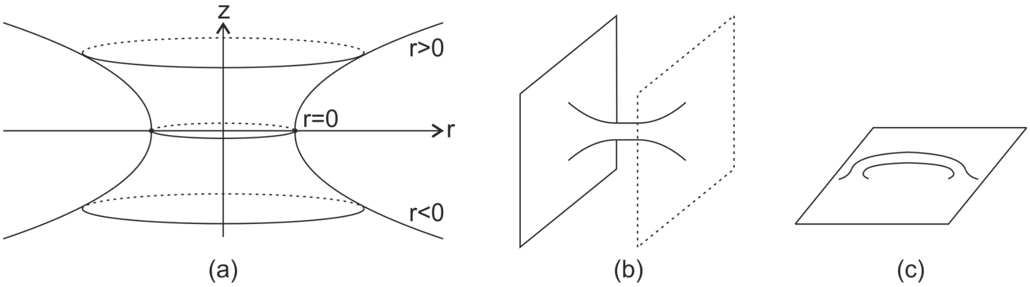


Figure 7.2: Wormhole metrics

A metric such as this one in which geodesics end is called *geodesically incomplete.* One can maximally extend this geometry to make it geodesically complete by allowing r to run to negative numbers by extending it to $r \in (-\infty,\infty)$ instead of $r \in [0,\infty)$. From the relation between r and z follows that $r<0$ corresponds to $z<0$. This leads to a so-called wormhole metric as shown in figure 7.2. In (a) the extension to negative values of r and z leads to a geodesically complete metric which asymptotically goes to $\mathbb{M}^4$ for $r\to\infty$ but also for $r\to-\infty$. To go from the $\mathbb{M}^4$ corresponding to $r\to\infty$ to the $\mathbb{M}^4$ corresponding to $r\to-\infty$ one has to go through $r=0$. This is an example of a wormhole which leads to two asymptotic distinct spaces connected by a tube or doorway as in (b) and not the kind of wormholes in science fiction with a shortcut as illustrated in (c).

These are smooth geometries which therefore have no infinite curvature and one might ask whether they can exist. Unfortunately to solve $G_{\mu\nu} = 8\pi G T_{\mu\nu}$ these geometries require $\rho<0$ which only comes from vacuum energy. The universe does have vacuum energy but it is uniform on large scales, and these solutions are obviously not. There is non-uniform vacuum energy from quantum fluctuations, but this might only lead to a quantum sized wormhole.

7.6 Schwarzschild Black Holes in Kruskal Coordinates

Given the Schwarzschild geometry $ds^2 = -(1-\frac{2GM}{r})dt^2 + (1-\frac{2GM}{r})^{-1}dr^2 + r^2 d\Omega^2$ in Schwarzschild coordinates with $r>2GM$ one can explore the geodesic completeness with the new set of coordinates

$$T_I = \left(\frac{r}{2GM}-1\right)^{\frac{1}{2}} e^{\frac{r}{4GM}} \sinh\left(\frac{t}{4GM}\right) \qquad R_I = \left(\frac{r}{2GM}-1\right)^{\frac{1}{2}} e^{\frac{r}{4GM}} \cosh\left(\frac{t}{4GM}\right) \tag{7.1}$$

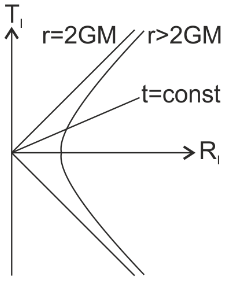
called *Kruskal coordinates.* Since $t \in (-\infty, \infty)$ and $r > 2GM$, the new coordinates are defined for $T_I \in (-\infty, \infty)$ and $R_I \in (0, \infty)$ and

$$ds^2 = \frac{32G^3M^3}{r}\, e^{-\frac{r}{2GM}}\, (-dT_I^2 + dR_I^2) + r^2\, d\Omega^2$$

defines the geometry. In this form with r and not only R_I it is easier to see what happens to the event horizon.

There are three important aspects of Kruskal coordinates:

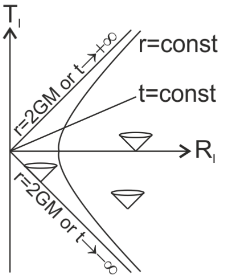
- Lines of constant r satisfy $-T_I^2 + R_I^2 = (1 - \frac{r}{2GM})\, e^{\frac{r}{2GM}}$ and are therefore half hyperboles because the right-hand side of the equation is negative for $r > 2GM$. In particular for $r = 2GM$ this becomes $T_I = \pm R_I$.
- Lines of constant t satisfy $T_I = R_I \tanh\bigl(\frac{t}{4GM}\bigr)$ and are straight lines radiating from the origin because $\frac{dT_I}{dR_I} = \tanh\bigl(\frac{t}{4GM}\bigr) =$ constant. In particular $t \to \pm\infty$ becomes $T_I \to \pm R_I$.
- Light cones open at 45° everywhere on a plot with the axes T_I and R_I since $ds^2 = 0 = (-dT_I^2 + dR_I^2)$ for radial motion such that $\frac{dT_I}{dR_I} = \pm 1$.



The geodesics can start at $r > 2GM$ and go below $r = 2GM$. Therefore the Kruscal coordinates can be extended

$$T_{II} = \left(1 - \frac{r}{2GM}\right)^{\frac{1}{2}} e^{\frac{r}{4GM}} \cosh\left(\frac{t}{4GM}\right)$$
$$R_{II} = \left(1 - \frac{r}{2GM}\right)^{\frac{1}{2}} e^{\frac{r}{4GM}} \sinh\left(\frac{t}{4GM}\right)$$

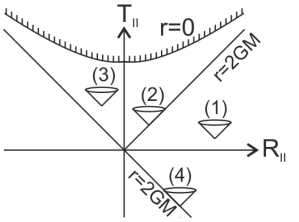
to $r < 2GM$. For $t \in (-\infty, \infty)$ and $r \in (0, 2GM)$ one gets $T_{II} \in (-\infty, \infty)$ and $R_{II} \in (-\infty, \infty)$ and

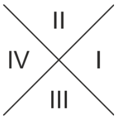


$$ds^2 = \frac{32G^3M^3}{r}\, e^{-\frac{r}{2GM}}\, (-dT_{II}^2 + dR_{II}^2) + r^2\, d\Omega^2$$

defines this geometry which is the same as for T_I and R_I. Constant r gives $T_{II}^2 - R_{II}^2 = (\frac{r}{2GM} - 1)\, e^{\frac{r}{2GM}}$ and in particular $T_{II}^2 - R_{II}^2 = 1$ for the singularity $r = 0$. Constant t means $T_{II} = R_{II} \coth\bigl(\frac{t}{4GM}\bigr)$ with $T_{II} \to \pm R_{II}$ for $t \to \pm\infty$. The light cones are obviously still open at 45° because the metric has not changed.

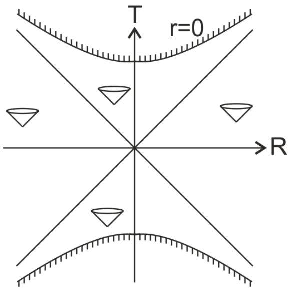
Outside of $r = 2GM$ as (1) it is possible to either remain outside for ever or to wander past $2GM$. Once at $r = 2GM$ as (2) there is no escape and one ends inside as (3) where one can only travel towards the singularity $r = 0$. There are geodesics which terminate at finite time in the future since they hit the singularity at $r = 0$. There is nothing one can or should do about them. In fact the points of termination of geodesics in more general contexts is used to define singularities.



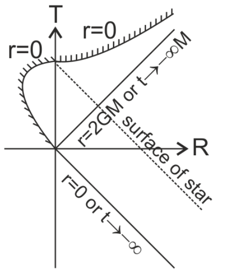


There are also geodesics such as those marked with (4) which, when traced backwards in time, terminate in a finite proper time. The geometry can therefore be still geodesically extended to region III with $T_{III} = -T_{II}$ and $R_{III} = -R_{II}$ and further to region IV with $T_{IV} = -T_I$ and $R_{VI} = -R_I$.

In both new regions ds^2 is the same and the light cones are therefore still open at 45°. Moreover there is now a singularity in the past with a so-called *white hole* out of which things can only escape but never enter. The backwards extended geodesics terminate on this singularity which is acceptable. The maximally extended Schwarzschild geometry with all four regions shows another example of a wormhole geometry. However this wormhole is non-traversable because getting in between the right and the left side means ending up at $r = 0$. Coming out of the while hole one has a choice of where to go but it is a most irrevocable decision.



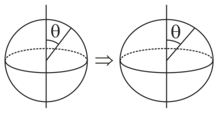
This picture of a black hole relies on $t \in (-\infty, +\infty)$, but there are two problems with the idea of an eternal black hole. On one hand has the universe a finite age, and on the other hand are astrophysical black holes born from stellar collapse. Once the surface hits $r = 0$, there is the usual black hole geometry. Before that, the details will be governed by the particulars of the interior solution.



The Kruskal diagram with $t \in (-\infty, +\infty)$ is a solution of General Relativity, and physicists spent and still spend a lot of time exploring these solutions despite the fact that they are not realistic in our universe. However one can amend the Kruskal diagram for the more realistic scenario of black holes emerging from stellar collapse. The bottom part and the left side disappear such that there is no white hole anymore. On the left side of the T-axis the singularity is growing in mass, but not in size. The final end point of the collapse is where $r = 0$ crosses the T-axis and where the surface of the star becomes zero. The interior is everything below the surface of the star. The horizon grows from $r = 0$ to $r = 2GM$ starting before the entire star is below $2GM$.

7.7 Rotating Black Holes, the Kerr Geometry and Penrose Diagrams

One can start with a spherically symmetric object and then let it spin by giving it non-zero angular momentum L about an axis through its center. The object induces a Schwarzschild metric (6.6) before it starts spinning and induces a different geometry when it is spinning. This geometry has first been described by Kerr. A difference to foresee is that spinning at a constant rate will still be time-independent and therefore stationary but no longer static such that cross-terms $dt\, dx^i$ are to be expected.



If the spin axis is aligned with the poles, then one should expect a dependence on θ from the squashing of the sphere. However there is still no dependence on ϕ. Note also that this is no translation invariance along the axis aligned with the poles, and it is therefore not cylindrical.

In 1963 Kerr found the solution for what is called the *Kerr metric*

$$\begin{aligned} ds^2 = &-\left(1 - \frac{2GMr}{\rho^2}\right) dt^2 - \frac{2GMar\sin(\theta)^2}{\rho^2}\, 2\, d\phi\, dt + \frac{\rho^2}{\Delta}\, dr^2 + \rho^2\, d\theta^2 \\ &+ \frac{\sin(\theta)^2}{\rho^2}\left((r^2+a^2)^2 - a^2\Delta\sin(\theta)^2\right) d\phi^2 \end{aligned} \tag{7.2}$$

in *Boyer-Lindquist coordinates* $\{t, r, \theta, \phi\}$ where $a = \frac{L_\phi}{M}$, $\Delta(r) = r^2 - 2GMr + a^2$, $\rho^2(r,\theta) = r^2 + a^2\cos(\theta)^2$. Some comments:

- For $a \to 0$ the Kerr metric becomes the Schwarzschild metric and the Boyer-Lindquist coordinates become the Schwarzschild coordinates.
- For $r \to \infty$ with M and a fixed the metric becomes Minkowski spacetime such that asymptotically it becomes flat spacetime.
- For $M \to 0$ with a fixed such that angular momentum L_ϕ also goes to zero, the geometry becomes Minkowski space in oblate spheroidal coordinates. In details this means that

  $$ds^2 = -dt^2 + \frac{r^2 + a^2\cos(\theta)^2}{r^2 + a^2}\, dr^2 + (r^2 + a^2\cos(\theta)^2)\, d\theta^2 + (r^2 + a^2\sin(\theta)^2)\, d\phi^2$$

  which is just $ds^2 = -dt^2 + dx^2 + dy^2 + dz^2$ with $x = \sqrt{r^2+a^2}\sin(\theta)\cos(\phi)$, $x = \sqrt{r^2+a^2}\sin(\theta)\sin(\phi)$ and $z = r\,\cos(\theta)$.
- The cross-term $d\phi\, dt$ is expected since the rotation is in ϕ.

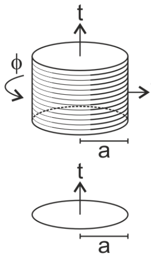
Similar to $r = 2GM$ and $r = 0$ in the Schwarzschild metric with Schwarzschild coordinates where $r = 2GM$ turned out to be a coordinate singularity while $r = 0$ is a real singularity, the values $\rho^2 = 0$ and $\Delta = 0$ in the denominator need careful analysis.

The first singularity $\rho = 0$ is a time curvature singularity where the curvature scalar diverges. For $\rho^2 = r^2 + a^2\cos(\theta)^2$ to become zero, both $r = 0$ and $\theta = \frac{\pi}{2}$ are required. This is in contrast with the Schwarzschild metric where $r = 0$ makes the geometry diverge for any θ and ϕ. Before looking at $\rho = 0$

it makes sense to analyze $r = 0$ first alone because it turns out to be rather surprisingly non-trivial. The Kerr metric becomes

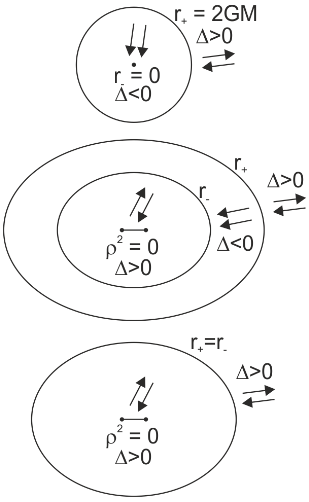
$$ds^2|_{r=0} = -dt^2 + (a\cos(\theta)\,d\theta)^2 + (a\sin(\theta))^2\,d\phi^2 = d\tilde{r}^2 + \tilde{r}^2\,d\phi^2$$

with $\tilde{r} = a\sin(\theta)$ and $\theta \in [0,\pi]$ such that $\tilde{r} \in [0,a]$.



So $r = 0$ is actually a volume in $2+1$ dimensions which is cylindrical in $\{t, \tilde{r}, \phi\}$ while in the Schwarzschild metric $ds^2|_{r=0} = -dt^2$ is just a line along t. Going back to $\rho = 0$ adds the condition $\theta = \frac{\pi}{2}$ which is at $\tilde{r} = a$. This is the surface of the cylinder such that the singularity $\rho = 0$ is spatially a ring which is then extended in time. Note that the region inside of the ring is non-singular. This is surprising but it should have been expected because there is an additional piece of information. The Schwarzschild geometry has the mass M as the only one parameter which defines the geometry, and $r = 0$ describes therefore a single point in space. The Kerr geometry has the angular momentum L_ϕ in addition to the mass M as parameter such that there is an additional piece of information. The fact that there is spinning is reflected in the singularity $\rho = 0$ which becomes a circle with a radius determined by the angular momentum. The deformation of the point to a ring is analogous to the deformation of the sphere to the spinning oblate spheroid.

The second singularity $\Delta = 0$ is not a curvature singularity but a coordinate singularity. Similarly to the Schwarzschild case it indicates the presence of the horizon. However $\Delta(r) = r^2 - 2GMr + a^2 = 0$ as a quadratic equation leads to $r_\pm = GM \pm \sqrt{G^2M^2 - a^2}$ corresponding to two horizons.



The sign of Δ determines the sign in front of dr^2 and determines therefore the behavior of r. For $\Delta > 0$, r is spatial and can increase or decrease, but if $\Delta < 0$, r is timelike so only moves in one direction because in physics one can only move into one direction in time. (In the Schwarzschild case r behaves spacelike outside of the horizon and timelike inside while t behaves timelike outside and spacelike inside.) Varying a shows:

- For $a = 0$ the two solutions are $r_- = 0$ and $r_+ = 2GM$ and correspond to Schwarzschild singularity and Schwarzschild horizon, respectively.
- For $a^2 < G^2M^2$ the geometry is called sub-extremal and $r_+ > r_- > 0$. The singularity marked $\rho^2 = 0$ is now a ring shown in the figure as a straight line in the center.
- For $a^2 = G^2M^2$ the black hole with its geometry is called extremal and $r_+ = r_- = GM$.
- For $a^2 > G^2M^2$ the geometry is called over-extreme and has no horizon. This is just a naked singularity. It will turn out that one cannot make a that big.

One possibility is to maximally extend the Kerr geometry with adopted Kruskal-type coordinates, but *conformal diagrams* also called *Penrose diagrams* are a more powerful tool. For Minkowski spacetime $\mathbb{M}^4$

$$ds^2 = -dt^2 + dr^2 + r^2\,d\Omega^2$$

with $t \in (-\infty, +\infty)$ and $r \in [0,\infty)$ one can use coordinates

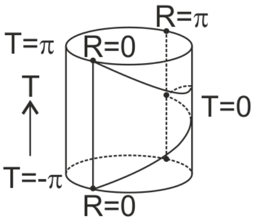
$$T = \tan^{-1}(t+r) + \tan^{-1}(t-r) \qquad\qquad R = \tan^{-1}(t+r) - \tan^{-1}(t-r)$$

where $0 \leq R < \pi$ and $|T| < \pi - R$ such that both coordinates have finite ranges. Then in

$$ds^2 = \frac{1}{(\cos(T)+\cos(R))^2}\left(-dT^2 + dR^2 + \sin(R)^2\,d\Omega\right) \qquad\qquad ds^2 = \frac{1}{\omega(T,R)^2}\,d\tilde{s}^2$$

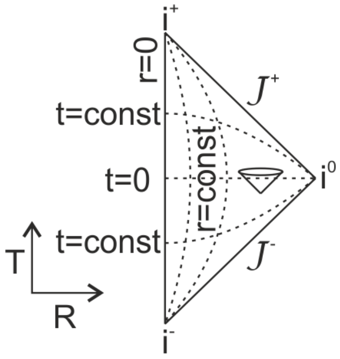
R is behaving like an angle as already hinted by the range $0 \leq R < \pi$. The fraction in front of $d\tilde{s}^2$ is called a conformal factor which is positive and multiplies the metric, and the metric $d\tilde{s}^2$ itself which is finite is called the conformally related geometry. Because when $ds^2 = 0$ also $d\tilde{s}^2 = 0$, light cones in the $ds^2 = 0$ geometry are also light cones in the $d\tilde{s}^2 = 0$ geometry. Thus studying light cones in the conformally related geometry gives the same information as one would get by studying them in the original geometry, but the conformally related geometry is finite in size and can therefore be drawn completely.

To visualize it, θ and ϕ are suppressed and T and R can be drawn as a cylinder. Because of $|T| < \pi - R$ as its range, T can vary between $-\pi$ and $+\pi$ for $R = 0$ but can only take the value 0 for $R = \pi$. The whole geometry is between the two curves which can be flattened out. The metric is



$$ds^2 = -dT^2 + dR^2$$

and the space is flat Minkowski spacetime.



Unwrapping gives the Penrose diagram for flat spacetime $\mathbb{M}^4$:

i^+ = timelike future ∞
i^- = timelike past ∞
$\Rightarrow$ All timelike geodesics ($m > 0$) begin at i^- and end at i^+.
i^0 = spacelike ∞
$\Rightarrow$ All spacelike geodesics end at i^0.
$\mathcal{J}^+$ = future null ∞
$\mathcal{J}^-$ = past null ∞
$\Rightarrow$ All lightlike geodesics ($m = 0$) begin at $\mathcal{J}^-$ and end at $\mathcal{J}^+$.
The coordinates t and r are the coordinates in flat spacetime $\mathbb{M}^4$ where the spatial part with r is in polar coordinates.

The Penrose diagram for a Schwarzschild black hole is shown in figure 7.3. The right half of (a) looks on the right side the same as the one for Minkowski spacetime. However $r = 0$ is no longer a vertical line but a horizontal line such that there is no escape after crossing $r = 2GM$. This is called a timelike singularity because r is timelike behind $r = 2GM$ such that $-dr^2$, and this is why one can only move in one direction. Curves r constant connect i^- and i^+, and curves t constant connect i^0 and the unlabeled point on the left side. This diagram is not geodesically complete because there are geodesics whose origin are unclear, and figure (b) illustrates the maximally complete diagram where the part in (a) is called M and the part $\tilde{M}$ has been added together with the black and the white hole.

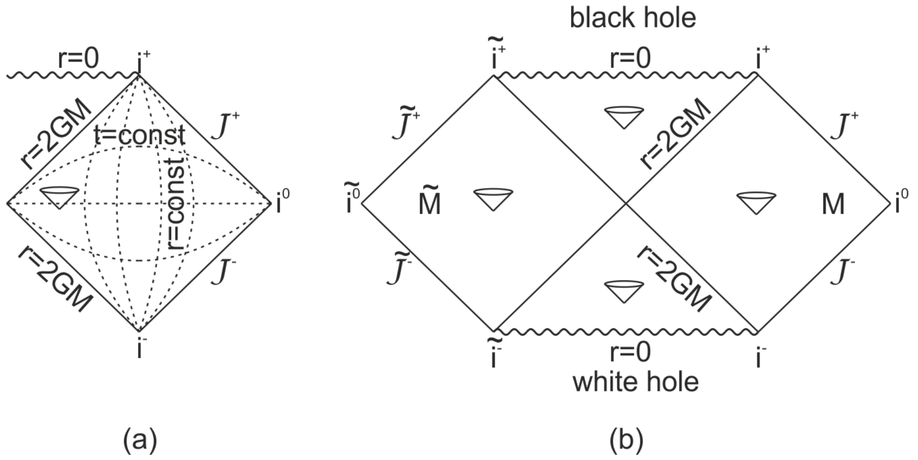


Figure 7.3: Penrose diagram for Schwarzschild black holes

From the white hole one has the choice whether one wants to go to the asymptotic universe M or to the other asymptotic universe $\tilde{M}$, but once in either of them there is no way back to the white hole nor to the other asymptotic universe. From both asymptotic universes there is a way to the black hole but once one has passed the horizon, there is no way back.

The Penrose diagram for the Kerr geometry corresponding to rotating black holes looks again different. If the black hole starts spinning but with $a^2 < G^2M^2$ there appear two horizons with r_- and r_+ where $r = 0$ becomes the horizon r_-, and the singularity itself becomes a ring. What is really interesting is the fact that outside r_+ and inside r_- one can move towards the singularity or away from it. However between r_- and r_+ one can only move toward the horizon r_-. When the black hole starts rotating faster the inner horizon grows until it becomes the same as the outer horizon a $a^2 = G^2M^2$. The middle region between the two horizons gets lost.

For the Kerr geometry the maximally extended solution turns therefore out to be more complicated than the Schwarzschild case because on one hand it looks, not too surprisingly, different for different choices of a^2 compared to G^2M^2 and on the other hand it looks, more surprisingly, different for different values of the coordinate θ. The reason for the dependence on θ is that to be a singularity $r = 0$ is not sufficient but also $\theta = \pi$ is necessary.

The case $a^2 = 0$ is the Schwarzschild case and is shown in figure 7.3. The over-extreme case $a^2 > G^2M^2$ with the naked singularity does not exist and there is no reason to study this case. Thus, the sub-extremal case $0 < a^2 < G^2M^2$ with the two possibilities $\theta = \frac{\pi}{2}$ and $\theta \neq \frac{\pi}{2}$ and the extremal case $a^2 = G^2M^2$ have to be analyzed.

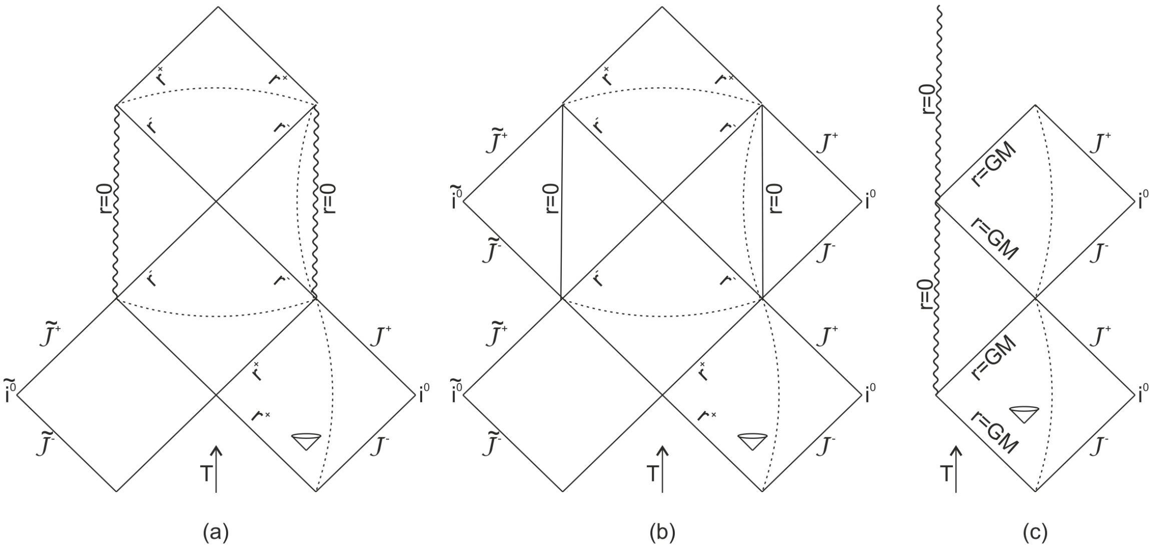


Figure 7.4: Penrose diagram for rotating Kerr black holes

In figure 7.4 the Penrose diagrams for the different cases are shown. Only the lines of constant r have been drawn as dashed lines, but not the lines of constant t which are perpendicular to them as in the case of the Schwarzschild black holes. Light cones are everywhere at 45°, and in all three cases one starts in the area where the light cone has been placed.

In the sub-extremal case $a^2 < G^2M^2$ with $\theta = \frac{\pi}{2}$ illustrated in (a) there is because ρ^2 becomes zero for $r = 0$ and $\theta = \frac{\pi}{2}$ a true singularity at $r = 0$, but this singularity is spacelike with $+dr^2$ and one can avoid it. If one is between the two horizons one can only proceed in one direction, but this is different for the direction one entered the zone between the two horizons. Coming from outside through the horizon r^+ one can only go through the horizon r^- to the inside, and similarly coming from inside through the horizon r^- one can only go through the horizon r^+ to the outside.

In the sub-extremal case with $\theta \neq \frac{\pi}{2}$ presented in (b) there is no real singularity at $r = 0$ because ρ^2 does not get zero. Also in this case one does therefore not have to go there, but if one does and passes the ring singularity one enters a different asymptotic spacetime. In both sub-extremal cases the entire solution is an infinitely long repeating pattern of these primitive regions.

In the extremal case with $a^2 = G^2M^2$ the two horizons become the same and one can go inside and exit the horizon, but one enters a different asymptotic spacetime. Thus one cannot throw something like a boomerang in and get it back out. One can pass a boundary $r = GM$ only in one direction.

Similarly to the Schwarzschild solution also the Kerr solution only exists in infinite time, but astrophysical black holes resulting from stellar collapses become existent at some point in time and have not existed forever. This is not a very realistic solution because an infinite number of asymptotic spacetimes are stitched together. However, when General Relativity predicts a geometry no matter how distinct it is from what is physically found in the universe, General Relativity is well behaved in this geometry.

The obvious problem of conservation of energy with several asymptotic spacetimes is not that dramatic. Observing something coming out of a horizon from the white hole, for example, looks like there comes energy in form of mass out of nowhere. One cannot see behind a horizon, however, the size of the horizon is dependent on M, and M is all the mass in the region behind the horizon. Thus one could observe the change of the horizon due to the loss of mass. Similarly if some mass passes the horizon of a black hole this just makes the black hole more massive. With different asymptotic universes the infinite number of them is all contained in the region behind a horizon. Thus throwing in some energy is never lost because all of these infinitely many universes are all behind this horizon, and throwing in some energy makes the horizon bigger forever. This infinite number of universes cannot be arranged as a circle because of causality. Time can never go back to a point in the past. Since these solutions of the Einstein equations are theoretical and have nothing to do with the real universe, all these thoughts are theoretical too.

There is a big difference between the way physicists work in General Relativity and in other areas. Usually they start from a set of differential equations and some initial conditions in order to find what happens in the future. Particles move and interact while time passes, and one can ask how the situation is after some time. In General Relativity it is different because the whole future is part of the solution. A geometry solving Einstein's equation is complete from the past to the future. Therefore one cannot take two black holes explored so far and ask what happens if they merge because merging them is not in their future. In astrophysics, of course, real black holes merge and doing so create gravitational waves.

7.8 Gravitational Waves

There are three essential parts to any kind of radiation but here formulated for gravitational radiation:

1. Find and solve a source free equation.
2. Detection: How do waves influence matter.
3. Creation: How does matter creates waves.

The first point needs a solution of the Einstein equations without sources, the second point means solving the geodesic equation in the geometry found within the first point, and the third point asks for solving Einstein's equations with source. The detection of gravitational radiation took a long time because it is typically very weak and matter must be very involved to create them. Gravitational waves have a small amplitude and one can make use of that with some approximation techniques in the first and second point.

For the first part linearized General Relativity similar to finding the Newtonian limit is used, but this cannot be not a static approximation because a static approximation does not make sense for waves. One starts with a flat geometry and some small fluctuations

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}(X) \qquad\qquad g^{\mu\nu} = \eta^{\mu\nu} - \eta^{\mu\alpha}\eta^{\nu\beta}h_{\alpha\beta}$$

where $h_{\mu\nu}(X)$ is therefore small, and one inserts this metric into the source free Einstein equations $R_{\mu\nu}$ in trace-reverse form. The Christoffel symbols are

$$\Gamma^{\rho}_{\alpha\beta} = \frac{1}{2}\eta^{\rho\sigma}\left(\partial_\alpha h_{\beta\sigma} + \partial_\beta h_{\sigma\alpha} - \partial_\sigma h_{\alpha\beta}\right)$$

since $\partial_\lambda \eta_{\mu\nu} = 0$ and $\eta^{\mu\alpha}\eta^{\nu\beta}h_{\alpha\beta}$ would be $O(h^2)$. The Riemann curvature tensor is

$$R^{\rho}{}_{\mu\sigma\nu} = \partial_\sigma \Gamma^{\rho}_{\nu\mu} - \partial_\nu \Gamma^{\rho}_{\sigma\mu} + \Gamma^{\rho}_{\sigma\lambda}\Gamma^{\lambda}_{\nu\mu} - \Gamma^{\rho}_{\nu\lambda}\Gamma^{\lambda}_{\sigma\mu}$$

where the last two terms can be ignored because they are $O(h^2)$. The Ricci tensor becomes

$$\begin{aligned} R_{\mu\nu} = R^{\rho}{}_{\mu\rho\nu} &= \partial_\rho \Gamma^{\rho}_{\nu\mu} - \partial_\nu \Gamma^{\rho}_{\rho\mu} \\ &= \frac{1}{2}\eta^{\rho\sigma}\left[\partial_\rho\partial_\nu h_{\mu\sigma} + \partial_\rho\partial_\mu h_{\sigma\nu} - \partial_\rho\partial_\sigma h_{\nu\mu} - \partial_\nu\partial_\rho h_{\mu\sigma} - \partial_\nu\partial_\mu h_{\sigma\rho} + \partial_\nu\partial_\sigma h_{\rho\mu}\right] \\ &= \frac{1}{2}\eta^{\rho\sigma}\left[\partial_\rho\partial_\mu h_{\sigma\nu} - \partial_\rho\partial_\sigma h_{\nu\mu} - \partial_\nu\partial_\mu h_{\sigma\rho} + \partial_\nu\partial_\sigma h_{\rho\mu}\right] \end{aligned}$$

and one can define $V_\mu = \partial_\rho h^{\rho}{}_{\mu} - \frac{1}{2}\partial_\mu h^{\rho}{}_{\rho}$ and its partial derivative $\partial_\nu V_\mu = \partial_\nu\partial_\rho h^{\rho}{}_{\mu} - \frac{1}{2}\partial_\nu\partial_\mu h^{\rho}{}_{\rho}$ where $h^{\rho}{}_{\rho}$ is the trace. Then with the definition $\Box \equiv \partial_\rho\partial^\rho$ which is $\partial_\rho\partial^\rho = -\frac{\partial^2}{\partial t^2} + \vec{\nabla}^2$ one gets

$$R_{\mu\nu} = \frac{1}{2}\left[-\Box h_{\mu\nu} + \partial_\mu V_\nu + \partial_\nu V_\mu\right] = 0$$

for the Ricci tensor, but one can do better. Gauge freedom allows despite the fact that one has to choose certain coordinates so that $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$, one can still make coordinate changes which preserve $\eta_{\mu\nu}$ but will generally change the form of $h_{\mu\nu}$.

If one transforms to $X^{\mu'} = X^\mu + \delta^\mu(x)$ with a small δ^μ the metric changes to

$$g_{\mu\nu} \to g_{\mu'\nu'} = \frac{\partial X^\mu}{\partial X^{\mu'}} \frac{\partial X^\nu}{\partial X^{\nu'}} g_{\mu\nu} \qquad\qquad g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu} \to g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu} - \partial_\mu\delta_\nu - \partial_\nu\delta_\mu$$

where $\delta_\mu = \eta_{\mu\nu}\delta^\nu$. Compare $h_{\mu\nu} \to h'_{\mu\nu} = h_{\mu\nu} - \partial_\mu\delta_\nu - \partial_\nu\delta_\mu$ with $A_\mu \to A'_\mu = A_\mu - \partial_\mu\Phi$ in electromagnetism. A useful aspect of gauge freedom is that the physical degrees of freedom do not change. In this case the physical curvature $R^\mu{}_{\nu\lambda\rho}$ is unchanged such that solutions to $R_{\mu\nu} = 0$ remain solutions. One can use here a gauge such that $V_\mu \to V'_\mu = \partial_\alpha h^\alpha{}_\mu - \frac{1}{2}\partial_\mu h^\alpha{}_\alpha = 0$. From $R_{\mu\nu} = -\frac{1}{2}\Box h_{\mu\nu} = 0$ follows $\Box h_{\mu\nu} = 0$.

One can immediately write down a plane wave solution

$$h_{\mu\nu} = a_{\mu\nu}\, e^{iK_\lambda X^\lambda}$$

where one can think of $a_{\mu\nu}$ as the polarization and the amplitude of the wave. Feeding this into $\Box h_{\mu\nu} = -K_\lambda K^\lambda h_{\mu\nu} = 0$ gives $K_\lambda K^\lambda = 0$ because $h_{\mu\nu} = 0$ is a solution but not a very interesting one. Thus K^λ is a null vector, and one can write it as $K^\lambda = (|\vec{k}|, \vec{k})$ with $K_\lambda K^\lambda = -|\vec{k}|^2 + \vec{k}\cdot\vec{k} = 0$. With the frequency $\omega = |\vec{k}|$ and the wavelength $\lambda = \frac{2\pi}{\omega}$ the wave travels with phase velocity $v = \lambda\frac{\omega}{2\pi} = 1$ in the direction $\vec{k}$. Thus, the gravitational wave travels with the speed of light as one would expect because the gravitational fluctuations are massless.

The symmetric 4×4 matrix $a_{\mu\nu}$ describes amplitude and polarization of the wave and can be simplified because using any δ_μ such that $\Box\delta_\mu = 0$ these four functions can be used to make any four components of $h_{\mu\nu}$ vanish identically. Choosing $h_{ti} = 0$ and $h^\mu{}_\mu = 0$ (traceless) or $a_{ti} = 0$ and $a^\mu{}_\mu = 0$ represents three plus one terms such that the ten degrees of freedom of the symmetric matrix $a_{\mu\nu}$ are reduced to six. With the previous gauge condition V_μ follows

$$\begin{aligned} V_t &= \partial_\rho h^\rho{}_t - \frac{1}{2}\partial_t h^\rho{}_\rho = \partial_t h^t{}_t = i\omega a_{tt}\, e^{iK_\lambda X^\lambda} = 0 &\qquad \Rightarrow &\qquad a_{tt} = 0 \\ V_i &= \partial_\rho h^\rho{}_i - \frac{1}{2}\partial_i h^\rho{}_\rho = \partial_j h^i{}_i = ik_j a_{ji}\, e^{iK_\lambda X^\lambda} = 0 &\qquad \Rightarrow &\qquad K^j a_{ji} = 0 \end{aligned}$$

where the equation $K^j a_{ji} = 0$ means that the waves are transverse or, in other words, that the spatial wave vector $\vec{k}$ is perpendicular to the polarization tensor.

If the wave vector is chosen to be $K^\mu = (\omega, 0, 0, \omega)$ with the spatial part $\vec{k} = (0, 0, \omega)$ then $a_{zi} = 0$ follows from the transversality condition. Thus only two independent components of $a_{\mu\nu}$ remain which are $a_{xx} = -a_{yy}$ (traceless) and $a_{xy} = a_{yx}$ (symmetric). This additional choice is called transverse-traceless gauge, and

$$h_{\mu\nu}(X) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & a & b & 0 \\ 0 & b & -a & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} e^{iK_\lambda X^\lambda} \tag{7.3}$$

is the final form of $h_{\mu\nu}(X)$. In the linearized theory on gets more general solutions by adding solutions of this form. This is, however, not possible in the original version of General Relativity which is highly non-linear. With the solution in (7.3) the part of finding a source free wave equation is done.

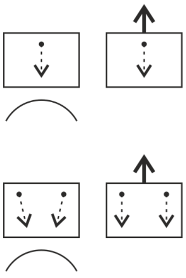
The next part is the detection of gravitational waves. With the metric solution $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}(X)$ where $\eta_{\mu\nu}$ is flat spacetime and $h_{\mu\nu}(X)$ is defined by (7.3) one can explore how test particles respond to this time-dependent geometry using the geodesic equation (5.9)

$$\frac{d^2X^\mu}{d\lambda^2} + \Gamma^\mu_{\alpha\beta}\frac{dX^\alpha}{d\lambda}\frac{dX^\beta}{d\lambda} = 0 \qquad \Rightarrow \qquad \frac{dU^\mu}{d\tau} + \Gamma^\mu_{\alpha\beta}\, U^\alpha\, U^\beta = 0 \qquad \Rightarrow \qquad \frac{dU^\mu}{d\tau} = -\Gamma^\mu_{\alpha\beta}\, U^\alpha\, U^\beta$$

for timelike and $U^\mu = \frac{dX^\mu}{d\tau}$. If the particle is at rest for $\tau = 0$ such that $U^\mu(0) = (1, 0, 0, 0)$ then

$$\frac{dU^\mu}{d\tau}(0) = -\Gamma^\mu_{00} = -\frac{1}{2}\eta^{\rho\sigma}(\partial_0 h_{\sigma 0} + \partial_0 h_{0\sigma} - \partial_\sigma h_{00}) = 0$$

since $h_{ti} = h_{tt} = 0$ and this means that if the particle begins at rest, it remains at rest as the wave passes. However, this is just a statement that the coordinate position of the mass is unchanged because the particle is at rest and the acceleration is zero. Thus maybe two test masses and the distance between them show an effect.

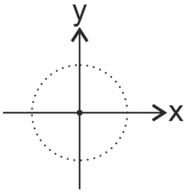


Actually, one could have anticipated the need for at least two particles from the equivalence principle. Observing only one particle does not allow to distinguish whether the lab is in the gravitational field of the earth or being accelerated. Only if one observes two particles one can detect tidal forces because the two particles come closer because they move towards the center of the earth while in the case of acceleration they move on parallel paths and keep the distance they initially had. Detecting curvature is impossible with only one test mass because one can always find coordinates in which the particle is and stays at rest. With two test masses one can see whether the distance between them remains constant or not.

For one mass at the spatial coordinate $(0, 0, 0)$ and another mass at spatial coordinate $(\varepsilon, 0, 0)$ one finds for the distance

$$\begin{aligned}\int \sqrt{ds^2} &= \int \sqrt{g_{\mu\nu} dX^\mu dX^\nu} = \int_0^\varepsilon \sqrt{g_{xx}} dx \\ &\approx \sqrt{g_{xx}(x=0)}\,\varepsilon = \sqrt{1 + h_{xx}(x=0)}\,\varepsilon \approx \left[1 + \frac{1}{2} h_{xx}(x=0)\right] \varepsilon = \left[1 + a e^{iK_\lambda X^\lambda}\right] \varepsilon\end{aligned}$$

and this varies with time. Despite the fact that one particle remains at $x = 0$ and the other at $x = \varepsilon$, the invariant distance between them changes because this value varies with time. This is the difference between the coordinates and the physical reality. In the chosen coordinates the two masses do not move, but in reality they move with respect to each other.



To get a better idea of what a gravitational wave does, the values a and b in (7.3) are chosen such that a is small and b is zero, and the wave travels along the z-axis. This gives

$$h_{\mu\nu}(X) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & a & 0 & 0 \\ 0 & 0 & -a & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \sin(kz - \omega t)$$

for the real part. A system of masses is set up in the xy-plane at $z = 0$ such that one mass is at the center and the others build a ring around it. The gravitational wave comes along the z-axis perpendicular to the xy-plane. The metric at $z = 0$ is

$$ds^2|_{z=0} = -dt^2 + [1 - a\sin(\omega t)]\, dx^2 + [1 + a\sin(\omega t)]\, dy^2$$

and with $X = (1 - \frac{1}{2} a \sin(\omega t))x$ and $Y = (1 + \frac{1}{2} a \sin(\omega t))y$ the metric becomes

$$ds^2|_{z=0} = -dt^2 + dX^2 + dY^2$$

plus terms of order $O(a^2)$. This is now flat Minkowski space, and one can visualize the geometry with Euclidean intuition. The result is called the plus-polarization (+polarization) presented in figure 7.5 (a).

If one instead chooses the values a and b in (7.3) such that b is small and a is zero, while the wave still travels along the z-axis, the real part of the perturbation of the metric is

$$h_{\mu\nu}(X) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & b & 0 \\ 0 & b & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \sin(kz - \omega t)$$

and corresponds to

$$ds^2|_{z=0} = -dt^2 + dx^2 + 2b\sin(\omega t) dx dy + dy^2$$

for the distance. With $X = x + \frac{1}{2}b\sin(\omega t)y$ and $Y = y + \frac{1}{2}b\sin(\omega t)x$ the metric becomes

$$ds^2|_{z=0} = -dt^2 + dX^2 + dY^2$$

because $dX = dx + \frac{1}{2}b\sin(\omega t)dy$ and $dY = dy + \frac{1}{2}b\sin(\omega t)dx$ as well as $dX^2 \approx dx^2 + b\sin(\omega t)dxdy$ and $dY^2 \approx dy^2 + b\sin(\omega t)dxdy$. This is again flat Minkowski space, and the result is called the cross-polarization (×polarization) shown in figure 7.5 (b).

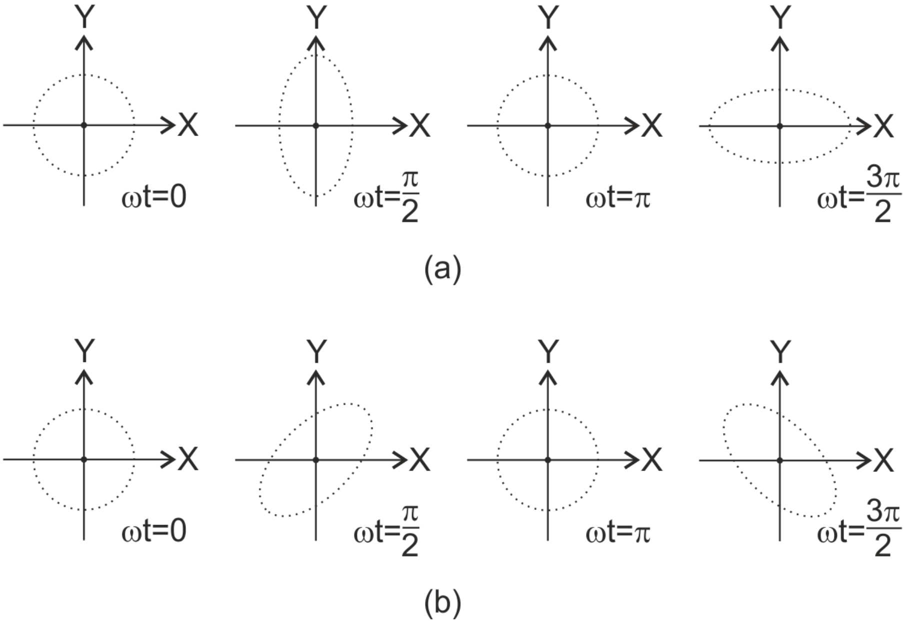


Figure 7.5: Polarization of a gravitational wave

These are the two independent polarization states of the plane gravitational wave. The polarization of the electromagnetic wave can similarly be decomposed into an x-polarization and a y-polarization. The difference though is that the electromagnetic wave is invariant if one flips it by 180° while the gravitational wave is invariant if one flips it by 90°. One can tie that to the fact that the graviton has spin two and the photon has spin one.

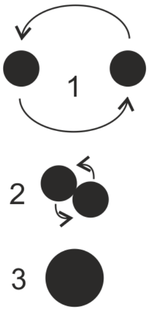
To really detect gravitational waves one could use several such rings of masses because one does not know the direction in which the wave comes and the ring should be perpendicular to this direction. With a ruler one could measure how the ring changes. The ruler does not expand and contract the same way because the above analysis used the geodesic equation for free test particles, and the atoms building the ruler are not free but also experience electromagnetic binding forces which swamp the gravitational distortion. Tiny displacement in physics are not measured with rulers but with interferometers.

The LIGO (Laser Interferometer Gravitational-Wave Observatory) uses four kilometer long Michelson interferometers with mirrors attached to free test masses which are actually hanging but are free to swing. The accuracy is in the order of 10^{-21} and the change of length is in the order of 10^{-18} m. So much noise has to be eliminated that quantum fluctuations must be filtered out. There are other projects planned. The LISA (Laser Interferometer Space Antenna) project uses satellites in space with a length scale for the arms of $5 \cdot 10^6$ km, and the PTR (Pulsar Timing Arrays) will observe irregularities in what should be periodic signals from pulsars.

The last question related to gravitational waves is how they get created. Thus one has to solve Einstein's equations in the presence of sources, and one cannot use small approximations because one wants to see a signal big enough to be measured. One can create electromagnetic waves by moving charges as in an antenna to get uniform radiation, but for gravitational wave production there is nothing popping energy into the system to make it steady state. One has a time-dependent system which produces the radiation.

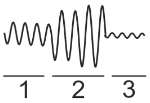
If two black holes come close then they merge and become a single black hole. Such an event has been observed by LIGO. Studying realistic gravity wave generation is difficult and will not be shown here, but there are two interesting features. One is the possibility of multipole expansion and the other is the observable signals from the binary mergers of two black holes.

Firstly, just like any other form of radiation, one can take the far field limit and do a multipole expansion of the power distribution. For both electromagnetic and gravitational waves, the monopole contribution vanishes because of conservation of charge and mass. The leading electromagnetic term is dipole. For General Relativity, however, also the dipole term vanishes because of conservation of angular momentum and the ability to coordinatize to zero the center of mass motion. Thus the lowest term is quadropole.



Secondly, appreciable signals can arise from binary mergers. In particular, when massive black holes merge they can release gravitational wave energies in the order of the mass of the sun. To analyze a merger, the problem is often broken up into stages. Everything could in principle be done numerically, but Einstein's equations are hard and one would have to simulate over a large region to get far-field behavior.

Two black holes merge in three phases. The first phase is called *inspiral*, and one can use post-Newtonian approximations (linear approximations) to address the two-body problem. The second phase is called *merger*, and one uses numerical calculations to handle it. The third and last phase is called *ringdown* where the resulting black hole still wobbles before it settles down to a Kerr black hole, and one uses single-body black hole perturbation theory for calculations.



The complete profile is often called the *chirp* characteristic of the event which takes place in about 0.5 s. The strength of the signal is different for the three phases. One of the fascinating things about black hole mergers compared to other merger events is the ringdown signature. Since ringdown only happens for black holes, its observation is a direct observation of black holes.

7.9 Thermodynamics and Other Features of Black Holes

There are interesting results, some are proven and some are not, but there is non-trivial evidence that they are true. A set of *singularity theorems* which were largely developed by Hawking and Penrose in the sixties belong to them. One might think that a perfect collapse to a point singularity as in the Schwarzschild case or a ring singularity as in the Kerr case is a feature of the high degree of symmetry assumed which would probably not occur in realistic cases which are perturbed. The singularity theorems suggest otherwise. They essentially use the notion of a trapped surface that forms during collapse but before a singularity has formed. These are similar to event horizons (though technically distinct) and force the collapsing matter inside of them to decreasing r. The important part is that even though the physics at singularities cannot be described by General Relativity, the trapped surfaces (and event horizons for that matter) are, and so General Relativity predicts its own shortcoming.

Another interesting result is the *cosmic censorship conjecture* which is an unproven but well supported idea that any singularity that results from collapse will always be hidden behind an event horizon. Some motivation for this comes from the singularity theorems themselves. This does not completely preclude the existence of naked singularities because not all singularities come from collapse.

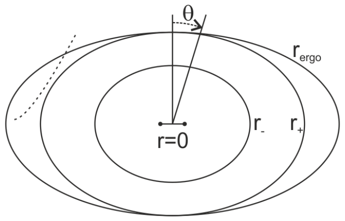
The *no-hair theorem* states that stationary, asymptotically flat black holes are completely characterized by their mass m, charge Q and angular momentum J. This is pretty amazing because it says that all of the complexity of a macroscopic system is essentially lost if it collapses. These quantities are three numbers and characterize a black hole completely. Actually most black holes are electrically neutral such that two numbers characterize them.

In the *area theorem* Hawking showed that the area of an event horizon can never decrease when assuming the weak energy condition, which is one of the studied constraints on the sources (existing matter) and essentially requires $\rho \geq 0$ for the energy density. For Schwarzschild black holes this theorem is obvious because adding mass to it increases M and therefore also $r = 2GM$. For other cases of black holes this is more difficult to prove.

Usually test particles are so small that they do not change the geometry created by the big sources. For black holes this is different. There is a pretty good understanding of how to identify conserved quantities for test particles by using Killing vectors and 4-momentum, but this is not clear how to define conserved quantities for a geometry such as Minkowski spacetime, a Schwarzschild black hole or a Kerr black hole. It is not obvious what is meant by the energy or momentum of a geometry. This is a subtle topic and there are a couple of approaches useful in different situations.

One approach is to use Komar integrals to define conserved quantities if one has a Killing vector field for a geometry. If one uses $K^\mu = (1,0,0,0)$ for t-independence, one finds that for $\mathbb{M}^4$ the energy is $E = 0$ and for Schwarzschild and Kerr geometry the energy is $E = M$ which is the mass of the black hole. This is interesting for the Kerr geometry because of the angular momentum which does not contribute to the energy. Using the Killing vector $R^\mu = (0,0,0,1)$ for angular momentum of the Kerr geometry shows that J is conserved.

The Kerr geometry provides highly non-trivial tests and implications for the preceding results. Kerr black holes with $a^2 > G^2M^2$ such that they are over-extreme, for example, would violate the cosmic censorship conjecture and show a naked singularity if they existed. Spinning black holes can have a large angular momentum very near G^2M^2. Thus one can start from the extremal case $a^2 = G^2M^2$ and feed it a bit of mass δM with angular momentum δJ which is, informally speaking, larger than the mass which is possible by increasing the impact parameter or the speed. This looks like it would push the black hole to over extreme, but it does not. It always scatters and is never absorbed by the black hole as one can show with rigorous calculations. The over-extremal case is therefore not possible.



The Kerr geometry has another important feature. Considering an object at rest outside of the black hole using some thrusters or other means such that $U^\mu = (U^0, 0, 0, 0)$ in Boyer-Lindquist coordinates $\{t, r, \theta, \phi\}$ but

$$U_\mu U^\mu = g_{00} U^0 U^0 = -\left(1 - \frac{2GMar}{\rho^2}\right) {U^0}^2 = -1$$

and this implies that for $(1 - \frac{2GMar}{\rho^2}) < 1$ no object can remain at rest. Inside of the black hole not being able to remain at rest is expected, but this is a region outside which is called ergosphere region. With $\rho^2 = r^2 + a^2 \cos(\theta)^2$ the condition is

$$\left(1 - \frac{2GMar}{r^2 + a^2\cos(\theta)^2}\right) < 1 \qquad \Rightarrow \qquad r < r_{\text{ergo}}(\theta) = GM + \sqrt{G^2M^2 - a^2\cos(\theta)^2}$$
$$\geq r_+ = GM + \sqrt{G^2M^2 - a^2}$$

and the effect is called "frame-dragging". If being at rest is not possible, one can obviously also not escape radially out but only to the side.

There is another important aspect of the ergosphere region. Considering a Killing vector $T^\mu = (1,0,0,0)$ for t-independence with

$$E_0 = m_0 \left(1 - \frac{2GMar}{\rho^2}\right) \frac{dt}{d\tau} + \frac{m_0\, 2GMar}{\rho^2} \frac{d\phi}{d\tau}$$

where the first term is positive if $r > r_{\text{ergo}}$ and negative otherwise while the second term is always positive because a and $d\phi$ have the same sign. Therefore E_0 is always positive if $r > r_{\text{ergo}}$, but E_0 can be positive or negative for $r < r_{\text{ergo}}$. This means that if an object is inside of the ergosphere and has $E_0 < 0$, then it cannot escape. If an object is outside of r_{ergo} with $E_0 > 0$ and enters the ergosphere where it splits into two pieces. If the two pieces have the energies $E_1 + E_2 = E_0$ such that $E_1 > E_0 > 0$ and consequently $E_2 < 0$, then the piece with E_2 is trapped and the piece with E_1 can leave the ergosphere with more energy then the whole object had initially. The energy of the black hole must therefore have decreased and lost mass. (Objects in physics usually have positive energy but this is in flat space, and it is because of the geometry that an object can have negative energy.)

To get this so-called Penrose process to work such that the negative E_2 be absorbed and the positive E_1 be on a trajectory that leaves the ergosphere, one can show that

$$J_2 \leq \frac{E_2}{\Omega_H} \qquad\qquad \Omega_H = \frac{a}{r_+^2 + a^2}$$

where Ω_H is the angular velocity of the horizon. Since $E_2 < 0$ the two quantities J_2 and Ω_H must have opposite signs, or, in other words, the absorbed object must have angular momentum opposite in direction of the black hole. This means that the absorption of the piece with E_2 reduces the angular momentum of the black hole by $\delta J_{\mathrm{BH}} = J_2$.

This process requires an ergosphere region. Therefore the limit of this is when $J_{\mathrm{BH}} = 0$ and the Kerr black hole becomes a Schwarzschild black hole which does not have an ergosphere. The question is whether the horizon area gets smaller which would not be allowed according to Hawking's area theorem. The area of the horizon is

$$A_H = \int \sqrt{\det \gamma}\, d\theta\, d\phi = 4\pi(r_+^2 + a^2) = 8\pi G^2 M^2 + 8\pi\sqrt{G^4 M^4 - G^2 M^2 a^2}$$

where γ_{ij} is the induced metric on the horizon from ds^2 with $r = r_+$ and $dr = dt = 0$ and $M^2 a^2$ can be replaced by J^2. If one varies δM and δJ one finds

$$\delta A_\mu = \frac{8\pi G a}{\Omega_H \sqrt{G^2 M^2 - a^2}}\,(\delta M - \Omega_H \delta J)$$

where the left factor is positive and the right factor can also not be negative because δM is E_2 and $\delta J \leq \frac{\delta M}{\Omega_H}$. This shows that $\delta A_H \geq 0$ and that the area theorem is obeyed.

The fact that A_H cannot decrease according to the area theorem resembles the fact that entropy cannot decrease. Thus one can try to identify the entropy with the area of the horizon. An argument for this comes from Beckenstein. If a black hole had no observable entropy one could take an external system with entropy S_0 and upon throwing it into the black hole decrease the entropy of the observable universe, thus violating the second law of thermodynamics. To preserve the second law of thermodynamics, black holes must admit an observable entropy. One might think to use the mass M of the black hole to correlate with entropy, but the Penrose process allows M to decrease in certain cases. It is only A_H which is a suitable proxy for entropy. But this already shows something deep. While entropy usually scales with the volume of a system, in this case it scales with the area. This points to the holographic nature of gravity, since information from four dimensions is captured by a three-dimensional surface.

Quantifying this idea from the Kerr case gives

$$\delta M = \frac{K}{8\pi G}\,\delta A + \Omega_H\,\delta J \qquad \text{with} \qquad K = \frac{\sqrt{G^2 M^2 - a^2}}{2GM(GM + \sqrt{G^2 M^2 - a^2})}$$

where K is the so-called surface gravity of the black hole, or roughly how strong the gravitational pull is near the horizon. Comparing this with $dE = T\,dS - P\,dV$ and associating $E = M$, $-P\,dV = \Omega_H\,\delta J$, $T\,dS = \frac{K}{8\pi G}\,\delta A$ one might be tempted to identify $T = \frac{K}{8\pi G}$ and $dS = \delta A$, but in truth the split is not obvious.

Hawking considered Quantum Field Theory in the curved geometry near the horizon of a black hole. He was not doing quantum gravity which is still not completely understood, but he was doing perfectly well-defined Quantum Field Theory with minimal coupling. Out of the vacuum one can get pairs of particles such as an electron and a positron which annihilate each other after some time. If this happens close to the horizon then one of the two particles can go behind the horizon and the other can escape to infinity. This naturally makes the black hole smaller and smaller, and the horizon shrinks. This does not contradict the area theorem because of the weak energy condition. Quantum fluctuations have $\rho < 0$, and the negative energy is not negative energy because of the geometry. From outside one sees the so-called Hawking radiation, and black holes can evaporate.

Even if one does not trust Quantum Field Theory in curved space, one could get the same result with Quantum Field Theory in flat space and then apply the equivalence principle. If one is in flat space at rest and start accelerating, then one sees a stream of particles coming in the opposite direction of the acceleration. This so-called Unruh effect is the consequence of uniform acceleration in flat space where an observer does not experience $\mathbb{M}^4$ but a Rindler spacetime. Quantizing a field in terms of Rindler time is very different than with Minkowski time. The end result is that uniformly accelerated observers in the vacuum of $\mathbb{M}^4$ see a thermal distribution of all allowed particle types dominated by the lowest mass coming at them with the temperature $T = \frac{a}{2\pi}$ where a is their acceleration. Using the equivalence principle allows to replace a by K and find $T = \frac{K}{2\pi}$ and hence $dS = \frac{\delta A}{4G}$ and therefore $S = \frac{A}{4G}$.

This shows that the temperature of a black hole is $T = \frac{K}{2\pi}$, but perhaps the most surprising aspect of Hawking's result is that black holes seem to radiate. Conservation of energy implies that the black hole is losing energy and hence mass in this process. In the Schwarzschild case it is obvious that the horizon area is decreasing. However, black holes evaporate at a small rate especially compared to any accretion. For microscopic black holes it leads to them being very short-lived.

This leads to the *black hole information paradox.* In an otherwise empty universe and given a long enough time in principle any black hole will completely evaporate. If there are two non-rotating equal masses such as a male and a female cow which both turn into a black hole. At this stage one can still think that the information whether it is the male or the female black hole is hidden behind the horizon, but when both black holes evaporated completely there is only radiation with the same temperature and the information is lost. Resolving this puzzle will almost certainly require a well-understood quantum theory of gravity. The holographic principle is a first step.

8 Cosmology

8.1 Friedmann-Robertson-Walker Cosmologies

In the above solutions of Einstein's equations with or without sources symmetries helped to simplify the problems. This usually meant idealizing to a single source as for Schwarzschild or Kerr black holes or exploring asymmetric behavior as for gravitational waves. There is yet another method to get Einstein's equations to simplify and that is by applying them to the entire universe and smooth over non-uniformities. This is the starting point of cosmology. To do so, one has to use $T_{\mu\nu} \neq 0$, must not assume t-independence because the universe evolves with time, and identify what symmetries are present to choose appropriate coordinates.

The symmetries of spacetime in the cosmological context are spatial homogeneity which means translation invariance in space and spatial isotropy which means rotation invariance in space. (These two symmetries are not the same because living on the surface of a cylinder, for example, exhibits homogeneity but no isotropy, and living in a Schwarzschild geometry offers isotropy at the center but no homogeneity.) Together these two symmetries imply that at any point in space one observes rotational invariance and therefore that there is no center of the universe. It also means that the spatial geometry is maximally symmetric.

For the spatial coordinates so-called comoving coordinates are used to describe spacetime. They are adopted to the rest frame of the source, even if it expands or shrinks with time. That is, if the proper distance between two objects increases because of spacetime expansion, then the coordinate separation will remain fixed.

Putting things together allows to define the metric as

$$ds^2 = -dt^2 + R^2(t)\,\gamma_{ij}(u)\,du^i\,du^j \tag{8.1}$$

with $i, j \in \{1, 2, 3\}$ before one starts solving Einstein's equations. The factor $R^2(t)$ has dimension length squared, and the factor $d\sigma^2 = \gamma_{ij}(u)\,du^i\,du^j$ as the t-independent spatial geometry is therefore dimensionless. This metric will preserve whatever spatial symmetry is imposed.

An observer using coordinates adopted to a different reference frame (even with constant velocity) will see a different metric with different symmetries. For example on earth one notices a dipole anisotropy in the cosmic microwave background due to the motion of the earth relative to the overall rest frame of the universe. The maximally symmetric spatial geometry leads to the curvature tensor and its derivation needed in Einstein's equations

$$R_{ijkl} = k(\gamma_{ik}\gamma_{jl} - \gamma_{il}\gamma_{jk}) \qquad \Rightarrow \qquad R_{jl} = \gamma^{ki}\,R_{ijkl} = 2k\gamma_{jl} \qquad \Rightarrow \qquad R = \gamma^{lj}\,R_{jl} = 6k$$

but only for the spatial indices. (In a maximally symmetric geometry one can calculate the curvature tensor algebraically without the need for Christoffel symbols and so on.) This is actually to be expected because if the space is maximally symmetric then the curvature must be the same everywhere. Otherwise the different points in space would have different curvature making them distinguishable.

One can categorize the possible spatial geometries by the sign of the constant k

$$\begin{array}{lllll} k=0 & \Rightarrow & d\sigma^2 = d\chi^2 + \chi^2\, d\Omega_2^2 & \Rightarrow & \mathbb{R}^3 \text{ flat} \\ k>0 & \Rightarrow & d\sigma^2 = d\chi^2 + \sin(\chi)^2\, d\Omega_2^2 & \Rightarrow & \mathrm{S}^3 \text{ closed} \\ k<0 & \Rightarrow & d\sigma^2 = d\chi^2 + \sinh(\chi)^2\, d\Omega_2^2 & \Rightarrow & \mathrm{H}^3 \text{ open} \end{array} \tag{8.2}$$

in polar coordinates. Because the three cases do not make it easy to switch between them, one defines

$$d\chi = \frac{d\bar{r}}{\sqrt{1-k\bar{r}^2}} \qquad \Rightarrow \qquad d\sigma^2 = \frac{d\bar{r}^2}{1-k\bar{r}^2} + \bar{r}^2\, d\Omega_2^2$$

with $k \in \{0, -1, +1\}$. Connecting this back to the three cases in (8.2) gives

$$\begin{array}{llll} k=0 & d\chi = d\bar{r} & \Rightarrow & \chi = \bar{r} \\ k=-1 & d\chi = \dfrac{d\bar{r}^2}{1+\bar{r}^2} & \Rightarrow & \chi = \sinh^{-1}(\bar{r}) \\ k=+1 & d\chi = \dfrac{d\bar{r}^2}{1-\bar{r}^2} & \Rightarrow & \chi = \sin^{-1}(\bar{r}) \end{array}$$

and the resulting metric describing spatially homogeneity and isotropy with time dependence becomes

$$ds^2 = -dt^2 + R^2(t)\left[\frac{d\bar{r}^2}{1-k\bar{r}^2} + \bar{r}^2\, d\Omega_2\right] \tag{8.3}$$

which is called *Robertson-Walker metric.* With some fixed length R_0 and the definitions

$$a(t) \equiv \frac{R(t)}{R_0} \qquad r \equiv R_0\, \bar{r} \qquad K \equiv \frac{k}{R_0^2}$$

where $a(t)$ is a dimensionless scale factor, r the distance with a dimension, and K the spatial curvature also with a dimension, the metric (8.3) becomes

$$ds^2 = -dt^2 + a(t)^2\left[\frac{dr^2}{1-K\, r^2} + r^2\, d\Omega_2^2\right] \tag{8.4}$$

where the two unknowns $a(t)$ and K have to be determined by Einstein's equations.

Assuming to be at rest with respect to the overall notion of the sources such that $U^\mu = (1,0,0,0)$ the energy-momentum tensor for a perfect fluid source is

$$T_{\mu\nu} = (\rho + p)\, U_\mu\, U_\nu + p\, g_{\mu\nu} = \begin{pmatrix} \rho & 0 & 0 & 0 \\ 0 & & & \\ 0 & & g_{ij}\, p & \\ 0 & & & \end{pmatrix}$$

where $g_{ij} \neq \gamma_{ij}$ because g_{ij} includes $a(t)$.

Einstein's equations in trace-reversed form is $R_{\mu\nu} = 8\pi G(T_{\mu\nu} - \frac{1}{2} g_{\mu\nu} T)$. Using the Robertson-Walker metric in the form (8.4) with the perfect fluid energy-momentum tensor gives

$$-3\frac{\ddot{a}}{a} = 4\pi G(\rho + 3p) \qquad \frac{\ddot{a}}{a} + 2\left(\frac{\dot{a}}{a}\right)^2 + 2\frac{K}{a^2} = 4\pi G(\rho - p)$$

for the 00 and the ij term, respectively. By combining them to get rid of the second derivative $\ddot{a}$ one obtains the equation

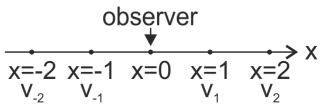
$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{K}{a^2} \qquad H^2 = \frac{8\pi G}{3}\rho - \frac{K}{a^2} \tag{8.5}$$

which is called *Friedmann equation* and where $H(t) \equiv \frac{\dot{a}(t)}{a(t)}$ is called the *Hubble parameter* which is positive for an expanding and negative for a contracting universe. Solutions of this equation are called *Friedmann-Robertson-Walker cosmologies* and are very good first approximations of the universe.

If one sees an object in the universe and asks how fast is it moving then the physical velocity of this object is

$$v_{\text{physical}} = \frac{dl_{\text{physical}}}{dt} = \frac{d[a(t)\,l_{\text{coord}}]}{dt} = \dot{a}\,l_{\text{coord}} = \frac{\dot{a}}{a}\,l_{\text{physical}} = H\,l_{\text{physical}}$$

where l_{physical} is the physical distance while the coordinate distancde l_{coord} is a constant because the coordinates have been chosen this way. The result $v_{\text{physical}} = H\,l_{\text{physical}}$ that the physical velocity of an object in the universe seen is proportional to the physical distance, is called *Hubble's law.* This means that the universe is expanding.



Hubble's law describes relative motion in expanding spacetime. There is an important difference between an explosion and an expansion. If an observer is at $x = 0$ in a one-dimensional space with $H = 1\,\text{s}^{-1}$ then the speed of objects in this expanding space at position $x = 1$ is $v_1 = 1$, at position $x = 2$ is $v_2 = 2$, and at $x = -2$ is $v_{-2} = 2$ but in opposite direction such that $v_x = x$ for the numerical value. After an explosion the speed is the same for any piece independent of its distance to the observer where the explosion has taken place such that $v_1 = v_2$, for example. The expansion has no center corresponding to homogeneity because the observer at any position sees the same expansion as the observer at $x = 0$, but the explosion has a center and therefore no homogeneity because somebody at $x = 1$ sees different things happening than the observer at $x = 0$. The important point is that spacetime itself is expanding.

The expansion of spacetime is because of gravity, but there are other forces such as electromagnetic forces which keep things together. These forces keep atoms, for example, at a fix distance. Only things which are only impacted by gravity but no other forces separate according to Hubble's law. Thus the ruler to measure the distance does not expand.

The Friedmann equation contains $a(t)$, $\rho(t)$ and K which are needed to complete solving Einstein's equations. The current value of the Hubble parameter H and the value K are measurable by observing the universe, and ρ is a bit more subtle and has several contributions. Since $a(t)$ depends on t and $\rho(t)$ depends on t, one can express how ρ varies with a as $\rho(a)$.

Conservation of energy-momentum $\nabla_\mu T^{\mu\nu} = 0$ or $g_{\alpha\nu} T^\mu{}_\alpha = 0$ gives $\nabla_\mu T^\mu{}_0 = -\frac{d\rho}{dt} - 3\frac{\dot{a}}{a}(\rho + p) = 0$ for $\alpha = 0$. From assuming an equation of state of the form $p = \omega\,\rho$ for a constant ω it follows

$$0 = -\dot{\rho} - 3\frac{\dot{a}}{a}(1+\omega)\rho \quad \Rightarrow \quad \frac{\dot{\rho}}{\rho} = -3(1+\omega)\frac{\dot{a}}{a} \quad \Rightarrow \quad \ln(\rho) = -3(1+\omega)\,\ln(a) \quad \text{or} \quad \rho(t) \propto a(t)^{-3(1+\omega)}$$

with these cases

- Matter (or dust) satisfies $p_M = 0$ and therefore $\omega = 0$ such that $\rho_M(t) \propto a^{-3}$ follows which expresses dilution of fixed particle number due to volume expansion.
- Radiation (or highly relativistic matter) satisfies $p_R = \frac{1}{3}\rho_R$ and therefore $\omega = \frac{1}{3}$ such that $\rho_R(t) \propto a^{-4}$ which expresses volume dilution a^{-3} and redshift in the direction of motion a^{-1}.
- Vacuum ($T_{\mu\nu} \propto g_{\mu\nu}$) satisfies $p_V = -\rho_V$ and therefore $\omega = -1$ such that $\rho_V(t) \propto a^0$ which is constant. (One cannot move relative to a vacuum.)

The result is $\rho_{\text{tot}} = \rho_M + \rho_R + \rho_V$, but if one defines $\rho_C \equiv -\frac{3K}{8\pi G a^2}$ for curvature such that $\rho_C \propto a^{-2}$ then Friedmann's equation becomes

$$H^2 = \frac{8\pi G}{3}\sum_i \rho_i$$

for $i \in \{M, R, V, C\}$. If the universe is at some time dominated by one of these with $\rho_{\text{dom}} \propto a^{-n}$ then

$$H^2 = \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}a^{-n} \qquad \Rightarrow \qquad \frac{da}{dt} = \sqrt{\frac{8\pi G}{3}}a^{1-\frac{1}{2}} \qquad \Rightarrow \qquad a^{\frac{1}{2}-1}\,da = \sqrt{\frac{8\pi G}{3}}\,dt$$

$$\Rightarrow \qquad a(t) = \left(\frac{n}{2}\sqrt{\frac{8\pi G}{3}}\right)^{\frac{2}{n}} t^{\frac{2}{n}}$$

such that

$$\rho_{\text{dom}} = \begin{cases} \rho_M \propto a^{-3} & \Rightarrow a(t) \propto t^{\frac{2}{3}} \\ \rho_R \propto a^{-4} & \Rightarrow a(t) \propto t^{\frac{1}{2}} \\ \rho_C \propto a^{-2} & \Rightarrow a(t) \propto t \\ \rho_V \propto a^{0} & \Rightarrow a(t) \propto e^{Ht} \end{cases}$$

where ρ_V can be derived from $\frac{da}{a} = H\,dt$ with $H = \sqrt{\frac{8\pi G}{3}}$ and $\ln(a) = H\,t$.

If the values of ρ_i depends differently on a then the expanding universe has been dominated by them differently at different times. For small a at early times ρ_R has dominated, for large a at late times ρ_V will dominate, and in between first ρ_M and later ρ_C dominate. A consequence is that $a(t \to 0) \to 0$ and this means that the universe started with a big bang unless ρ_V dominated at early time which is highly unlikely.

8.2 The Universe

A Friedmann-Robertson-Walker universe dominated at early times by anything other than vacuum energy must have begun with a *big bang*, and this is of course a naked singularity. The cosmic censorship conjecture does not apply because the big bang is not the result of a collapse.

In order to understand the universe one needs to know more about $a(t)$, K and $\rho(t)$ in (8.5). Certain quantities can be measured but others such as vacuum energy cannot and must come out of calculations. Cosmology is in some sense an observable science as opposed to an experimental one. In fact one may argue that much of General Relativity is the same since one cannot create any significant sources, though one can experiment by observing test masses. Furthermore, cosmology is the study of the time-varying history of the universe, and unlike the criterion for any good experiment, it would not repeat itself.

The question is what one can actually observe. Much of the precise knowledge is the byproduct of astrophysics. The detailed study of stellar models including both nuclear and gravitational effects has provided a pretty clear prediction of how certain stars should behave such as their luminosity, size, and emission spectra. Observing these so-called *standard candles* and distortion from their predicted features delivered information on the non-trivial geometry through which their light is moving and therefore on the slope of the universe.

If one can predict an emission spectrum for a standard candle and then observe one that is shifted towards infrared, one can relate the size of the redshift to the relative sizes of the universe between the time of the emission long time ago and the observation now. The frequency ω_o observed should be related to the frequency ω_e of emission inversely of the way the scale factor observed a_o is related to the scale factor of emission a_e because expansion leads to a redshift. Formally this means $\frac{\omega_o}{\omega_e} = \frac{a_e}{a_o}$. The redshift factor z is defined in terms of the observed and emitted wave lengths λ_o and λ_e, respectively, as

$$z \equiv \frac{\lambda_o - \lambda_e}{\lambda_e} = \frac{\lambda_o}{\lambda_e} - 1 = \frac{\omega_e}{\omega_o} - 1 = \frac{a_o}{a_e} - 1 \qquad\qquad a_e = \frac{a_o}{1+z}$$

and the observed value is $z > 0$ and therefore $a_o > a_e$ showing that the universe is expanding. One can approximate $a_e \approx a_o + (t_e - t_o)\dot{a}_o$

The observed current value of the Hubble parameter is $H_o = \frac{\dot{a}_o}{a_o}$. Because $f(x_2) \approx f(x_1) + (t_2 - t_1) f'(x_1)$ for x_1 and x_2 close together such that $a_e \approx a_o + (t_e - t_o)\dot{a}_o$ and $\frac{a_e}{a_o} \approx 1 + (t_e - t_o)H_o$ and because $\frac{1}{1-x} \approx 1 + x$ one can use

$$z = \frac{a_o}{a_e} - 1 \approx \frac{1}{1 + (t_e - t_o)H_o} - 1 = \frac{1}{1 - (t_o - t_e)H_o} - 1 \approx (t_o - t_e)H_o = d\,H_o$$

assuming the light traveled at $c = 1$ such that $t_o - t_e = d$ is the distance. Thus if z is known for a star from the redshift of its spectrum and if one knows how far away this star is, one can determine H_o. The question remains how one measures the distance d. Before one can measure the distance one should know what is meant by “distance” in cosmology, and this is by no means a trivial question as scientists in astrophysics have learned.

There are at least five distinct notions of distance:

- The coordinate distance is useful in computations but it is not “physical” in the sense that it has no physical meaning.
- The equal time distance to a distant object specifying how far it is from the earth now is less directly tied to observation.
- The observed distance to a distant object indicates its distance at the time of light emission that is observed now.
- The angular separation is the distance between two distant sources at approximately the same distance from the earth.
- The angular size of a source is the length across a distant source which does in contrast to angular separation not expand with time since the object is bound by larger forces.

There is also the luminosity distance defined as

$$L = 4\pi\, d_L^2\, F \qquad\qquad d_L = \sqrt{\frac{L}{4\pi\, F}}$$

where F is the measured flux. This distance is useful for standard candles, but for small values of z the different definitions of distance correlate with each other and give pretty much the same answer. The luminosity L is known because one knows what a star is made of because one knows the astrophysics of this type of star. The observed flux F of the star is a quantity that can be directly measured on the earth.

For low redshift observations one can use $d = d_L$ and combine with z measurements determine a value for H_o as

$$H_o = 70 \pm 10\,\frac{\text{km}}{\text{s}\,\text{Mpc}} \qquad \Rightarrow \qquad d_H = \frac{c}{H_o} = 4.55 \cdot 10^6\,\text{pc} \qquad t_H = \frac{1}{H_o} = 14.4 \cdot 10^9\,\text{years}$$

with $1\,\text{pc} = 3.086 \cdot 10^{13}\,\text{km}$. The distance d_H is the Hubble distance which is effectively the size of the universe, and t_H is then the age of the universe. This is however just the observable universe, but the real universe might be much larger, and for large redshifts things get much more complicated.

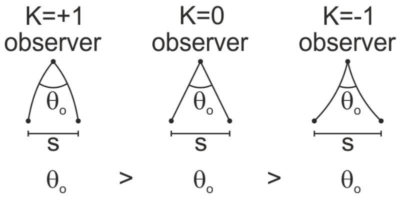
Turning to the other quantities ρ and K to be determined, one defines

$$\Omega_i \equiv \frac{8\pi G}{3H^2}\rho_i = \frac{\rho_i}{\rho_{\text{crit}}} \qquad\qquad \rho_{\text{crit}} \equiv \frac{3H^2}{8\pi G}$$

for $i \in \{M, R, V, C\}$ where the critical energy density ρ_{crit} is the total energy density needed so that $K = 0$ and the universe is therefore spatially flat. The condition is if $\sum_i \Omega_i = 1$ then $K = 0$ because the energy density determines the curvature according to Einstein’s equations. As one makes the measurements at the current time the respective quantities are Ω_{i_o}, ρ_{i_o} and H_o. They are numbers between zero and one.

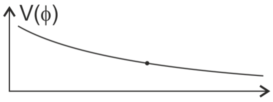
To get Ω_{M_o} one looks at a cluster and uses local gravitational effects to infer mass. Then one uses density of clusters to extrapolate to large scale. The result is $\Omega_{M_o} = 0.3 \pm 0.1$. For Ω_{R_o} the cosmic microwave background arises from relic photons after last scattering once electrons and protons cooled to form electrically neutral atoms such that the universe became electromagnetically transparent. One measures a thermal distribution with $T = 2.73^\circ$ K and hence gets $\Omega_{R_o} \approx 10^{-4}$. Because $\rho_M \propto a^{-3}$ and $\rho_R \propto a^{-4}$ this result showing $\rho_{M_o} \gg \rho_{R_o}$ makes sense.

For the curvature Ω_{C_o} one can predict anisotropies over a length scale of s from the understanding of the cosmic microwave background which should not be completely uniform. One can observe the angular size of the anisotropies, and one can determine the spatial curvature by comparing these values to s. The observation indicates that $K = 0$ such that $\Omega_{C_o} = 0$. This means that from $\sum_i \Omega_{i_o} = \Omega_{M_o} + \Omega_{R_o} + \Omega_{C_o} + \Omega_{V_o} = 1$ follows $\Omega_{V_o} = 0.7 \pm 0.1$.



8.3 Puzzles in Cosmology

The result $\Omega_{M_o} = 0.3$ is determined from observed gravitational dynamics by measuring rotational velocities of spiraling galaxies and inferring the mass needed to hold them together. However one can also take note of how much matter can be seen or how much luminous matter there is. The amount of matter one sees in the universe is $\Omega_{B_o} = 0.04 \pm 0.02$ largely due to baryonic matter (protons, neutrons). To balance Ω_{M_o} the conclusion is that there must be so-called *dark matter* with $\Omega_{D_o} \approx 0.26$. Dark matter is obviously non-luminous as the name indicates, and it also most likely does not consist of massive compact halo objects such as black holes, white dwarfs or neutron stars although one could guess that black holes, for example, would be good candidates for dark matter. It is also most likely not baryonic because the program of big bang nucleosynthesis is remarkably good at predicting the relative abundance seen today, but also easy to screw up by modifying its assumptions. It is also most likely not relativistic and therefore not hot.



The term *dark energy* was coined by analogy with dark matter to exhibit the lack of complete understanding of what makes up $\Omega_{V_o} = 0.7$. One expects and can estimate a contribution from the zero-point energies of quantum fields in the Standard Model. This contribution acts much like a cosmological constant term $\Lambda\, g_{\mu\nu}$ in Einstein's equations, but there is a coincidence problem which is that ρ_V is constant while ρ_M dilutes with a^{-3} and today $\Omega_{M_o} \sim \Omega_{V_o}$. Today Ω_{M_o} should be neglectable compared to Ω_{V_o}, but incidentally they have now nearly the same value. A method to address this is to introduce a slowly varying scalar field ϕ into Einstein's equations with a potential $V(\phi)$ that is shallow. The Friedmann equation becomes $\ddot{\phi} + 3H\dot{\phi} + \frac{dV}{d\phi} = 0$ where the term $3H\dot{\phi}$ acts like friction damping the evolution of ϕ. In Einstein's equations $T_{\mu\nu}$ includes $-V(\phi)\, g_{\mu\nu}$, and this is just like $\Lambda\, g_{\mu\nu}$ if $V(\phi)$ is nearly constant. Such a term acts like a vacuum.

From this follows the so-called *cosmological constant problem.* With regards to the value of Λ one can actually get an estimate on its expected value by considering various contributions. The universe is filled with quantum fields with zero-point energies, and whenever a symmetry is spontaneously broken by a field taking a non-symmetric expectation value driven by some effective potential, the height of the potential represents the energy released. The Higgs mechanism contributes $10^{44}\,\mathrm{eV}^4$ and the various zero-point energies of the Standard Model add another $10^{108}\,\mathrm{eV}^4$ giving a $\rho_{\text{theory}} \approx 10^{108}\,\mathrm{eV}^4$, but $\Omega_{V_o} = 0.7$ gives a $\rho_{\text{observed}} \approx 10^{-12}\,\mathrm{eV}^4$ which is 120 orders of magnitude off.

Also the fact that the universe is comprised almost exclusively of matter with almost no anti-matter is a puzzling problem known under the name *baryon asymmetry.* The Standard Model seems to favor nearly equal production of each. Surprising is that even though this looks like a maximally one-sided distribution today, if one traces it back to early times in the universe, it actually only amounts to a difference in matter and anti-matter of one part in a billion. This means that early when the universe was composed of a lot of matter and anti-matter which as it cooled annihilated away and what one sees now is that tiny initial difference. However getting a small asymmetry is even harder than getting none or a maximal one. This is still an open question.

The following three additional puzzles have been more or less addressed by the inflationary model:

1. Flatness problem: Flat geometry with $\Omega = 1$ ($K = 0$) is an unstable solution. Thus the fact that we observe $\Omega = 1$ is either very incidental or something is going on that is not known.
2. Horizon problem: In a universe of finite age there can exist regions at certain times which have never been in causal contact such that they are separated by a cosmological horizon. At the time the cosmological microwave background was formed, the universe was large enough to have many causally disconnected regions, but the cosmological microwave background is observed to be remarkably uniform.
3. Relic (or monopole) problem: Whenever a gauge symmetry with a U(1) subfactor is spontaneously broken, one would expect the production of at least one topological defect in form of a magnetic monopole per causally connected domain. But magnetic monopoles have not been observed.

All three of these problems have been solved by supposing that the universe, in its past, underwent some period of exponential expansion ($a \propto e^{\#t}$ as opposed to power law $a \propto t^{\#}$) which is called *inflation.* During this inflationary growth first proposed by Guth, $\Omega = 1$ actually becomes a stable fixed point and so inflation does the “fine-tuning” for the humans. If inflation occurred before the cosmic microwave background was formed then everything could have been causally connected prior and now only appear to be based on power law growth. If inflation occurred after monopoles were formed, then their density could be “deflated” to nearly zero such that monopoles are too rare to be observable.

Appendix A: Detailed Examples of Transformations

Transformation from Cartesian to Spherical Polar Coordinates

This example of a transformation shows how Cartesian coordinates (x, y, z) in flat Euclidean space $\mathbb{R}^3$ are transformed into spherical polar coordinates (r, θ, φ) which can be specified as

$$x = r\,\sin(\theta)\,\cos(\varphi) \qquad y = r\,\sin(\theta)\,\sin(\varphi) \qquad z = r\,\cos(\theta)$$

expressed for the case that (r, θ, φ) coordinates are given. With

$$g_{ij} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad g_{ij} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2\,\sin(\theta)^2 \end{pmatrix}$$

as the metric in Cartesian coordinates on the left and the metric in spherical polar coordinates on the right, the line elements are

$$ds^2 = g_{ij}\,dx^i\,dx^j = dx^2 + dy^2 + dz^2 \qquad ds^2 = g_{ij}\,dx^i\,dx^j = dr^2 + r^2 d\theta^2 + r^2\,\sin(\theta)^2 d\varphi^2$$

for Cartesian coordinates on the left and for spherical polar coordinates on the right. It is legitimate to specify the geometry of the space either by giving the metric or by giving the line element.

Starting from the trivial metric for Cartesian coordinates the metric in spherical polar coordinates can be determined. The transformation $(x, y, z) \rightarrow (r, \theta, \varphi)$ changes the metric as

$$g_{ij} \rightarrow g_{i'j'} = \frac{\partial x^i}{\partial x^{i'}} \frac{\partial x^j}{\partial x^{j'}} g_{ij}$$

with

$$\frac{\partial x^i}{\partial x^{i'}} = \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} & \frac{\partial x}{\partial \varphi} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} & \frac{\partial y}{\partial \varphi} \\ \frac{\partial z}{\partial r} & \frac{\partial z}{\partial \theta} & \frac{\partial z}{\partial \varphi} \end{pmatrix} = \begin{pmatrix} \sin(\theta)\cos(\varphi) & r\,\cos(\theta)\cos(\varphi) & -r\,\sin(\theta)\sin(\varphi) \\ \sin(\theta)\sin(\varphi) & r\,\cos(\theta)\sin(\varphi) & r\,\sin(\theta)\cos(\varphi) \\ \cos(\theta) & -r\sin(\theta) & 0 \end{pmatrix}$$

such that

$$\begin{aligned} g_{i'j'} = \left(\frac{\partial x^i}{\partial x^{i'}}\right)^T g_{ij} \left(\frac{\partial x^j}{\partial x^{j'}}\right) = & \begin{pmatrix} \sin(\theta)\cos(\varphi) & \sin(\theta)\sin(\varphi) & \cos(\theta) \\ r\,\cos(\theta)\cos(\varphi) & r\,\cos(\theta)\sin(\varphi) & -r\sin(\theta) \\ -r\,\sin(\theta)\sin(\varphi) & r\,\sin(\theta)\cos(\varphi) & 0 \end{pmatrix} \\ & \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ & \begin{pmatrix} \sin(\theta)\cos(\varphi) & r\,\cos(\theta)\cos(\varphi) & -r\,\sin(\theta)\sin(\varphi) \\ \sin(\theta)\sin(\varphi) & r\,\cos(\theta)\sin(\varphi) & r\,\sin(\theta)\cos(\varphi) \\ \cos(\theta) & -r\sin(\theta) & 0 \end{pmatrix} \end{aligned}$$

multiplied in matrix notation gives the above metric in spherical polar coordinates.

In $\mathbb{R}^3$ there is invariance under rotations meaning that ds^2 as well as g_{ij} are unchanged by a rotation. (This is also true in Special Relativity where the invariance is under rotations and boosts.) In this example the space is obviously flat but the line element and the metric are not invariant under the transformation $(x, y, z) \rightarrow (r, \theta, \varphi)$. The reason is that this transformation is not a rotation. The definition of SO(3) is that $R^T\,g_{ij}\,R = g_{ij}$ (and the definition of SO(1,3) is that $\Lambda^T\,\eta_{\mu\nu}\,\Lambda = \eta_{\mu\nu}$) but this is exactly what has been done, and the result was not g_{ij} for the coordinates $\{x, y, z\}$ but a different g_{ij} namely the one for (r, θ, φ). The transformation is therefore a coordinate transformation but not a rotation, and this shows that not all matrices with sine and cosine are rotations.

Lorentz Transformation in General Coordinates

To demonstrate the relation between arbitrary coordinate transformations in General Relativity and the Lorentz transformation in Special Relativity

$$V^{\mu} \to V^{\mu'} = \frac{\partial X^{\mu'}}{\partial X^{\mu}} V^{\mu} \Rightarrow V^{\mu} \to V^{\mu'} = \Lambda^{\mu'}{}_{\mu} V^{\mu}$$

the Lorentz transformation

$$\Lambda^{\mu'}{}_{\mu} = \begin{pmatrix} \cosh(\phi) & -\sinh(\phi) & 0 & 0 \\ -\sinh(\phi) & \cosh(\phi) & 0 & 0 \\ 0 & 0 & \cos(\theta) & -\sin(\theta) \\ 0 & 0 & \sin(\theta) & \cos(\theta) \end{pmatrix}$$

which is a boost in x and a rotation in the yz-plane.

With $dX^{\mu} \to dX^{\mu'} = \Lambda^{\mu'}{}_{\mu} dX^{\mu}$ which is in matrix form

$$\begin{pmatrix} \cosh(\phi) & -\sinh(\phi) & 0 & 0 \\ -\sinh(\phi) & \cosh(\phi) & 0 & 0 \\ 0 & 0 & \cos(\theta) & -\sin(\theta) \\ 0 & 0 & \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} dt \\ dx \\ dy \\ dz \end{pmatrix} = \begin{pmatrix} \cosh(\phi)\, dt - \sinh(\phi)\, dx \\ -\sinh(\phi)\, dt + \cosh(\phi)\, dx \\ \cos(\theta)\, dy - \sin(\theta)\, dz \\ \sin(\theta)\, dy + \cos(\theta)\, dz \end{pmatrix} = \begin{pmatrix} dt' \\ dx' \\ dy' \\ dz' \end{pmatrix}$$

the coordinate transformations are

$$\begin{aligned} dt' &= \cosh(\phi)\, dt - \sinh(\phi)\, dx & \qquad dy' &= \cos(\theta)\, dy - \sin(\theta)\, dz \\ dx' &= -\sinh(\phi)\, dt + \cosh(\phi)\, dx & \qquad dz' &= \sin(\theta)\, dy + \cos(\theta)\, dz \end{aligned}$$

explicitly written out.

It must be possible to get $\Lambda^{\mu'}{}_{\mu}$ in the form of Special Relativity from $\frac{\partial X^{\mu'}}{\partial X^{\mu}}$ in the form of General Relativity. This gives

$$\frac{\partial X^{\mu'}}{\partial X^{\mu}} = \begin{pmatrix} \frac{\partial t'}{\partial t} & \frac{\partial t'}{\partial x} & \frac{\partial t'}{\partial y} & \frac{\partial t'}{\partial z} \\ \frac{\partial x'}{\partial t} & \frac{\partial x'}{\partial x} & \frac{\partial x'}{\partial y} & \frac{\partial x'}{\partial z} \\ \frac{\partial y'}{\partial t} & \frac{\partial y'}{\partial x} & \frac{\partial y'}{\partial y} & \frac{\partial y'}{\partial z} \\ \frac{\partial z'}{\partial t} & \frac{\partial z'}{\partial x} & \frac{\partial z'}{\partial y} & \frac{\partial z'}{\partial z} \end{pmatrix} = \begin{pmatrix} \cosh(\phi) & -\sinh(\phi) & 0 & 0 \\ -\sinh(\phi) & \cosh(\phi) & 0 & 0 \\ 0 & 0 & \cos(\theta) & -\sin(\theta) \\ 0 & 0 & \sin(\theta) & \cos(\theta) \end{pmatrix} = \Lambda^{\mu'}{}_{\mu}$$

as expected.

This result is not surprising because the form of coordinate transformations in General Relativity is much more general than the form in Special Relativity, but it is good do see in an example that the transformations in General Relativity includes those of Special Relativity. It also shows that the resulting matrix is constant and has therefore no dependences on dt, dx, dy, dz.

Appendix B: General Relativity in the Rear View Mirror

The starting point is Special Relativity which is a framework for doing physics that provided the same value of the speed of light to all observers. This principle following from electrodynamics with Maxwell's equations, coupled with the general relativity principle that the laws of physics should appear the same to all inertial observers, has the consequence that one has to give up the concept of an absolute time where everyone agrees on exact time sequence and instead adopt a unified spacetime where certain generalized rotations called boosts mix the spatial and time axes. This allows a set of simultaneous events in one frame to have spatially dependent time ordering in another frame, and makes it necessary to redefine causality from simple time-ordering with an absolute time to a new form in terms of light cones in the four-dimensional spacetime.

Exploring the relativity principle led to many interesting conclusions:

- The Poincaré group SO(1,3)$\ltimes$T^4 with the six Lorentz transformations and the four translations is the relevant symmetry group and all such transformations are specified by constants and relate Cartesian to Cartesian coordinate systems.
- All quantities from three dimensions had to be generalized to four dimensions including energy and momentum which effectively unifies them to a 4-vector.
- The notion of linear algebra with of scalars, vectors and matrices had to be generalized to tensors including scalars and vectors but also higher-rank tensors.
- The construction of densities led to the conclusion that a four-dimensional energy-momentum density is given by the energy-momentum tensor.
- Consistent tools were obtained for describing the behavior of massless particles because classical mechanics cannot handle them.
- The overarching principle from which much of this follows is that Special Relativity is a theory on the four-dimensional Minkowski spacetime whose metric in Cartesian coordinates is diagonal with different signs for the time component than for the three spatial components.

As one knows today much more about Special Relativity and Particle Physics than Einstein knew one does not have to follow Einstein's logic to generalize Newtonian gravity but can develop General Relativity with the understanding of the other forces. In hindsight it is known that Minkowski spacetime is a spacetime in which gravity and therefore also curvature is absent. Special Relativity is the theory of a vanishing energy-momentum tensor corresponding to the absence of gravity. The matrices for the Poincaré group are all built out of constants in Minkowski spacetime.

One might wish to generalize this framework to one that accommodates local and therefore position dependent or even arbitrary coordinate transformations. The consequences are that the metric can vary depending on position and that the derivative has to be redefined to the covariant derivative taking into account that also the base vectors change. As a consequence the transformations are no longer built out of constants. The Christoffel connection in the definition of the covariant derivative plays a similar role as the gauge fields for other forces.

All of this can be done in Minkowski spacetime, but one can also allow the geometry to become dynamical by allowing the gauge field in the form of the Christoffel connection to become dynamical. To do so one introduces a gauge field kinematic term identified by the Riemann curvature tensor. The resulting theory is that of a dynamical spacetime geometry and allows arbitrary coordinates. It inherits simple dynamical principles from Minkowski spacetime such as that the straight line for free particles becomes a geodesic motion in curved geometries. This theory can be applied to describe gravity on any spacetime subject to the one critical restriction that it be smooth and locally equivalent to Minkowski spacetime which means manifolds and therefore no singularities. This restriction encodes Einstein's equivalence principle.

However General Relativity predicts the formation of spacetime geometries which are not smooth manifolds such as collapses to black holes. Though General Relativity breaks down at the singularity, but it is fully applicable outside of it. This means that one can use it to explore the geometry inside and outside of the horizon which leads to all kinds of bizarre results and outstanding problems. Finally, one can use this tool to explore the whole universe leading to cosmology.