9 Gravity and the geometrization of physics

There exist excellent textbooks on Einstein's theory of gravity, ranging from non-technical introductions [329, 469, 187, 10, 231] to technically detailed ones [264, 367, 390, 55, 205, 414, 103, 548, 210, 131, 164, 66, 135, 96, 398, 506, 272, 315, 380, 342], as well as on tensor differential calculus in curved spaces [☞ [508, 62, 563, 210] to begin with], which is typically regarded as a prerequisite for a technical mastery of the material. The purpose of this chapter then cannot compete with these rich and detailed sources nor with textbooks on black holes and wormholes [103, 543], gravity in general and cosmology [418, 481, 419, 28, 558], and the interested Reader is wholeheartedly directed to this literature.

Complementing these resources, the general theory of relativity as a theory of (classical, i.e., non-quantum) gravity is here presented in comparison with Yang–Mills gauge theories from Chapters 5–8, thus continuing the unifying guiding idea that led us to this point; approaches to *quantum* gravity will be addressed in Chapter 11.

9.1 Einstein's equivalence principle and gauge symmetry

Most books that discuss general relativity and gravity – regardless of the technical level – start off with A. Einstein's principle of equivalence. Complementing this historically standard approach, gravity and general relativity may also be described and even "discovered" by (1) carefully examining the possible spacetime geometries as frameworks for real physical observations as done by R. Geroch [205]; (2) exploring the appearance and use of multi-valued fields [☞ magnetic monopole, Section 5.2.3] in a variety of physical models as done by H. Kleinert [315]; or (3) modeling the familiar gravitational and inertial phenomena from the point of view of a particle theory virtuoso as done by R. P. Feynman [164].

Borrowing from these approaches, we do start with Einstein's equivalence principle, but show that it is conceptually identical to the idea of gauge symmetry employed in Chapters 5 and 6. Thus it fits perfectly in the unifying "business card" of Nature, Table P.1 on xiii. Using the same concepts developed in Chapters 5 and 6, this lets us identify the analogue of the gauge vector potentials, construct a Lagrangian for them and derive Einstein's equation (9.44), below.

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

9.1.1 Inertial vs. gravitational mass

It was pointed out in Digression 8.1 that the full Lorentz transformations (including rotations and boosts; see Section 3.1.1) are a symmetry of the well-established Maxwell equations (5.72), while the Galilean group is a symmetry of Newtonian mechanics. While the Galilean group is the $c \to \infty$ limit of the Lorentz group, it is not a subgroup, and the two frameworks cannot be coherently combined, so as to describe the electrodynamics of moving electric charges. As is well known, the $c \to \infty$ limit of the Maxwell equations (so they would exhibit the Galilean group of symmetries) is unphysical: light propagates at finite speed. We are thus left with Nature's choice, relativistic physics.

The framework of relativistic physics, however, leaves a curious dichotomy regarding the concept of mass: On one hand, we have a simple mathematical result (3.36), which equates the Lorentz-invariant magnitude of its 4-momentum with the mass of an object, which is in turn identified (3.28)–(3.30) with the "inertial mass" familiar from non-relativistic mechanics. This mass is the ratio $m = \frac{|\vec{F}|}{|\vec{a}|}$, where $\vec{F}$ is a force applied to an object, $\vec{a}$ its resulting acceleration, where all observations are made in a coordinate system where the object was initially at rest, and we may even consider the limit where $\vec{F}$ and so also $\vec{a}$ are arbitrarily small.

On the other hand, in Newton's universal law of gravity, the mass of an object determines how strongly the gravitational attraction acts upon it – and there is no a-priori reason for this "gravitational mass" to be the same as the "inertial mass." That is to say, there remains the logical possibility that inertial effects upon an object may not be proportional to the same "mass" as are the gravitational effects, which is something Nature can – and does – decide for us: They are indeed one and the same.

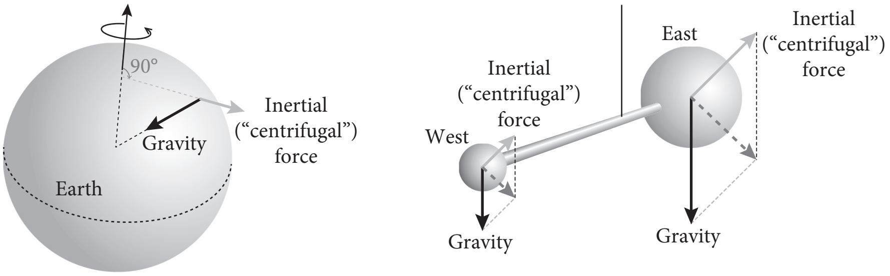

Figure 9.1 The classic Eötvös experiment: balancing the dumbbell horizontally compares the gravitational forces, while balancing it in the obtuse upward direction (grey arrows) compares inertial forces.

Experiments to this end have been carried out since around 1885, at first by Eötvös Loránd, where two substantial masses connected by a rod are balancing, suspended by a thin thread. The gravitational force acts towards the center of the Earth, while the inertial ("centrifugal") force due to Earth's rotation acts away from the axis of rotation, at an obtuse angle from the gravitational force. By aligning the horizontally balanced dumbbell initially in the east–west direction, all forces acting on each massive object are perpendicular to the connecting rod, and any difference in the sum of forces acting on one object vs. the other will produce a torque and twist the dumbbell from the initial east–west alignment. No matter what variety of the "eastern" and the "western" object in this torsion dumbbell were tried, the gravitational and the inertial forces were always found to be in the same proportion, thus proving the equality of the "inertial" and the "gravitational" mass, by now to the precision (relative error) of 10^{-11} [462].

Another logical possibility, that antiparticles [☞ Section 2.3.7] and particles *repel* each other by gravity, is easily dispelled in similarly high-precision experiments with elementary particles

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

such as the neutral kaons [☞ Section 4.2.3]: Since the decay eigenstates (4.65), $|K^0_S\rangle := \frac{1}{\sqrt{2}}(|K^0\rangle - |\overline{K}{}^0\rangle)$, and $|K^0_L\rangle := \frac{1}{\sqrt{2}}(|K^0\rangle + |\overline{K}{}^0\rangle)$, are linear combinations of the particle and its antiparticle and beams of K^0_S and K^0_L propagate in Earth’s gravitational field between creation and detection, a difference in the sign of the masses of K_0 and $\overline{K}_0$ would have to show. The experiments indeed do have the requisite precision, and indicate that K_0 and $\overline{K}_0$ have a positive (attracting) “gravitational” mass [164].

Between his seminal papers on special and general relativity, 1905–16, Einstein of course did not know about kaons, but must have been aware of the Eötvös-type experiment and its variations. He must have also been aware of the physically unnatural restriction to inertial coordinate systems within the special theory of relativity, as well as the fact that changes in the gravitational field could not propagate faster than the speed of light. To all of these issues, he came up with a single and elegant solution:

> ***Conclusion 9.1 (Principle of Equivalence)*** *Not only are the “inertial” and the “gravitating” masses equal, but inertial and gravitational physical effects are in fact* ***identical****.*

Tracing Einstein’s line of thought in the popular as well as most standard textbook presentations repeatedly brings up the example of a person in an enclosure such as an elevator with no windows. While at rest at the ground floor, the person in the elevator feels Earth-normal gravity. While the elevator accelerates upward, the inertial effect is added to the gravitational effect, and the person experiences an increase in their weight – which a scale will readily verify is quite real. During the constant motion between the floors, the weight experienced returns to Earth-normal. Finally, while the elevator decelerates when reaching the destination floor above, the person experiences a decrease in their weight. In fact, this much can be easily reasoned simply from Newton’s third law: the force measured by the scale on which the person in the elevator stands doesn’t care whether the reaction (with which it holds the person from falling through) balances the gravitational or the kinematic acceleration.

Extrapolating from these very familiar experiences, one can easily imagine a person within an enclosure, who would not be able to tell whether the experienced weight (or lack thereof) is a consequence of the gravitational force of some nearby planet, or the fact that the enclosure (perhaps a rocket ship) is moving in an accelerated fashion. Indeed, this is clearly true as long as the considered accelerated motion and related inertial forces and the gravitational forces are confined to one direction.

Even certain simple arrangements with additional forces and accelerations in additional directions easily permit such a dual interpretation. Consider for example a person at the North Pole, observing the motion of a so-called “spherical” pendulum, such as a bundle of keys attached to a keychain that the person holds firmly. With the Earth’s rotational axis passing through the person’s hand holding the keychain, the keys would be moving under the influence of three types of forces:

1. the gravitational force ($\vec{F}_g$), vertically downward to a very good approximation;
2. the horizontal “centrifugal” force ($\vec{F}_{cf}$), directed away from the axis of Earth’s rotation;
3. the horizontal Coriolis force ($\vec{F}_C$), at every instant perpendicular to both the axis of Earth’s rotation and to the direction of motion of the keys.

Exactly the same effects would be observed by a person in an accelerating rocket ship that additionally rotates about the direction of its linear motion – such “co-rotating” non-inertial coordinate systems were considered on p. 84, so as to exclude them from the Definition 3.1 on p. 84 of inertial coordinate systems; see the left-hand pair of illustrations in Figure 9.2.

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

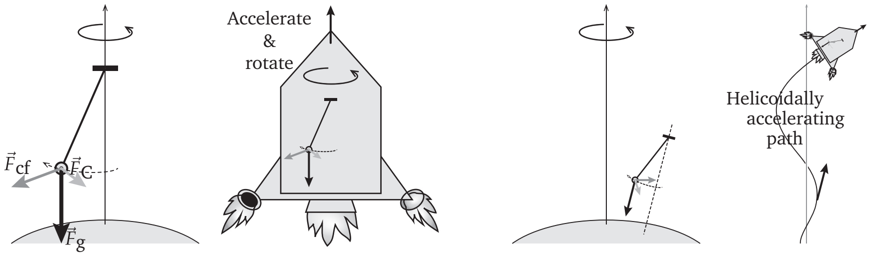

Figure 9.2 Two rotating pendula and the corresponding co-rotating accelerating coordinate systems.

Finding an appropriately accelerating coordinate system to be equivalent to an arrangement with more and more complicated systems of forces and accelerations of course becomes more and more complicated. For example, if the person with the swinging keychain were to move away from the North Pole, the direction of the gravitational acceleration would no longer coincide with the axis of Earth's rotation – as is the case with Foucault's pendulum in Paris, France. Effectively, the direction of gravitational acceleration for that person co-rotates about the axis of Earth's rotation, with which it also forms a nonzero angle. The corresponding accelerating coordinate system would then have to accelerate in a direction that forms the complementary angle with the Earth's axis of rotation, and precesses about it, thus accelerating along an expanding helicoidal path; see the right-hand pair of illustrations in Figure 9.2.

Any mechanical system under the influence of a *homogeneous* gravitational field is already perfectly equivalent in Newtonian mechanics to making the same mechanical system *uniformly* accelerate. Einstein's equivalence principle (Conclusion 9.1) is, however, fully general and applies to all physical phenomena, not just mechanics. W. Pauli then showed in his inimitable swift (and parsimonious) fashion, that this principle implies [414, Section 53]:

1. The influence of Newtonian (weak-field) gravity on a slowly moving object is determined by a scalar potential.
2. The gravitational field of stars causes a red shift in their spectral lines.
3. Even in a static gravitational field, light rays do not follow a geodesic in the 3-dimensional sense, but in the 4-dimensional spacetime sense: light rays are bent by gravity.

We will discuss the first of these results below, after introducing the requisite technical details.

9.1.2 Spacetime geometry and general coordinate transformations

As Geroch shows in detail [205, pp. 67–165], for every arrangement and scenario of particles moving in gravitational fields, there is a co-moving *spacetime geometry*. These are coordinate systems, each with four coordinates x^μ, $\mu = 0,1,2,3$, and a specified metric, $g_{\mu\nu}(\mathrm{x})$ of signature $(1,3)$; see Definition 3.2 on p. 89, we will explore some of the more interesting ones in some detail in Section 9.3. However, unlike in Chapter 3, these coordinates are inherently curvilinear in most applications, as should be clear from the example in the right-hand illustrations of Figure 9.2.

Away from certain exceptional locations (singularities) to be discussed in Section 9.3 and in sufficiently small regions (so-called patches) of spacetime, these inherently curvilinear coordinates can always be related to the Cartesian coordinates, much as every smooth curve can be approximated by its tangent. In Cartesian coordinates, the generalization of Pythagoras' theorem to spacetime [☞ relations (3.15)–(3.17)] defines the (spatial) so-called line element:

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

$$\mathrm{d}s^2 := -c^2\mathrm{d}\tau^2 = \mathrm{d}x^\mu(-\eta_{\mu\nu})\mathrm{d}x^\nu. \tag{9.1}$$

The relation (3.11c) then provides the expression in arbitrary coordinates $x^\mu \mapsto y^\mu = y^\mu(\mathrm{x})$:

$$\mathrm{d}s^2 := \mathrm{d}x^\mu(-\eta_{\mu\nu})\mathrm{d}x^\nu = \mathrm{d}y^\rho \underbrace{\Big(\frac{\partial x^\mu}{\partial y^\rho}\Big)(-\eta_{\mu\nu})\Big(\frac{\partial x^\nu}{\partial y^\sigma}\Big)}_{}\mathrm{d}y^\sigma = \mathrm{d}y^\rho\, g_{\rho\sigma}(\mathrm{y})\,\mathrm{d}y^\sigma, \tag{9.2}$$

$$g_{\rho\sigma}(\mathrm{y}) := \Big(\frac{\partial x^\mu}{\partial y^\rho}\Big)(-\eta_{\mu\nu})\Big(\frac{\partial x^\nu}{\partial y^\sigma}\Big), \quad \text{the metric tensor.} \tag{9.3}$$

> ***Comment 9.1*** *Note that the overall sign of the metric tensor (9.2) is opposite from the overall sign of the metric tensor (3.19). This unfortunate difference in conventions stems from the fact that the metric tensor (9.2) in general relativity defines a* ***distance****, while the expression (3.17) defines the proper* ***time*** *of a particle that moves in spacetime.*

The analogous computation for an *arbitrary* invertible coordinate substitution $y^\mu \to z^\mu(\mathrm{y})$ produces

$$g_{\mu\nu}(\mathrm{y}) = \frac{\partial z^\rho}{\partial y^\mu}\frac{\partial z^\sigma}{\partial y^\nu}\, g_{\rho\sigma}(\mathrm{z}), \tag{9.4}$$

proving that the metric tensor $g_{\mu\nu}$ is a rank-2, type-$(0,2)$ tensor.[1] More precisely, we define:

> ***Definition 9.1*** *Coordinate system transformations $x^\mu \to y^\mu(\mathrm{x})$ that are* **(1)** *unambiguously invertible, and* **(2)** *preserve the space/time character (signature) of spacetime [☞ Definitions 3.2 on p. 89 and 3.3 on p. 90] are* ***general coordinate transformations****.*

Unless otherwise stated, we only consider coordinate transformations that belong to this class.

Using the matrix notation, relation (9.4) may be written as

$$[\,g_{..}(\mathrm{x})\,] = [\,\frac{\partial \mathrm{z}}{\partial \mathrm{x}}\,][\,g_{..}(\mathrm{z})\,][\,\frac{\partial \mathrm{z}}{\partial \mathrm{x}}\,]^T, \tag{9.5}$$

where the superscript T denotes matrix transposition.[2] Computing the determinants produces

$$g(\mathrm{x}) = \Big(\det\Big[\frac{\partial \mathrm{z}}{\partial \mathrm{x}}\Big]\Big)^2 g(\mathrm{z}), \quad \text{where} \quad g(\mathrm{x}) := \det\big[g_{..}(\mathrm{x})\big]. \tag{9.6}$$

Since the metric tensor in spacetime has an odd number of negative eigenvalues,[3] it follows that the determinant of the metric tensor is negative, and

$$\sqrt{-g(\mathrm{x})} = \det\Big[\frac{\partial \mathrm{z}}{\partial \mathrm{x}}\Big]\sqrt{-g(\mathrm{z})} \tag{9.7}$$

[1] According to definition (9.2) of the quantity ds as a *distance* – which for purely spatial 4-vectors must agree with the familiar notion of the Euclidean distance – and owing to the "particle" convention (3.19) features the relative difference in the overall sign between $\eta_{\mu\nu}$ and $g_{\mu\nu}$: in flat spacetime, $g_{\mu\nu} \to -\eta_{\mu\nu}$. Both quantities are, however, called metric tensors, and the Reader is expected to read from the context which of the two conventions are used.

[2] The careful Reader will note that in the matrix representation of the components $g_{\rho\sigma}(\mathrm{z})$ one of the two indices must be counting rows while the other then must be counting columns. In the contraction with the matrices of partial derivatives in relation (9.4), the upper index (on the z-coordinate) in one of these two matrices must count columns (being contracted with the rows of $[g_{\rho\sigma}]$), but in the other it must count rows, whence the matrix representation of one of these matrices of partial derivatives is necessarily transposed in comparison with the other one.

[3] The general coordinate transformations, by Definition 9.1, preserve the signature, i.e., the numbers of positive and negative eigenvalues of the metric matrix.

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

is a real ***scalar density*** of ***weight*** -1. The weight of $\sqrt{-g}$ being -1 signifies that it transforms oppositely from the 4-fold differential (which then is a scalar density of weight $+1$):

$$\mathrm{d}^4\mathrm{x} = \det\left[\frac{\partial\mathrm{x}}{\partial\mathrm{y}}\right]\mathrm{d}^4\mathrm{y}, \tag{9.8}$$

which is computed straightforwardly (B.37) in Appendix B.2.1.

Conclusion 9.2 *The result (9.7) and the computation (B.37) in Appendix B.2.1 then imply that*

$$\sqrt{-g(\mathrm{x})}\,\mathrm{d}^4\mathrm{x} = \sqrt{-g(\mathrm{z})}\,\mathrm{d}^4\mathrm{z} \tag{9.9}$$

is an invariant with respect to the general coordinate transformations [☞ Definition 9.1 on p. 319], and provides the ***invariant (differential) 4-volume element****.*

Given the metric tensor $g_{\mu\nu}(\mathrm{y})$, the *inverse metric tensor* is defined by matrix inversion:

$$g^{\mu\nu}(\mathrm{y}) \;:\quad g^{\mu\nu}(\mathrm{y})\,g_{\nu\rho}(\mathrm{y}) \overset{!}{=} \delta^{\mu}_{\rho} \overset{!}{=} g_{\rho\nu}(\mathrm{y})\,g^{\rho\mu}(\mathrm{y}), \tag{9.10}$$

point-by-point $\mathrm{y} = (y^0, \dots, y^3)$ in spacetime. Since

$$0 = \partial_\sigma\big(\delta^{\mu}_{\rho}\big) = \partial_\sigma\big(g_{\rho\nu}\,g^{\rho\mu}\big) = \big(\partial_\sigma g_{\rho\nu}\big)g^{\rho\mu} + g_{\rho\nu}\big(\partial_\sigma g^{\rho\mu}\big), \tag{9.11}$$

it follows that

$$\big(\partial_\sigma g^{\lambda\mu}\big) = -g^{\rho\mu}g^{\lambda\nu}\big(\partial_\sigma g_{\rho\nu}\big). \tag{9.12}$$

In turn, derivatives of the determinant $g = \det[g_{..}]$ are computed using the Jacobi relation:

$$\partial_\rho g = g\,g^{\mu\nu}\,\partial_\rho g_{\mu\nu}, \tag{9.13}$$

from which it follows that

$$\partial_\rho\sqrt{-g} = -\tfrac{1}{2}\frac{\partial_\rho g}{\sqrt{-g}} = -\tfrac{1}{2}\sqrt{-g}\,\big(g_{\mu\nu}\,\partial_\rho g^{\mu\nu}\big) = \tfrac{1}{2}\sqrt{-g}\,\big(g^{\mu\nu}\,\partial_\rho g_{\mu\nu}\big). \tag{9.14}$$

For more detail, see Appendix B.2.3.

9.1.3 Einstein's equivalence principle as a gauge principle

Reconsider an object such as $\Psi(\mathrm{x})$, the wave-function used to describe an electron in Chapter 5. As discussed in detail in Section 5.1 and employed throughout Chapters 5–7, the (complex) function $\Psi(\mathrm{x})$ perforce contains unphysical information and is physically equivalent to $e^{i\varphi(\mathrm{x})}\Psi(\mathrm{x})$, where the phase $\varphi(\mathrm{x})$ is an undetermined function over spacetime. Consequently, the rate of change of $\Psi(\mathrm{x})$ in spacetime is computed not using partial derivatives, but using gauge-covariant derivatives (5.13), i.e., (5.117): $D_\mu := \partial_\mu + \frac{i}{\hbar c}A_\mu Q$. Since $\Psi(\mathrm{x})$ depends on the spacetime point both explicitly and also through the undetermined phase $\varphi(\mathrm{x})$, the partial derivative in D_μ computes the rate of change in spacetime owing to the explicit dependence on spacetime, while the $\frac{i}{\hbar c}A_\mu Q$ terms provides the "correction" owing to the indirect dependence via the undetermined phase, $\varphi(\mathrm{x})$.

The discussion in Section 9.1.1 showed that Einstein's principle of equivalence is itself equivalent to the statement that the difference between gravitational and inertial effects is purely a difference in the mathematical description, i.e., a difference in the *choice* of the coordinate system. Section 9.1.2 then formalizes the notion of spacetime geometry, as a spacetime coordinate system together with the corresponding metric, and *changing* this choice is accomplished by means of a general coordinate transformation.

As discussed by Pauli [414, p. 150], besides the technical aspects of the general coordinate transformations as formalized by tensor calculus, the key physical import of Einstein's principle of equivalence as provided in Conclusion 9.1 on p. 317 is its universal nature. That is, the equality of the various gravitational and inertial effects holds not only for certain (say, mechanical) phenomena, but for all physical phenomena. Therefore, there can be no physical distinction between them, and gravitational and inertial effects are not merely equal, but identical.

However, this insistence on universality is implied by the completely general (applicable to all of fundamental physics!) first, "conceptual" notion of unification as specified in part (a) of Conclusion 8.1 on p. 300. Under the umbrella of this overarching unifying principle, Einstein's equivalence principle is equivalent to

> ***Conclusion 9.3 (Gauge principle of coordinate equivalence)*** *General coordinate transformations [☞ Definition 9.1 on p. 319] can have no physically measurable consequences – and so must be symmetries [☞ Appendix A.1.3].*

In turn, this is conceptually identical to the gauge principle as employed in Chapters 5–7, except that the principle is here applied to the parametrization of spacetime, rather than to the abstract phases of wave-functions as in Chapters 5–7. Also, general coordinate transformations are typically nonlinear; this renders any gauge theory relating to general coordinate transformations intrinsically more complicated than the gauge theories considered in Chapters 5–7. We will explore the parallels and the differences between Yang–Mills gauge theory as discussed in Chapters 5–7 and general relativity, and will develop a selection of topics within general relativity specifically to that end. The Reader should, however, be aware of other possible approaches to gravity (some of them not entirely unrelated to the approach adopted herein), such as "gauge gravity" [451, 276] or "emergent gravity" [486, 315], to name a few.

Nevertheless, the *conceptual* similarity between the gauge principle as employed in Chapters 5–7 and the gauge equivalence principle (Conclusion 9.3 on p. 321) is striking:

1. Positions (in space of phases vs. in spacetime):
   (a) The choice of the overall phase of a wave-function is not observable; relative phases of different summands in a linear combination of wave-functions *are* observable.
   (b) The position of an object in spacetime is not observable; relative positions of different objects – distances between them – *are* observable.
2. Local (gauge) symmetry (changing the "position"):
   (a) Changing the choice of the overall phase of a wave-function locally in spacetime, i.e., by amounts that differ from point to point in spacetime.
   (b) Changing the choice of the coordinate system locally in spacetime, i.e., by (nonlinear) general coordinate transformations.
3. Gauge-covariant derivative operators (see below):
   (a) Correct the computation of the rate of change in spacetime to compensate for the spacetime variations in the choice of the undetermined phase.
   (b) Correct the computation of the rate of change in spacetime to compensate for the spacetime variations (nonlinearity) in the spacetime coordinate system itself.
4. Gauge interactions and curving trajectories (see below):
   (a) Gauge potentials and fields interact with test particles and curve their trajectories.
   (b) Spacetime is curved by the presence of matter, and curves the trajectories of test particles (including light).

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

9.1.4 Exercises for Section 9.1

✎ **9.1.1** *Show that, when y^μ are also Cartesian spacetime coordinates, the relation (9.3) implies that $g_{\rho\sigma}(\mathrm{y}) = -\eta_{\rho\sigma}$.*

✎ **9.1.2** *Show that, when both x^μ and y^μ are Cartesian spacetime coordinates, $\frac{\partial x^\mu}{\partial y^\rho}$ must be a Lorentz transformation as discussed in Section 3.1.1.*

✎ **9.1.3** *Prove (9.9).*

✎ **9.1.4** *Prove the result (9.14).*

9.2 Gravity vs. Yang–Mills interactions

Having identified in Section 9.1.2 the key elements by which tensor algebra as used in Chapter 3 generalizes to the general spacetime geometries (Appendix B.2 has more details), we turn to employing the gauge symmetry concept from Chapters 5–7 to general coordinate transformations. In particular, given a 4-tuple of contravariant components $A^\mu(\mathrm{x})$ of a vector field as well as a 4-tuple of covariant components $B_\mu(\,x)$ of another vector field, we quote the definition of the covariant derivatives:

$$\text{result (B.55):}\quad D_\mu A^\rho := \big[\partial_\mu A^\rho + \Gamma^\rho_{\mu\nu} A^\nu\big] \quad\text{and}\quad D_\mu B_\nu := \big[\partial_\mu B_\nu - \Gamma^\rho_{\mu\nu} B_\rho\big]. \tag{9.15}$$

As shown in Appendix B.2, the second term in these derivatives compensates for the fact that the frame of reference, i.e., system basis vectors in a curvilinear coordinate system, varies point-to-point in spacetime. They also ensure that these derivatives are covariant with respect to general coordinate transformations:

$$\big(\widetilde{D_\mu A^\rho}(\mathrm{y})\big) = \frac{\partial x^\nu}{\partial y^\mu}\frac{\partial y^\rho}{\partial x^\sigma}\big(D_\nu A^\sigma(\mathrm{x})\big) \quad\text{and}\quad \big(\widetilde{D_\mu B_\rho}(\mathrm{y})\big) = \frac{\partial x^\nu}{\partial y^\mu}\frac{\partial x^\sigma}{\partial y^\rho}\big(D_\nu B_\sigma(\mathrm{x})\big), \tag{9.16}$$

and covariant derivatives of vectors transform as rank-2 proper tensors. That is, these covariant derivatives behave with respect to general coordinate transformations *identically* as do the gauge-covariant derivatives (5.7), (5.117) and (6.6) with respect to the local (gauge) symmetry of Yang–Mills type models described in Chapters 5–7.

It should then present no surprise that the necessary introduction of the $\Gamma^\rho_{\mu\nu}$-dependent "correcting" terms in the covariant derivatives (9.15) – to accommodate for the spacetime variable coordinatization of the spacetime geometry – will result in a gauge interaction. Furthermore, the results (9.48)–(9.49) below will identify this interaction as gravity.

9.2.1 The metric connection and the Christoffel symbol

The formal characterization (B.66) of the covariant derivative is formally identical to the general form (5.10), i.e., (6.6); its action on a type-(p,q) tensor is given by the general relation owing to the definition (B.40):

$$(D_\mu\,\mathbb{T})^{\nu_1\cdots\nu_p}_{\rho_1\cdots\rho_q} = (\partial_\mu T^{\nu_1\cdots\nu_p}_{\rho_1\cdots\rho_q}) + \sum_{i=1}^{p}\Gamma^{\nu_i}_{\mu\sigma_i}\,T^{\nu_1\cdots\sigma_i\cdots\nu_p}_{\rho_1\cdots\cdots\cdots\rho_q} - \sum_{i=1}^{q}\Gamma^{\sigma_i}_{\mu\rho_i}\,T^{\nu_1\cdots\cdots\cdots\nu_p}_{\rho_1\cdots\sigma_i\cdots\rho_q}; \tag{9.17}$$

see also Appendix B.2.3. The well-known special cases are (9.15) and the rank-2 case:

$$(D_\mu\,\mathbb{T})_{\nu\rho} = \partial_\mu T_{\nu\rho} - \Gamma^\sigma_{\mu\nu}T_{\sigma\rho} - \Gamma^\sigma_{\mu\rho}T_{\rho\sigma}. \tag{9.18}$$

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

Notice: the precise index notation of the covariant derivative action on tensor densities depends on the rank and type of those tensor densities, as then also does the action of the Levi-Civita connection 4-vector $\mathbb{\Gamma}_\mu$, i.e., the Christoffel symbol $\Gamma^\rho_{\mu\nu}$.

It follows that the symbol $\Gamma^\rho_{\mu\nu}$ transforms *inhomogeneously* – and so is not a tensor:

$$\Gamma^\rho_{\mu\nu}(\mathrm{y}) = \frac{\partial x^\sigma}{\partial y^\mu}\frac{\partial x^\tau}{\partial y^\nu}\frac{\partial y^\rho}{\partial x^\kappa}\Gamma^\kappa_{\sigma\tau}(\mathrm{x}) + \frac{\partial y^\rho}{\partial x^\sigma}\frac{\partial^2 x^\sigma}{\partial y^\mu \partial y^\nu}, \tag{9.19}$$

exactly as in the case of gauge 4-vector potentials in the (abelian) electrodynamics (5.89) and non-abelian chromodynamics (6.6b). At a first glance, the inhomogeneous term in the expressions (5.89) and (6.6b) is proportional to $(\partial_\mu \boldsymbol{\varphi}) = (\partial_\mu U)U^{-1}$, which may seem different from the second term in the result (9.19). However, using the matrix notation

$$[U]^\rho{}_\sigma = \frac{\partial y^\rho}{\partial x^\sigma}, \qquad \text{we have} \qquad \frac{\partial y^\rho}{\partial x^\sigma}\frac{\partial^2 x^\sigma}{\partial y^\mu \partial y^\nu} = [U]^\rho{}_\sigma \frac{\partial}{\partial y^\mu}[U^{-1}]^\sigma{}_\nu, \tag{9.20}$$

which then fully agrees with $(\partial_\mu \boldsymbol{\varphi}) = (\partial_\mu U)U^{-1} = -U(\partial_\mu U^{-1})$, up to a conventional sign of the phase “angle” $\boldsymbol{\varphi}$.

> ***Comment 9.2*** *The transformations $U = \left[\frac{\partial \mathrm{y}}{\partial \mathrm{x}}\right]$ employed here are general coordinate transformations [☞ Definition 9.1 on p. 319], which form a (gauge) group only in a restricted sense.[4] The physical manifestations of the theory in which $\mathbb{\Gamma}_\mu$ is the gauge potential and U the gauge transformation will be identified below as gravity; see equations (9.48)–(9.49).*

One may also construct the so-called the connection (differential) 1-forms[5]

$$\mathbb{A} := \mathrm{d}x^\mu \mathbb{A}_\mu, \qquad \text{i.e.,} \qquad \mathbb{\Gamma} := \mathrm{d}x^\mu \mathbb{\Gamma}_\mu. \tag{9.21}$$

Since $\mathbb{A} = \mathrm{d}x^\mu A^a_\mu Q_a$ and Q_a are elements of the *algebra* of the gauge group, one says that $\mathbb{A}$ is valued in the gauge algebra. Similarly, $\mathbb{\Gamma}$ is a differential 1-form with values in the algebra of the group of transformations (B.41); the covariant differential $\mathrm{d}x^\mu D_\mu$ is also-called the Levi-Civita connection.

> ***Conclusion 9.4*** *As the algebra of a group is essentially specified by linearizing (A.9), it follows that $\mathbb{\Gamma}$ may be regarded as a differential 1-form that takes values in the algebra of transformations of the tangent 4-plane (at any given spacetime point) into itself, which is the algebra of the Lorentz group, $Spin(1,3)$. Although no spinor appears in this discussion, the Lorentz group of course must act unambiguously on spinors also, whereupon we write $Spin(1,3)$ instead of $SO(1,3)$ [☞ discussion about relations (5.45)–(5.48)].*

However, note the difference: For the Yang–Mills gauge symmetries in Chapters 5–7, the unitary operator of the symmetry transformation, $U := \exp\{i g_c \varphi^a(\mathrm{x}) Q_a/\hbar\}$, depends on (the coordinates of) the spacetime point $\mathrm{x} = (x^0, \dots, x^3)$ but describes a change in parametrizing another, abstract space of generalized phases of wave-functions. Within our present context, $[U]^\rho{}_\sigma = \frac{\partial y^\rho}{\partial x^\sigma}$ depends on the spacetime point x, but simultaneously describes the change in the coordinate parametrization (basis elements) of that *very same* spacetime. Besides, the coordinate transformations $x^\mu \rightarrow y^\nu = y^\nu(\mathrm{x})$ are nonlinear in general. This conceptual as well as literal nonlinearity

[4] The binary combination of two transformations exists only when they “concatenate”: $\frac{\partial x^\mu}{\partial y^\nu}\frac{\partial z^\rho}{\partial x^\mu} = \frac{\partial z^\rho}{\partial y^\nu}$ and $\frac{\partial x^\mu}{\partial y^\nu}\frac{\partial y^\nu}{\partial z^\rho} = \frac{\partial x^\mu}{\partial z^\rho}$, but a product such as $\frac{\partial x^\mu}{\partial y^\nu}\frac{\partial z^\nu}{\partial w^\rho}$ does not simplify as a closed binary operation. This structure curiously reminds us of the so-called “renormalization group,” see Section 5.3.3 on p. 210☝.

[5] Instead of $\mathrm{d}x^\mu$, one may of course use any arbitrary basis elements, e^μ, resulting also in 1-forms, albeit not differential. The use of the $\mathrm{d}x^\mu$-basis is however standard, as it provides a connection with differential and integral calculus.

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

provides the root of all differences between (Yang–Mills) gauge theories and the general theory of relativity, viewed as a gauge theory.

This difference also reflects in the following: The gauge vector potential (6.6c) has a matrix representation:

$$\mathbb{A}_\mu := A_\mu^a\, Q_a \qquad \to \qquad [\mathbb{A}_\mu]_\alpha{}^\beta. \tag{9.22}$$

The gauge vector potential for general coordinate transformations (9.20) is the Levi-Civita connection 4-vector,

$$\mathbb{\Gamma}_\mu \qquad \to \qquad [\mathbb{\Gamma}_\mu]_\nu{}^\rho, \tag{9.23}$$

that acts upon a vector according to the relations (9.15), in perfect analogy with the action of the chromodynamics gauge vector potential (9.22) upon a quark wave-function:

$$[\mathbb{A}_\mu \cdot \Psi]^\alpha = [\mathbb{A}_\mu]_\beta{}^\alpha\, \Psi^\beta \qquad \leftrightarrow \qquad [\mathbb{\Gamma}_\mu \cdot V]^\rho = \Gamma^\rho_{\mu\,\nu}\, V^\nu. \tag{9.24}$$

Note, however, that the chromodynamics gauge potentials are matrices in the abstract space of (color) phases and covariant vectors in real spacetime. By contrast, the Christoffel symbol is a matrix in the very same spacetime wherein it is also a connection 4-vector. What is more, it is not hard to show that (see, e.g., the derivation of (B.59) in Appendix B.2.3)

$$\Gamma^\rho_{\mu\nu} = \tfrac{1}{2} g^{\rho\sigma} \Big[\frac{\partial g_{\sigma\nu}}{\partial x^\mu} + \frac{\partial g_{\mu\sigma}}{\partial x^\nu} - \frac{\partial g_{\mu\nu}}{\partial x^\sigma} \Big]. \tag{9.25}$$

That is, the gauge potential for general coordinate transformations, the Levi-Civita connection 4-vector $\mathbb{\Gamma}_\mu$, can be derived from the metric tensor (9.3),[6] which thereby serves as a gauge "*pre*-potential." In Yang–Mills gauge theories, no such thing exists.

In turn, relation (9.25) is equivalent to the result

$$D_\mu\, g_{\nu\rho} = 0 \qquad \Leftrightarrow \qquad D_\mu\, g^{\nu\rho} = 0. \tag{9.26}$$

That is, the metric tensor and its inverse are "covariantly constant," so (9.25) may just as well be derived from either of the two relations (9.26). Again, Yang–Mills gauge theories contain no such nontrivial "covariantly constant" object.

Thus, while the electric and magnetic fields may be obtained as derivatives of an electromagnetic potential (5.15)–(5.73) A_μ, this potential cannot be obtained as a derivative of some more fundamental prepotential. Similarly, chromodynamics fields $\mathbb{F}_{\mu\nu} = F^a_{\mu\nu} Q_a$ can also be expressed in terms of a chromodynamics potential (6.15) $\mathbb{A}_\mu = A^a_\mu Q_a$, but these potentials cannot be expressed in terms of something more fundamental yet. In sharp contrast, the Christoffel symbol $\Gamma^\rho_{\mu\nu}$ *may be and is* expressed in terms of a derivative of the metric tensor (9.25) and the inverse metric tensor. From relations (9.25) it also follows that the Christoffel symbol is symmetric with respect to the exchange of the indices

$$\Gamma^\rho_{\mu\nu} = +\Gamma^\rho_{\nu\mu}. \tag{9.27}$$

In the Yang–Mills gauge vector potentials $[A_\mu]_\alpha{}^\mu$, an analogous symmetrization (here, for $\mu \leftrightarrow \alpha$) simply makes no sense at all: μ and α indicate basis elements in completely different spaces.

[6] Strictly speaking, this is true only in the absence of fermions. With fermions present, one uses the so-called Palatini formalism, wherein the metric tensor and the Levi-Civita connection 4-vector $\mathbb{\Gamma}_\mu$ are independent.

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

Digression 9.1 Some useful consequences of the relations (9.25)–(9.26) are

$$\frac{\partial g_{\mu\nu}}{\partial x^\sigma} = \Gamma^\rho_{\mu\sigma} g_{\rho\nu} + \Gamma^\rho_{\nu\sigma} g_{\rho\mu}, \qquad \frac{\partial g^{\mu\nu}}{\partial x^\sigma} = -g^{\mu\rho}\Gamma^\nu_{\sigma\rho} - g^{\nu\rho}\Gamma^\mu_{\sigma\rho}, \tag{9.28a}$$

$$\Gamma^\mu_{\mu\nu} = \frac{\partial}{\partial x^\nu} \ln\big(\sqrt{-g}\big), \qquad g := \det[g_{..}]; \quad g < 0 \text{ because of signature } (1,3), \tag{9.28b}$$

where we used the relation

$$\frac{\partial g}{\partial g_{\mu\nu}} = g\, g^{\mu\nu}, \quad \text{so that} \quad \frac{\partial g}{\partial x^\rho} = g\, g^{\mu\nu} \frac{\partial g_{\mu\nu}}{\partial x^\rho}. \tag{9.28c}$$

The signature is the number of positive and negative eigenvalues of the metric tensor [☞ discussion about the expression (3.19) and Definition 3.3 on p. 90].

Digression 9.2 Also, definition (9.17) produces the following oft-used results:

$$\text{grad}(f)_\mu := D_\mu f = (\partial_\mu f); \tag{9.29a}$$

$$\text{curl}(V_{.})^{\rho\sigma} := \varepsilon^{\mu\nu\rho\sigma} D_\nu V_\mu = \varepsilon^{\mu\nu\rho\sigma}(\partial_\nu V_\mu); \tag{9.29b}$$

$$\text{curl}(V^{\cdot})^{\rho\sigma} := \varepsilon^{\mu\nu\rho\sigma} D_\mu (g_{\nu\lambda} V^\lambda) = \varepsilon^{\mu\nu\rho\sigma}\partial_\mu (g_{\nu\lambda} V^\lambda); \tag{9.29c}$$

$$\text{div}(V^{\cdot}) := D_\mu V^\mu = \frac{1}{\sqrt{-g}}\big(\partial_\mu(\sqrt{-g}\, V^\mu)\big); \tag{9.29d}$$

$$\text{div}(V_{.}) := D_\mu (g^{\mu\nu} V_\nu) = \frac{1}{\sqrt{-g}}\big(\partial_\mu(\sqrt{-g}\, g^{\mu\nu} V_\nu)\big); \tag{9.29e}$$

$$\Box f := D_\mu (g^{\mu\nu} D_\nu f) = \frac{1}{\sqrt{-g}}\Big[\partial_\mu\big(\sqrt{-g}\, g^{\mu\nu}(\partial_\nu f)\big)\Big]. \tag{9.29f}$$

Note that, in 1+3-dimensional spacetime, the curl of a 4-vector is a rank-2 tensor. On the other hand, the spacetime analogue of $\vec{\nabla}^2 \vec{A} \equiv \vec{\nabla}(\vec{\nabla}\cdot\vec{A}) - \vec{\nabla}\times(\vec{\nabla}\times\vec{A})$ may be used to compute

$$\Box A^\mu = \left[g^{\mu\nu}\partial_\nu\left(\frac{\partial_\rho\big(\sqrt{-g}\, A^\rho\big)}{\sqrt{-g}}\right) + \frac{1}{\sqrt{-g}}\varepsilon^{\mu\nu\rho\sigma}\varepsilon^{\alpha\beta\kappa\lambda}\partial_\nu\left(\frac{(\partial_\alpha A_\beta)}{\sqrt{-g}}\, g_{\kappa\rho} g_{\lambda\sigma}\right)\right]. \tag{9.29g}$$

9.2.2 The curvature of spacetime

Finally, just as the gauge field $\mathbb{F}_{\mu\nu}$ is defined in relation (6.15) as the commutator of covariant derivatives, so too may the Riemann curvature tensor be defined:

$$\begin{aligned} R_{\mu\nu\rho}{}^\sigma &:= [\, D_\mu\, ,\, D_\nu\,]_\rho{}^\sigma = \big[(\delta^\sigma_\lambda \partial_\nu + \Gamma^\sigma_{\nu\lambda})\Gamma^\lambda_{\mu\rho}\big] - \big[(\delta^\sigma_\lambda \partial_\mu + \Gamma^\sigma_{\mu\lambda})\Gamma^\lambda_{\nu\rho}\big] \\ &= \partial_\nu \Gamma^\sigma_{\mu\rho} - \partial_\mu \Gamma^\sigma_{\nu\rho} + \Gamma^\sigma_{\nu\lambda}\Gamma^\lambda_{\mu\rho} - \Gamma^\sigma_{\mu\lambda}\Gamma^\lambda_{\nu\rho}. \end{aligned} \tag{9.30}$$

Note the formal similarity of the defining expression (9.30) and the definition of the gauge field for non-abelian gauge symmetry (6.15). However, unlike $\mathbb{F}_{\mu\nu}$ which is an antisymmetric rank-2 tensor and the components of which are matrices in the abstract space of phases, the Riemann tensor is a rank-4, type-$(1,3)$ tensor. Besides, it may be shown that [508, 62, 367, 548, 66, 96]

$$R_{\mu\nu\rho}{}^\rho = 0, \tag{9.31}$$

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

and that the closely related tensor

$$R_{\mu\nu\,\rho\sigma} := R_{\mu\nu\rho}{}^{\lambda}\, g_{\lambda\sigma} \tag{9.32a}$$

satisfies the relations:

$$\begin{aligned}
R_{\mu\nu\,\rho\sigma} &= -R_{\nu\mu\,\rho\sigma}, && && (9.32\text{b})\\
R_{\mu\nu\,\rho\sigma} &= -R_{\mu\nu\,\sigma\rho}, && && (9.32\text{c})\\
R_{\mu\nu\,\rho\sigma} &= +R_{\rho\sigma\,\mu\nu}, && && (9.32\text{d})\\
\varepsilon^{\lambda\nu\rho\sigma} R_{\mu\nu\,\rho\sigma} &= 0, && \text{1st Bianchi identity,} && (9.32\text{e})\\
\varepsilon^{\kappa\lambda\mu\nu} D_{\lambda} R_{\mu\nu\,\rho\sigma} &= 0, && \text{2nd Bianchi identity.} && (9.32\text{f})
\end{aligned}$$

This 2nd Bianchi identity (9.32f) is both formally and conceptually analogous to the Bianchi identity (5.87) in electrodynamics and (6.19) for non-abelian gauge fields.

Relation (9.31) is analogous to the requirement that in the expansion $\mathbb{F}_{\mu\nu} = F^{a}_{\mu\nu}\, Q_a$, the generators Q_a of non-abelian factors in the gauge group are traceless: $\mathrm{Tr}[Q_a] = [Q_a]_{\alpha}{}^{\alpha} = 0$. This is certainly true of the gauge field of the $SU(3) \times SU(2)_w$ group, and is not true precisely for the *abelian* electromagnetic $U(1)$ field $F_{\mu\nu}$. The Riemann tensor $R_{\mu\nu\rho}{}^{\sigma}$ may be regarded as a special rank-2 and type-$(0,2)$ tensor, the components of which are matrices and traceless rank-2 and type-$(1,1)$ tensors, $R_{\mu\nu\rho}{}^{\sigma} = [R_{\mu\nu}]_{\rho}{}^{\sigma}$, subject to the additional constraints (9.32b)–(9.32f). The fact that both $\mathbb{F}_{\mu\nu}$ and $R_{\mu\nu\rho}{}^{\sigma}$ are defined as commutators of appropriate covariant derivatives then guarantees the first of the relations, (9.32b). This similarity permits the interpretation of the Riemann tensor as a general coordinate transformation analogue of the tensor $\mathbb{F}_{\mu\nu}$. The components $R_{\mu\nu\rho}{}^{\sigma}$ are then interaction fields associated with general coordinate transformations, and in fact represent the general-relativistic generalization of the gravitational field; see below.

The very existence of the definition (9.32a) points to the difference between $R_{\mu\nu\rho}{}^{\sigma}$ and $[F_{\mu\nu}]\alpha^{\beta}$. For orthogonal and symplectic gauge groups,[7] their invariant quadratic forms would play the role of $g_{\lambda\sigma}$ and produce $[F_{\mu\nu}]_{\alpha\beta}$. Unitary groups (such as $SU(3)_c$) have no such tensor, and for them there can exist nothing analogous to definition (9.32a). Also, for unitary gauge groups there exist no analogues of the relations (9.32c)–(9.32e).

Furthermore, for Yang–Mills gauge fields, $[\mathbb{F}_{\mu\nu}]_{\alpha}{}^{\beta}$, there is no way to perform the contraction between one of the "matrix" indices α or β and one of the "tensor" indices μ or ν. In turn, the contractions that can be performed,

$$g^{\mu\nu}\mathbb{F}_{\mu\nu} \equiv 0, \qquad \left\{\begin{array}{lll} \mathrm{Tr}[\mathbb{F}_{\mu\nu}] & = [\mathbb{F}_{\mu\nu}]_{\alpha}{}^{\alpha} = 0, & \text{for semisimple Lie groups,}\\ \mathrm{Tr}[F_{\mu\nu}] & = F_{\mu\nu}, & \text{for } U(1) \text{ factors,} \end{array}\right. \tag{9.33}$$

are trivial: The first equality holds owing to the fact that $g_{\mu\nu} = +\, g_{\nu\mu}$ but $\mathbb{F}_{\mu\nu} = -\,\mathbb{F}_{\nu\mu}$. The second one follows from the fact that $\mathrm{Tr}[Q_a] \neq 0$ only for $U(1)$ factors.

The situation is, however, different for the Riemann tensor: neither is

$$\text{the Ricci tensor:} \qquad R_{\mu\rho} := R_{\mu\nu\rho}{}^{\nu}, \tag{9.34}$$

trivial, nor is its trace,

$$\text{the scalar curvature:} \qquad R := g^{\mu\rho}\, R_{\mu\rho} = g^{\mu\rho}\, R_{\mu\nu\rho}{}^{\nu}. \tag{9.35}$$

[7] Orthogonal and symplectic groups may be defined as the groups of linear transformations of some specified real vector space that preserve a (pseudo-)Euclidean, i.e., symplectic quadratic form, respectively [☞ Appendix A]. However, this invariant quadratic form does not determine the gauge potential of Yang–Mills theories with orthogonal and symplectic group of symmetries, unlike the fact that the relation (9.25) does determine the Christoffel symbol in terms of the metric.

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

It is also useful to know that, following Conclusion 9.4 on p. 323, we have that the differential 2-form[8]

$$\mathbb{R} := \big[\, \mathrm{d}x^\mu D_\mu\,,\, \mathrm{d}x^\nu D_\nu \,\big], \qquad \text{i.e.,} \qquad [\mathbb{R}]_\rho{}^\sigma := \mathrm{d}x^\mu\, \mathrm{d}x^\nu\, R_{\mu\nu\rho}{}^\sigma \tag{9.36}$$

also has values in the algebra of the Lorentz group *Spin*$(1,3)$.

Definition (9.30) shows that the components of the Riemann tensor $R_{\mu\nu\rho}{}^\sigma$ are derivatives of the second order (or are quadratic in derivatives of the first order) of the metric tensor components,[9] but it contains also the inverse metric tensor. $R_{\mu\nu\rho}{}^\sigma$ is therefore a nonlinear function of the metric tensor components, $g_{\mu\nu}$, but precisely of second order in spacetime derivatives of those components.[10] The same is then true also of the Ricci tensor (9.34), as well as the scalar curvature (9.35).

Yang–Mills gauge theories have nothing analogous to the expressions (9.34)–(9.35). There, the Lagrangian density (6.23) is found in the form $-\frac{1}{4}\,\mathrm{Tr}[\mathbb{F}_{\mu\nu}\,\mathbb{F}^{\mu\nu}]$, which is quadratic in the derivatives of $\mathbb{A}_\mu$. This Lagrangian density then yields equations of motion (6.24) that are analogous to Gauss's law for the electric field and Ampère's law for the electromagnetic field (6.37).

Analogously to the expression $-\frac{1}{4}\,\mathrm{Tr}[\mathbb{F}_{\mu\nu}\mathbb{F}^{\mu\nu}]$ in the Lagrangian density (6.23), the Hamilton action with the Riemann tensor would be proportional to the integral

$$\int \sqrt{-g}\, \mathrm{d}^4\mathrm{x}\; R_{\mu\nu\rho}{}^\sigma\, g^{\mu\kappa} g^{\nu\lambda}\, R_{\kappa\lambda\sigma}{}^\rho. \tag{9.37}$$

Since both $\sqrt{-g}\,\mathrm{d}^4\mathrm{x}$ and $R_{\mu\nu\rho}{}^\sigma\, g^{\mu\kappa} g^{\nu\lambda}\, R_{\kappa\lambda\sigma}{}^\rho$ are scalar quantities, this integral is invariant under general coordinate transformations. Varying this action by the components of the Christoffel symbol would, in the standard fashion, produce Euler–Lagrange equations of the second order in derivatives of the Christoffel symbol, $\mathbb{\Gamma}$. However, the Christoffel symbol is itself a derivative of the metric tensor, and varying this action by components of the metric tensor (which is more fundamental than the Christoffel symbol) would produce Euler–Lagrange equations of motion for the metric tensor components that are of the *fourth order* in spacetime derivatives, which agrees with neither classical (non-quantum) theory of gravity nor with experimental facts about gravity.

Fortunately – and completely unlike in Yang–Mills gauge theory – with the Riemann tensor it is possible to define another, so-called Einstein–Hilbert action:

$$\frac{c^3}{16\pi\, G_N} \int \sqrt{-g}\, \mathrm{d}^4\mathrm{x}\; R, \qquad \text{where } R \overset{(9.35)}{=} g^{\mu\rho}\, R_{\mu\nu\rho}{}^\nu. \tag{9.38}$$

The powers of the natural constants $c, \hbar$ and G_N in the prefactor are determined:

1. by requiring the Hamilton action to have the dimensions $\frac{ML^2}{T}$ [☞ Sections 1.2.3 and 1.2.2],
2. by definition (3.10) whereby $[\mathrm{d}^4\mathrm{x}] = L^4$ (note: $\mathrm{d}^4\mathrm{x} = c\,\mathrm{d}t\, \mathrm{d}^3\vec{r}$),[11]
3. by definitions (9.2), (9.25) and (9.30), from which it follows that $[g_{\mu\nu}] = 1$, $[\Gamma^\rho_{\mu\nu}] = L^{-1}$ and $[R_{\mu\nu\rho}{}^\sigma] = L^{-2}$, respectively.

The conventional numerical prefactor $\frac{1}{16\pi}$ simplifies many derivations and many final results. Varying this action by the metric tensor components produces [508, 62, 367, 548, 66, 96]

[8] When defining differential p-forms, one automatically uses the antisymmetric product of basis elements and without any notational distinction: $(\cdots \mathrm{d}x^\mu \mathrm{d}x^\nu \cdots) = -(\cdots \mathrm{d}x^\nu \mathrm{d}x^\mu \cdots)$.

[9] All told, every summand in the defining expression (9.30) contains precisely two spacetime derivatives.

[10] Unlike the quadratic, cubic or another expression of a relatively low degree, the components of the inverse metric tensor are by definition ratios of the determinants of various cofactors and the determinant of the entire metric tensor. A Taylor expansion in the components of the original metric tensor is then an infinite series, containing arbitrarily high powers of the components of the original metric tensor. This makes the inverse metric tensor, and then also the Riemann and other curvature tensors, *very* nonlinear.

[11] Some Authors imply $\mathrm{d}^4\mathrm{x} := \mathrm{d}t\, \mathrm{d}^3\vec{r}$, so that the prefactor in the action (9.38) has c^4 instead of c^3 as given here.

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

$$R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R = 0. \tag{9.39}$$

This system of differential equations, the Einstein equations, determines the metric tensor components as functions of the spacetime coordinates, and in the absence of all matter, i.e., in empty space. The combination $G_{\mu\nu} := R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R$ is called the Einstein tensor.

Already, writing the Einstein equations (9.39), with definitions (9.30) and (9.25), indicates the essential differences from Yang–Mills gauge theories: The differential equations (6.37) are at most cubic in the 4-vector potentials $\mathbb{A}_\mu$, while the Einstein equations (9.39) are *very* nonlinear in the metric tensor components. The definition of the Christoffel symbol and the scalar curvature involve the inverse metric tensor, the components of which are ratios of cubic polynomials in the components $g_{\mu\nu}$ and the determinant $\det[g_{\mu\nu}]$. This much more radical nonlinearity of the differential equations (9.39) – and also the action (9.38) from which the Einstein equations follow – is the root of the technical differences between the general theory of relativity and Yang–Mills gauge theories.

9.2.3 Coupling of gravity and matter

Finally, the operations so far defined may be combined and produce a relevant result for our present purposes:

Conclusion 9.5 *In the general case, Hamilton's action is*

$$S[\phi_i(\mathrm{x})] := \int \sqrt{-g}\, \mathrm{d}^4\mathrm{x}\, \mathscr{L}\big(\phi_i(\mathrm{x}), (D_\mu \phi_i(\mathrm{x})), \dots; \mathrm{x}\big), \tag{9.40}$$

$$g := \det[\boldsymbol{g}(\mathrm{x})], \quad \mathrm{d}^4\mathrm{x} := \tfrac{1}{4!} \varepsilon_{\mu\nu\rho\sigma} \mathrm{d}x^\mu \mathrm{d}x^\nu \mathrm{d}x^\rho \mathrm{d}x^\sigma, \tag{9.41}$$

where $\mathscr{L}$ is the "Lagrangian density" (in the sense of "Lagrangian per unit 4-volume"). In turn, both $\sqrt{-g}\,\mathrm{d}^4\mathrm{x}$ and $\mathscr{L}$ are scalars, i.e., invariants with respect to general coordinate transformations [☞ Definition 9.1 on p. 319].

Comment 9.3 *Lagrangian densities $\mathscr{L}\big(\phi_i(\mathrm{x}), (\partial_\mu \phi_i(\mathrm{x})), \dots; \mathrm{x}\big)$ constructed within the special-relativistic field theory may continue to be used, but "covariantizing" the derivatives, $\partial_\mu \mapsto D_\mu := \partial_\mu + \mathbb{\Gamma}_\mu$, where $\mathbb{\Gamma}_\mu$ is the formal Levi-Civita* ***connection*** *4-vector, which when acting on tensors may be represented by the Christoffel symbol (9.17).*

In the general case, the covariant derivative is $D_\mu = \partial_\mu + \mathbb{\Gamma}_\mu + \sum_k \frac{ig_k}{\hbar c} \boldsymbol{A}^{(k)}_\mu \cdot \boldsymbol{Q}^{(k)}$, where $\boldsymbol{Q}^{(k)}_{a_k}$ are generators of the kth factor in the Yang–Mills group of gauge symmetries with the coupling parameter g_k and gauge 4-vector potentials $A^{(k)}_\mu{}^{a_k}$.

In the general case, let $\mathscr{L}_M$ be the Lorentz-invariant Lagrangian density for any type of matter – here, "matter" denotes everything except the metric tensor $g_{\mu\nu}$, the Levi-Civita connection 4-vector potential $\mathbb{\Gamma}_\mu$, and the Riemann tensor $R_{\mu\nu\rho}{}^\sigma$ and quantities constructed from these. The corresponding model that is invariant with respect to general coordinate transformations has the Hamilton action

$$\int \sqrt{-g}\, \mathrm{d}^4\mathrm{x} \left[\frac{c^3}{16\pi\, G_N} R - \mathscr{L}_M\right], \tag{9.42}$$

where all the derivatives in the Lagrangian density $\mathscr{L}_M$ are "covariantized" as discussed in Comment 9.3 on p. 328. Varying this action by the components of the inverse metric tensor yields

$$\frac{\delta R}{\delta g^{\mu\nu}} + \frac{R}{\sqrt{-g}} \frac{\delta(\sqrt{-g})}{\delta g^{\mu\nu}} = -\frac{16\pi\, G_N}{c^3} \frac{1}{\sqrt{-g}} \frac{\delta(\sqrt{-g}\,\mathscr{L}_M)}{\delta g^{\mu\nu}}, \tag{9.43}$$

that is [508, 62, 367, 548, 66, 96],

$$R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R = \frac{8\pi\, G_N}{c^4} T_{\mu\nu}, \tag{9.44}$$

where the rank-2 and type-$(0,2)$ tensor

$$T_{\mu\nu} := -\frac{2c}{\sqrt{-g}}\,\frac{\delta(\sqrt{-g}\,\mathscr{L}_M)}{\delta g^{\mu\nu}} \tag{9.45}$$

has the physical meaning of the energy–momentum tensor density for the physical system described by the Lagrangian density $\mathscr{L}_M$.

Digression 9.3 Note that the inverse metric tensor and the metric tensor of course are not independent quantities, since the inverse metric tensor is defined so as to satisfy

$$g_{\mu\nu}\,g^{\rho\nu} = \delta^\rho_\mu, \qquad g_{\mu\nu} = +g_{\nu\mu} \quad\Rightarrow\quad g^{\mu\nu} = +g^{\nu\mu}. \tag{9.46a}$$

It then follows that varying the inverse metric tensor is not independent of varying the metric tensor itself:

$$0 = \delta(\delta^\rho_\mu) = \delta(g_{\mu\nu}\,g^{\rho\nu}), \tag{9.46b}$$

$$\Rightarrow \qquad \delta g^{\mu\nu} = -g^{\mu\rho}g^{\nu\sigma}\,(\delta g_{\rho\sigma}), \quad \text{and} \quad \frac{\delta}{\delta g^{\mu\nu}} = -g_{\mu\rho}g_{\nu\sigma}\,\frac{\delta}{\delta g_{\rho\sigma}}. \tag{9.46c}$$

Varying the action (9.42) by various fields that represent various "matter" degrees of freedoms produces the Euler–Lagrange equations of motion for these fields. As all the derivatives in the Lagrangian density $\mathscr{L}_M$ are covariantized, the resulting Euler–Lagrange equations of motion will, in the general case, depend on the Levi-Civita connection 4-vector $\mathbb{\Gamma}_\mu$ as well as on the metric $g_{\mu\nu}$. The Euler–Lagrange equations of motion and the Einstein equations (9.44) then form a coupled system of differential equations, which are certainly nonlinear in the metric tensor components.

Although such coupled systems of differential equations most often are not soluble in closed form, the geometric meaning of the Einstein equations (9.44) is very clear:

1. On the left-hand side, $R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R$ is a nonlinear expression in the metric tensor components, which is of precisely second order in spacetime derivatives; the left-hand side depends only on the metric tensor components and their spacetime derivatives.
2. On the right-hand side, $T_{\mu\nu}$ is the energy–momentum tensor density, which describes the spacetime (and general-relativistic) generalization of mass of the matter.

The differential equation (9.44) thus determines the metric tensor, for which the energy–momentum tensor density plays the role of the "source" – just as the differential equation representing Gauss's law determines the electric field for which electric charge density plays the role of the source, and Ampère's law determines the electromagnetic field for which the electric current density plays the role of the source.

What's more, comparing the Einstein equations with the differential equations representing the Gauss–Ampère laws is more than suggestive: it may be shown that the energy–momentum tensor density, $T_{\mu\nu}$, is indeed the Noether "current" density that corresponds to the continuous symmetry of spacetime translations.

Since the metric tensor is the quantity that determines the spacetime geometry, we have:

Conclusion 9.6 *Conceptually, the Einstein equations are perfectly analogous to Gauss's law for the electric field and Ampère's law for the electromagnetic field, and they determine the spacetime geometry, for which the energy–momentum tensor density of the present matter is the "source," i.e., the "driving force."*

That is: by virtue of its presence, matter curves spacetime.

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

Digression 9.4 Relation (9.24) gives a *formal* correspondence between Yang–Mills gauge theories and the general theory of relativity:

$$[\mathbb{A}_\mu]_\alpha{}^\beta \;\longleftrightarrow\; \Gamma^\rho_{\mu\nu}, \qquad \text{and so also} \qquad [\mathbb{F}_{\mu\nu}]_\alpha{}^\beta \;\longleftrightarrow\; R_{\mu\nu\rho}{}^\sigma. \tag{9.47a}$$

This formal correspondence is also qualitatively correct, and foremost in its geometric sense, where the tensors $\mathbb{F}_{\mu\nu}$ and $R_{\mu\nu\rho}{}^\sigma$ represent the curvature of the effective spacetime for the purposes of field propagation and particle motion.

However, in a strictly practical sense – the so-called "engineering" spirit of Section 9.3.4 that also permeates the discussion leading to Conclusion 9.6 – the formal correspondence (9.47a) is not appropriate.[12] The Einstein equations (9.44) identify the differential expression that is of second order in spacetime derivatives of the metric tensor with the energy–momentum tensor density $T_{\mu\nu}$ for that distribution of matter:

$$\Big\{R_{\mu\nu}-\tfrac{1}{2}g_{\mu\nu}\,R = \tfrac{1}{2}g^{\rho\sigma}(\partial_\mu\partial_\rho g_{\nu\sigma}+\partial_\nu\partial_\rho g_{\mu\sigma})+\cdots\Big\} = \frac{8\pi\,G_N}{c^4}T_{\mu\nu}. \tag{9.47b}$$

That system of differential equations is formally analogous to the Gauss–Ampère laws (5.88), expressed in terms of the gauge potential:

$$\Big\{(\Box A^\mu) - \eta^{\mu\nu}(\partial_\nu\partial_\rho A^\rho)\Big\} = \frac{1}{4\pi\epsilon_0}\,\frac{4\pi}{c}\,j_e^\nu. \tag{9.47c}$$

Comparing equations (9.47b) and (9.47c) implies the correspondence

$$A_\mu \;\longleftrightarrow\; g_{\mu\nu}, \qquad F_{\mu\nu} \;\longleftrightarrow\; \Gamma^\rho_{\mu\nu}, \qquad j_e^\mu \;\longleftrightarrow\; T_{\mu\nu}, \tag{9.47d}$$

which better fits this "engineering" sense. The differences between the correspondences (9.47a) and (9.47d) stem from the already mentioned differences, and foremost from the following facts:

1. Both in Yang–Mills gauge theories and in the general theory of relativity, the covariant derivative is defined so that $D_\mu - \partial_\mu \propto \mathbb{A}_\mu$, i.e., $D_\mu - \partial_\mu \propto \mathbb{\Gamma}_\mu$. However, $\mathbb{A}_\mu$ cannot be expressed as the derivative of anything "more fundamental," whereas $\mathbb{\Gamma}_\mu$ can: see equation (9.25).
2. Both in Yang–Mills gauge theories and in general theory of relativity, the curvature is defined as the commutator $[D_\mu, D_\nu]$. However, the Hamilton action for Yang–Mills gauge theory is quadratic in the curvature, while the Einstein–Hilbert action is linear in the (scalar) curvature (9.35).

Finally, the identity

$$R = -g^{\mu\nu}\big(\Gamma^\sigma_{\mu\rho}\Gamma^\rho_{\nu\sigma} - \Gamma^\rho_{\mu\nu}\Gamma^\sigma_{\rho\sigma}\big) + \partial_\mu\mathcal{K}^\mu \tag{9.47e}$$

shows the Einstein–Hilbert Lagrangian to be quadratic in $\mathbb{\Gamma}_\mu$, making it similar – though definitely not identical – to the Yang–Mills type Lagrangians (5.76) and (6.23), in further support of the "engineering" correspondence (9.47d).

[12] This practical sense is regarded "engineering" in the sense that the Gauss–Ampère laws may be used to find the desired electromagnetic field, by constructing the appropriate distribution of charges and currents. Analogously, the Einstein equations (9.44) may be used so that by constructing a particular distribution of matter one produces the desired gravitational field, and so also the spacetime of the desired curvature.

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

9.2.4 Geometry and Newtonian limit

In turn, if we take $\mathscr{L}_M = m\sqrt{g_{\mu\nu}\frac{\partial x^\mu}{\partial t}\frac{\partial x^\nu}{\partial t}}$,[13] which is the Lagrangian density [☞ definition L_0 in Digression 3.7 on p. 93, and defining equation (9.2)] for a particle that moves in spacetime with the metric tensor $g_{\mu\nu}$, then varying the action (9.42) by x^μ yields

$$\frac{\mathrm{d}^2 x^\rho}{\mathrm{d}t^2} + \Gamma^\rho_{\mu\nu}\frac{\mathrm{d}x^\mu}{\mathrm{d}t}\frac{\mathrm{d}x^\nu}{\mathrm{d}t} = 0. \tag{9.48}$$

These are the differential equations that determine the so-called geodesic (extremal) lines. In flat spacetime, $g_{\mu\nu} = -\eta_{\mu\nu}$ and the Christoffel symbol vanishes, so equation (9.48) gives $\ddot{x}^\mu = 0$, i.e., $x^\mu = x^\mu_0 + v^\mu_0 t$ gives straight lines in spacetime. Rearranging the second term we obtain the analogue of Newton's second law:

$$m\,\frac{\mathrm{d}^2 x^\rho}{\mathrm{d}t^2} = F^\rho_{\text{grav}} := -m\,\Gamma^\rho_{\mu\nu}\frac{\mathrm{d}x^\mu}{\mathrm{d}t}\frac{\mathrm{d}x^\nu}{\mathrm{d}t}, \tag{9.49}$$

where the right-hand side provides the gravitational force that curves the trajectory of the particle, the acceleration of which appears on the left-hand side.

> **Conclusion 9.7** *The possibility of reinterpreting essentially geometric information as essentially physical information*
>
> $$\left.\begin{array}{r}\textit{spacetime curvature}\\ \textit{appearing in equation (9.48)}\end{array}\right\} \quad\Leftrightarrow\quad \left\{\begin{array}{l}\textit{definition of the force and}\\ \textit{interaction in equation (9.49)}\end{array}\right. \tag{9.50}$$
>
> *points to the fundamental equivalence of these two ways of thinking and explaining natural phenomena.*

Of course, this is merely one of the simplest examples, but it should be clear that now even in the most general context – including also the Yang–Mills type of gauge interactions[14] [☞ Chapters 5 and 6] – the coupled system of the Einstein equations and the general-relativistically covariant Euler–Lagrange equations of motion may be reinterpreted:

1. either in a purely geometric sense, where objects move along geodesic (extremal) trajectories defined by the (charge/color/isospin-sensing) curvature of spacetime,[14]
2. or in a purely "physicsy" sense, where objects move under the influence of forces with which these objects affect one another.

It behooves us to keep in mind that this latter way of interpreting natural phenomena *implicitly* presupposes the existence of an "empty" spacetime in which these objects move. Therefore, the first, geometric way of interpretation is more economical, and represents the basis of "geometrizing" physics: the notion of force may be replaced by the notion of curvature in the (appropriately generalized) spacetime; see also Comment 3.2.

Starting from (9.48), following Pauli [414], we focus on a spatial component of x, $x^\rho \to x^k$, use that $x^0 = ct$, and assume that $g_{\mu\nu}$ deviates only slightly from its flat-space value, $-\eta_{\mu\nu}$, and obtain

$$\frac{\mathrm{d}^2 x^k}{\mathrm{d}t^2} \approx -c^2\,\Gamma^k_{00}, \tag{9.51}$$

[13] Here, t denotes an arbitrary parameter of the dimension of time, which grows monotonously along the worldline of the given particle.

[14] From this "geometrized" point of view, the various phases that are subject to gauge transformation are to be included in the "total spacetime." Since these phases vary over the usual spacetime, the resulting structure is a called a fiber bundle, where the spacetime-variable phases span the *fibers* over the *base* spacetime. The fiber-wise curvature is measured by the $\mathbb{F}_{\mu\nu}$-type tensors, and is detected only by particles that have the appropriate type of charge: electromagnetic, weak isospin or chromodynamic color.

where terms quadratic in the small deviations $\gamma_{\mu\nu} := (g_{\mu\nu}+\eta_{\mu\nu})$ have been dropped. Assuming furthermore that the components of the metric $g_{\mu\nu}$ are slowly varying in time so that time derivatives may be neglected,

$$\Gamma^k_{00} = \tfrac{1}{2} g^{k\sigma}\big(\partial_0 g_{\sigma 0} + \partial_0 g_{0\sigma} - \partial_\sigma g_{00}\big) \approx -\tfrac{1}{2} g^{k\ell}(\partial_\ell g_{00}). \tag{9.52}$$

In fact, since we must keep $\partial_\ell g_{00} = \partial_\ell(\gamma_{00}-1) = (\partial_\ell \gamma_{00})$, where γ_{00} is the small deviation, dropping terms that are second order in $\gamma_{\mu\nu}$ allows us to drop the (also small) contributions from:

1. off-diagonal terms from the ℓ-summation, and
2. the deviations in g_{kk} from $(-\eta_{kk} = 1)$, whereby $g^{kk} \to 1$.

This produces

$$\frac{\mathrm{d}^2 x^k}{\mathrm{d}t^2} \approx \tfrac{1}{2} c^2 (\partial_k \gamma_{00}), \qquad \text{i.e.,} \qquad \frac{\mathrm{d}^2 \vec{r}}{\mathrm{d}t^2} \approx \tfrac{1}{2} c^2 (\vec{\nabla}\gamma_{00}) \overset{!}{=} -\vec{\nabla}\Phi_{\mathrm{N}}, \tag{9.53}$$

and allows us to identify $-\frac{1}{2}c^2\gamma_{00} := -\frac{1}{2}c^2(g_{00}+1)$ with Newton's gravitational potential, such as $\Phi_{\mathrm{N}} = -G_N \frac{M}{r}$ for a point-like source of gravity of mass M, so the potential energy of the considered particle with mass m at a distance r from the gravitational source is $m\Phi_{\mathrm{N}} = -G_N \frac{mM}{r}$.

Much more detailed derivations of the Newtonian weak-field limit of gravity may be found in the literature; see for example Refs. [96, 95, 271, 58].

9.2.5 Exercises for Section 9.2

✎ **9.2.1** *Prove the relations in Digression 9.1 on p. 325.*

✎ **9.2.2** *Prove the relations in Digression 9.2 on p. 325.*

✎ **9.2.3** *Prove that the Riemann tensor has 20 independent degrees of freedom. (Hint: the rank-4 tensor itself of course has $4^4 = 256$ components. Show that the relations (9.32b) reduce this to 36, the relation (9.32d) further to 21, and relation (9.32e) to 20.)*

✎ **9.2.4** *Prove the relation (9.32f) using the definition (9.30) of $R_{\mu\nu\rho\sigma}$.*

✎ **9.2.5** *Prove that the Ricci tensor is symmetric: $R_{\mu\nu} = R_{\nu\mu}$.*

✎ **9.2.6** *Prove that the equations (9.48) are covariant, i.e., that a coordinate substitution changes these equations only up to a non-vanishing overall multiplicative factor.*

✎ **9.2.7** *Derive the Euler–Lagrange equations of motion for the n-plet of scalar fields $\phi_i(\mathrm{x})$ with the Lagrangian density*

$$\mathscr{L}[\phi_i] = \frac{1}{2} g^{\mu\nu}\delta_{ij}\,(D_\mu\phi^i)(D_\nu\phi^j) - \frac{m^2c^2}{2\hbar^2}\delta_{ij}\phi^i\phi^j. \tag{9.54}$$

(Hint: since ϕ^i are Lorentz-scalars, determine first the action of $D_\mu\phi^i$ from relation (9.17).)

✎ **9.2.8** *From the Lagrangian density (9.54), derive the energy–momentum tensor density, $T_{\mu\nu}$, and the system of Euler–Lagrange equations from the previous exercise coupled with the Einstein equations.*

9.3 Special solutions

Solutions of the Einstein equations (9.44) represent various spacetime geometries – various universes[15] – of which each one may serve as the background/arena in which all "other" physics happens, including the elementary particle physics as analyzed so far. Besides, the Einstein equations – as a system of differential equations for the metric tensor components – are nonlinear, making the existence of a growing class of exact solutions all the more interesting.

9.3.1 The Schwarzschild solution

Only a month after the publication of Einstein's general theory of relativity and gravitation, in 1915, Karl Schwarzschild published the first and best known exact solution to the Einstein equations. Six years later, the mathematician George David Birkhoff proved a theorem[16] whereby any spherically symmetric solution of the Einstein equations without matter (9.39) must be stationary and asymptotically flat, i.e., the geometry of the outer region of spacetime must be described by the Schwarzschild metric tensor (see Refs. [367, 264, 103, 548, 131] and also [128, 587, 127]), given here in spherical coordinates:

$$\textbf{Schwarzschild}\quad \begin{cases} [g_{\mu\nu}] = \operatorname{diag}\big(-f_S(r), \frac{1}{f_S(r)}, r^2, r^2\sin^2(\theta)\big), \\ \mathrm{d}s^2 = -f_S(r)c^2\mathrm{d}t^2 + \frac{1}{f_S(r)}\mathrm{d}r^2 + r^2\big(\mathrm{d}\theta^2 + \sin^2(\theta)\,\mathrm{d}\varphi^2\big), \end{cases} \tag{9.55a}$$

where

$$f_S(r) := \Big(1 - \frac{r_S}{r}\Big), \qquad r_S = \frac{2G_N\,M}{c^2}. \tag{9.55b}$$

As the metric tensor (9.55) satisfies the Einstein equations with $T_{\mu\nu} = 0$, it follows that the Schwarzschild solution describes *empty spacetime*, in the sense that this is a possible geometry of spacetime in the *absence* of any matter. The mass $M := \frac{c^2 r_S}{2G_N}$ that may be ascribed to the point-like object at the origin of the coordinate system then does not represent a particle of matter that is placed there, but is a characteristic of spacetime itself [☞ Digression 9.5 on p. 340], which for observers outside r_S is curved as if there existed an object of mass M.

The meaning of the Schwarzschild radius, r_S, is as follows: The well-known expression for the (first) escape velocity, i.e., the velocity of separation from a planet of mass M at a distance r from the center of the planet is

$$v_1 = \sqrt{\frac{2G_N\,M}{r}}. \tag{9.56}$$

It follows that the separation velocity at the Schwarzschild radius becomes $v_1(r_S) = c$. This literally means that Schwarzschild's solution (9.55) holds for $r \geqslant r_S$. For observers that are outside the Schwarzschild radius, objects that pass through the surface of the sphere of radius r_S can no longer return. This sphere is thus called the "event horizon" and effectively separates the exterior from the interior. As the same conclusion holds also for light, classical physics predicts that the interior of this horizon is completely black for observers in the exterior – whence the popular name "black hole." Formally, the metric tensor (9.55) is applicable also in the interior of the event horizon, but

[15] The distinction between a "spacetime geometry" and a "universe" – as the latter word is used in this chapter – is far from strict: the latter term is used merely to emphasize its *global* meaning. A "universe," after all, has an all-encompassing ring to it and so allows "spacetime geometry" to have either just a local reference, if desired, or a fully global one. In recent times however, the terms "multiverse" and "metaverse" came into vogue, denoting a collection – sometimes infinitely large – of universes [513, 514, 515, 557, for starters]. Especially when these universes within a multiverse are connected, the connotation of globalness of a single universe is restricted in some way or another, at the least. Herein, in turn, a "universe" will be used to denote a closed, isolated and geodesically complete spacetime, unless explicitly stated otherwise.

[16] It was recently discovered that this theorem, many years known under Birkhoff's name, was proven two years earlier (in 1919) by the Norwegian physicist Jørg Tofte Jebsen [297].

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

here the coordinate t becomes space-like and r becomes time-like; the physical meaning of this change remains uncertain, foremost because – at least within classical physics[17] – it is not possible to design an experiment (even a thought-experiment) with which one could compare the evolution of physical phenomena outside the event horizon with those unfolding within the horizon.

Singularities

The functional dependence of the Schwarzschild metric on the radius indicates that there exist two special places within the space with the geometry (9.55):

1. the Schwarzschild radius, where $f_S(r) = 0$, so the metric tensor has a singularity: the coefficient of the $\mathrm{d}t^2$ term vanishes, and the coefficient of the $\mathrm{d}r^2$ term diverges;
2. the coordinate origin, where $f_S(r)$ diverges, so the coefficient of the $\mathrm{d}t^2$ term diverges, and the coefficient of the $\mathrm{d}r^2$ term vanishes.

However, the metric tensor transforms under general coordinate transformations as a rank-2 and type-$(0,2)$ tensor, and it is not clear a priori if these special places are indeed singularities. As the metric tensor is of type $(0,2)$, this transformation has the form [☞ Definition B.2 on p. 511]

$$g_{\mu\nu}(\xi) = \frac{\partial\zeta^\rho}{\partial\xi^\mu}\, g_{\rho\sigma}(\zeta)\, \frac{\partial\zeta^\sigma}{\partial\xi^\sigma} \qquad \Longleftrightarrow \qquad \boldsymbol{g}' = \mathsf{U}^T \boldsymbol{g}\, \mathsf{U} \quad \text{(in matrix form)}, \tag{9.57}$$

which is *not* a similarity transformation. Thus, neither the characteristic polynomial, $\det[\boldsymbol{g} - \lambda\mathbb{1}]$, nor the eigenvalues of the matrix $\boldsymbol{g}$ are invariants. The only invariant that can be constructed from the metric tensor is $\delta_\mu^\rho = g_{\mu\nu} g^{\rho\nu}$, which produces no information about possible singularities.

However, depending on the first and second derivatives of the metric tensor components, the Riemann curvature tensor does contain information about their (non)analyticity, and one only needs to find a way to extract that information in an invariant fashion. The scalar curvature (9.35) is one such invariant. As the Riemann tensor has 20 independent degrees of freedom [☞ Exercise 9.2.3], this leaves precisely 19 independent invariants, but an explicit listing of such invariants remains an open problem☝. Now, there do exist two simple quadratic invariants

$$\|R_{\mu\nu}\|^2 := R_{\mu\nu}\, g^{\mu\rho} g^{\nu\sigma}\, R_{\rho\sigma} \qquad \text{and} \qquad \|R_{\mu\nu\rho}{}^\sigma\|^2 := R_{\mu\nu\rho}{}^\sigma\, g^{\mu\alpha} g^{\nu\beta} g^{\rho\gamma} g_{\sigma\delta}\, R_{\alpha\beta\gamma}{}^\delta, \tag{9.58}$$

of which the second, the so-called Kretschmann invariant for the Schwarzschild metric, equals

$$\|R_{\mu\nu\rho}{}^\sigma\|^2 = \frac{48 G_N{}^2 M^2}{c^4\, r^6}, \tag{9.59}$$

and is indeed divergent at the coordinate origin, $r = 0$. This proves that the coordinate origin is really a singularity of the geometry. The fact that neither the scalar curvature (9.35) nor the quadratic curvature invariants (9.58) diverge on the event horizon does not prove that the location $r = r_S$ is not a singularity. It remains, in principle, to check 17 other independent invariants; the divergence of any one of those invariants on the sphere $r = r_S$ would prove that the event horizon is a singularity. Unfortunately, as no list of 20 independent invariants is known, such a direct verification is not available in practice.[18]

[17] The quantum theory of gravity is not a complete theory, and this analysis is not without debate. However, in the early 1970s, Stephen Hawking was among the first to apply the "semi-classical" analysis and so discover that black holes radiate, emitting the so-called Hawking radiation. The same methods led to the derivation of the Bekenstein–Hawking formula according to which the entropy of a black hole is proportional to the surface area of the event horizon. A recent application of stringy methods and the gravity–gauge duality [☞ p. 443] discovered newer, and not just semi-classical results.

[18] Nor may this suffice even in principle: As discussed in Ref. [264, Section 8.1], because of the non-definiteness of the metric $g_{\mu\nu}$, there could exist singular solutions to the Einstein equations for which all invariant curvature polynomials (constructed from $g_{\mu\nu}$, $g^{\mu\nu}$, $\varepsilon_{\mu\nu\rho\sigma}$ and $R_{\mu\nu\rho}{}^\sigma$) are finite. Also, there do exist special solutions such as the Taub-NUT (Newman, Unti and Tamburino) solution, where the invariant curvature polynomials remain bounded but the spacetime contains incomplete geodesics within a compact neighborhood of the horizon.

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

Fortunately, Georges Lemaître discovered in 1933 that the coordinate substitution (introduced by Arthur Eddington in 1924, without noting the significance)

$$\mathrm{d}\tau := \mathrm{d}t + \sqrt{\frac{r_S}{r}}\,\frac{\mathrm{d}r/c}{(1-\frac{r_S}{r})}, \qquad \mathrm{d}\varrho := \mathrm{d}t + \sqrt{\frac{r}{r_S}}\,\frac{\mathrm{d}r/c}{(1-\frac{r_S}{r})} \tag{9.60a}$$

changes the appearance of the Schwarzschild metric tensor into

$$\mathrm{d}s^2 = -c^2\mathrm{d}\tau^2 + \Big(\frac{2r_S}{3(\varrho - c\tau)}\Big)^{\frac{2}{3}}\mathrm{d}\varrho^2 + r^2\big(\mathrm{d}\theta^2 + \sin^2(\theta)\mathrm{d}\varphi^2\big) \tag{9.60b}$$

and so clearly shows that the sphere $r = r_S$, i.e., $\varrho = \varrho_S := \big(\frac{2}{3}r_S + c\tau\big)$ is free of singularities.

Thus, the event horizon is a completely non-singular location in spacetime and the unlucky observer who drifts through it would notice nothing unusual in his immediate vicinity – except that he would not be able to return outside the event horizon. This phenomenon is often compared with the fact that fish that arrive too close to a waterfall can no longer return upstream.

In turn, the $r = 0$ location is indeed a real singularity [☞ equations (9.58)–(9.59)], and its existence explains the fact that the Schwarzschild solution describes empty space with no matter located within the event horizon, although the coordinate origin may be ascribed the mass $M = \frac{r_S\,c^2}{2G_N}$. More precisely, any Gaussian sphere that fully encloses the event horizon will detect a gravitational field as if within it there existed a mass M. However, such a Gaussian sphere can be shrunk down only as far as the event horizon; beyond that, no information could be extracted from the gravitational field detectors (scales) bedecking the Gaussian sphere. Mathematically, this unusual property stems from the nonlinearity of the Einstein equations and the singularity of the Schwarzschild solution of those equations. Physically, this indicates the self-interaction of the gravitational field – which is conceptually very similar to the self-interaction of non-abelian Yang–Mills gauge fields [☞ so-called "glueballs," discussed on p. 239], and this self-interaction mimics a material particle located at the origin. In fact, the formation of black holes may be described as a phase transition [148, 147] and even have a Landau–Ginzburg effective description [149], much like the Higgs effect [☞ Section 7.1]. However, unlike the fact that black holes have mass, no self-interacting non-abelian Yang–Mills gauge field configuration could exist that would exhibit a non-vanishing charge (color, isospin,. . .) at the origin.

There is, however, another important conceptual difference in describing and modeling Yang–Mills interactions and gravity:

1. The standard models of Yang–Mills interactions [☞ Chapters 5–7] are formulated in flat and infinitely large spacetime, which has the geometry of $\mathbb{R}^{1,3}$, i.e., real 4-dimensional spacetime with the flat metric $g_{\mu\nu} = -\eta_{\mu\nu}$.
2. Models of gravity generally involve a *choice* of a nontrivial metric $g_{\mu\nu} \neq -\eta_{\mu\nu}$, defined on a spacetime that need not at all have the simple structure of $\mathbb{R}^{1,3}$.

When modeling gravity, we *are* free to chose a spacetime where portions – such as singular points – are excised. If all singular points are excised, the remaining spacetime will be singularity-free, but this typically comes at a price: there will exist geodesic paths, solutions to equations (9.48), which tend towards the points that have been excised or are otherwise absent from the given spacetime. It then may or may not be possible to "fill in" (complete) this spacetime in a way that renders all geodesics complete and also (re-)introduces no singular points. Already this observation should make it clear that the (non-)singularity of spacetime is a rather delicate issue that cannot be resolved simply by identifying whether or not all curvature invariants (were one even able to enlist them all) are (non-)singular.

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

In addition, geodesic incompleteness is not the only way of detecting an incompleteness in the spacetime, and it is standard [367, 264] to distinguish at least three a-priori different notions of completeness and incompleteness as its logical negative:

> **Definition 9.2** *A spacetime is* ***geodesically complete*** *if every geodesic path can be extended infinitely within the given spacetime. One may further specify geodesic (in)completeness by restricting to* ***time-like****,* ***null*** *or* ***space-like*** *geodesics.*

This permits the logical possibility that a given spacetime with a given choice of metric is both time-like and null-geodesically complete, but contains incomplete space-like geodesics.

Besides considering geodesic paths as a continuous sequence of points, one may consider any other (discrete) Cauchy sequence of points; this leads to:

> **Definition 9.3** *A spacetime is* ***metrically complete*** *if every Cauchy sequence converges to a point within the given spacetime.*

For a positive-definite metric, it turns out that the geodesic and metric notions of (in)completeness are equivalent [317, 318]. However, the physically interesting case involves the Lorentzian metric of signature $(1,3)$, which is not positive-definite, and where this equivalence does not hold.

There is also another definition, due to C. Ehresmann (1957) and B. G. Schmidt (1971), which generalizes geodesic completeness: One considers all possible smooth (once differentiable) curves in a given spacetime and shows that the length of any such curve is finite in a given parametrization if and only if it is also finite in any other parametrization obtained by parallel transport. Variables parametrizing such curves in a 1–1 fashion are called *(generalized) affine parameters*. Curves with this class of parametrization define a *bundle*, which is then used in the definition [264]:

> **Definition 9.4** *A spacetime is* ***b-complete*** *if every once-differentiable curve of finite length as measured by a generalized affine parameter is within the given spacetime.*

If a finite once-differentiable curve with its end-point(s) contained in the spacetime is a geodesic, this geodesic is complete in the sense of Definition 9.2. If the metric is positive-definite, b-completeness coincides with metric completeness.

The metric is of course not positive-definite in the physically interesting Lorentzian spacetime, in which case it turns out that b-completeness of spacetime implies its geodesic completeness, but the converse is not true [264]. This prompts Hawking and Ellis to *define* a spacetime to be singularity free if it is b-complete, and concede that:

> . . . one might possibly wish to weaken this condition slightly, to say that space-time is singularity-free if it is only *non-spacelike b-complete*, i.e., if there is an end-point for all non-spacelike C^1 [once-differentiable] curves with finite length as measured by a generalized affine parameter.

Needless to say, a detailed analysis of singularities in spacetime geometry and the theory of gravity is much more involved than the purely algebraic considerations around equation (9.59) and certainly beyond our present scope. In addition, the study of gravitation, spacetime geometry, astrophysics and cosmology brings up the questions whether a singularity could dynamically develop within an initially non-singular spacetime, whether an initially singular spacetime could dynamically de-singularize, and how various singularities might interact with each other. The interested Reader is therefore directed to standard references [367, 264, 548, 66, 96], to begin with.

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

9.3.2 Charged and rotating solutions

In 1916–18, Hans Reissner and Gunnar Nordstrøm generalized the Schwarzschild solution to electrically charged black holes:

Reissner–Nordstrøm
$$\begin{cases} [g_{\mu\nu}] = \operatorname{diag}\big(-f_{RN}(r), \frac{1}{f_{RN}(r)}, r^2, r^2\sin^2(\theta)\big), \\ \mathrm{d}s^2 = -f_{RN}(r)\,c^2\mathrm{d}t^2 + \frac{1}{f_{RN}(r)}\mathrm{d}r^2 + r^2\big(\mathrm{d}\theta^2 + \sin^2(\theta)\,\mathrm{d}\varphi^2\big), \end{cases} \tag{9.61a}$$

where

$$f_{RN}(r) := \Big(1 - \frac{r_S}{r} + \frac{r_q^2}{r^2}\Big), \qquad r_q := \sqrt{\frac{q^2\, G_N}{4\pi\epsilon_0\, c^4}}. \tag{9.61b}$$

This solution has a horizon at the location where $g_{rr} \to \infty$, i.e., where $f_{RN}(r) = 0$:

$$r_\pm = \tfrac{1}{2}\Big(r_S \pm \sqrt{r_S^2 - 4r_q^2}\Big). \tag{9.62}$$

For $2r_q < r_S$, the concentric spheres of radii r_+ and r_- are two concentric horizons. When $2r_q = r_S$, the two horizons coincide, and this case is called the *extremal Reissner–Nordstrøm* solution. Using equations (9.55b) and (9.61b), the extremal case is characterized by the relation $q = \sqrt{4\pi\epsilon_0 G_N}\, M$. For two extremal Reissner–Nordstrøm solutions of the same-sign electric charge, the gravitational attraction precisely cancels the electrostatic repulsion and there is effectively no interaction. In the case when $2r_q > r_S$, i.e., when $q > \sqrt{4\pi\epsilon_0 G_N} M$ and the black hole is "overcharged," there are no horizons and the singularity at the coordinate origin would be visible to the observer at any distance.

Comment 9.4 *A singularity that is not enclosed by an event horizon is called "naked." The existence of naked singularities would violate Roger Penrose's cosmic censorship hypothesis (to wit, that every singularity is enclosed within an event horizon and is accessible to no "outside" observer). In accord with this hypothesis, it is* ***believed*** *that the gravitational collapse of matter cannot create naked singularities*☝.

The exact solution for a chargeless, static, spinning black hole was discovered by Roy Kerr only in 1963, and is now most often specified in the coordinates given by Robert H. Boyer and Richard W. Lindquist in 1967:

Kerr
$$\begin{cases} \mathrm{d}s^2 = -\Big(1 - \frac{r_S\, r}{\rho^2}\Big)c^2\mathrm{d}t^2 + \rho^2\Big(\frac{1}{\Delta}\mathrm{d}r^2 + \mathrm{d}\theta^2\Big) \\ \qquad + \Big(r^2 + \ell^2 + \frac{r_S\, r\, \ell^2}{\rho^2}\sin^2(\theta)\Big)\sin^2(\theta)\,\mathrm{d}\varphi^2 - \frac{2r_S\, r\,\ell\,\sin^2(\theta)}{\rho^2}c\,\mathrm{d}t\,\mathrm{d}\varphi, \end{cases} \tag{9.63a}$$

where

$$\ell := \frac{L}{Mc}, \qquad \rho := \sqrt{r^2 + \ell^2\cos^2(\theta)}, \qquad \Delta := r^2 - r_S\, r + \ell^2, \tag{9.63b}$$

and L is the angular momentum. Note that – unlike in the Schwarzschild (9.55) and Reissner–Nordstrøm (9.61) solutions – the Kerr metric tensor is not diagonal: the (ct, r, θ, φ) coordinates are not orthogonal in the Kerr geometry. This solution possesses two event horizons at the location where $g_{rr} \to \infty$, which gives two concentric spheres of radii

$$r_H^\pm = \tfrac{1}{2}\Big(r_S \pm \sqrt{r_S^2 - 4\ell^2}\Big), \tag{9.64}$$

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

of which r_H^+ is clearly *the* relevant event horizon for outside observers. In turn $g_{tt} \to 0$ occurs on the ellipsoids (adopting Visser's nomenclature [540]):

$$\textbf{ergosurface} \qquad r_E^\pm = \tfrac{1}{2}\Big[r_S \pm \sqrt{r_S^2 - 4\ell^2\cos^2(\theta)}\Big]. \tag{9.65}$$

The space between the outer one of these ellipsoids and the outer one of the spherical event horizons is called the *ergoregion*. Objects that enter through the outer ergosurface (9.65) must co-rotate with an angular speed of at least

$$\Omega = -\frac{g_{t\varphi}}{g_{\varphi\varphi}} = \frac{2r_S\, r\, \ell\, c}{\rho^2(r^2+\ell^2) + r_S\, r\, \ell^2 \sin^2(\theta)}, \tag{9.66}$$

even if this implies that they move faster than c, in reference to outside observers. Such superluminal motion, however, does not contradict the theory of relativity, as in a real sense the spacetime itself inside the ergoregion co-rotates akin to a radially accelerating conveyor belt, and objects are – in reference to this co-rotating spacetime – not moving faster than c.

However, since the ergosurface (9.65) is not a "one-way" event horizon, objects can dip into the ergoregion and come back out of it. As the motion during the passage through the ergoregion is faster than a "parallel" motion outside the ergoregion, such an object will draw energy from the spinning black hole. Indeed, consider a conveyer belt that passes through the ergoregion but loops back outside the ergoregion. The co-rotation within the ergoregion will thus drive the conveyor belt outside the ergoregion and so do useful work. This process of drawing energy from a spinning black hole is called the Penrose process, after Roger Penrose, who discovered this possibility. Also, there exist trajectories that pass through the ergoregion, which make it possible to travel backwards in time.

Two years later, in 1965, Ezra Newman found a generalization of the Kerr metric tensor, for an electrically charged spinning black hole:

$$\textbf{Kerr–Newman} \qquad \begin{cases} \mathrm{d}s^2 = -\dfrac{\Delta}{\rho^2}\Big(c\,\mathrm{d}t - \ell\sin^2(\theta)\,\mathrm{d}\varphi\Big)^2 + \rho^2\Big(\dfrac{1}{\Delta}\mathrm{d}r^2 + \mathrm{d}\theta^2\Big) \\ \qquad\qquad + \dfrac{\sin^2(\theta)}{\rho^2}\Big((r^2+\ell^2)\mathrm{d}\varphi - \ell c\,\mathrm{d}t\Big)^2, \end{cases} \tag{9.67a}$$

where

$$\ell := \frac{L}{Mc}, \quad \rho := \sqrt{r^2 + \ell^2\cos^2(\theta)}, \quad \Delta := r^2 - r_S\, r + \ell^2 + r_q^2, \quad r_q := \sqrt{\frac{q^2\, G_N}{4\pi\epsilon_0\, c^4}}, \tag{9.67b}$$

and L, M, and q are the angular momentum, the mass and the electric charge of the black hole. Just as the Kerr metric tensor (9.63), the Kerr–Newman metric tensor (9.67) is also not diagonal, and the (ct, r, θ, φ) coordinates are not orthogonal in the Kerr–Newman geometry.

Furthermore, direct computation proves that the location $\rho = 0$ is a true coordinate singularity for both the Kerr geometry (9.63) and the Kerr–Newmann solution (9.67), since the Kretschmann curvature invariant $\|R_{\mu\nu\rho}{}^{\sigma}\|^2$ defined in equations (9.58) diverges there. Given that the location $r = 0$ within the standard interpretation of the coordinates (r, θ, φ) is a single point, the result

$$\rho = 0 \quad \Leftrightarrow \quad r = 0, \text{ and } \big(\theta = \tfrac{\pi}{2} \text{ if } \ell \neq 0\big), \tag{9.68}$$

may appear puzzling, in that the coordinate location "$r = 0$ and $(\theta \neq \frac{\pi}{2}$ if $\ell \neq 0)$" is singular in neither the Kerr geometry nor the Kerr–Newmann geometry. This indicates that the coordinate locations "within the coordinate origin,"

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

$$O_* := \{r=0,\ \theta=\tfrac{\pi}{2},\ \varphi\in[0,2\pi]\} \quad\text{and}\quad O_\circ := \{r=0,\ \theta\neq\tfrac{\pi}{2},\ \varphi\in[0,2\pi]\}, \tag{9.69}$$

must be distinguished. This makes it obvious that the coordinates (r,θ,φ) must not be interpreted literally as the standard spherical coordinates for the Kerr and the Kerr–Newmann geometries, (9.63) and (9.67), respectively; R. Wald provides the standard argument for O_* to be interpreted as a ring-shaped singularity in these geometries [548, pp. 314–315]; see also [540]. Consequently, the whole coordinate region

$$O := \{r=0,\ \theta\in[0,\pi],\ \varphi\in[0,2\pi],\ \ \varphi\simeq\varphi+2\pi\} \tag{9.70}$$

must be regarded as a *null* 2-sphere standing in the place of the standard coordinate origin, and the singularity of the Kerr and the Kerr–Newmann geometry is then located on the equator of this null 2-sphere. This recalls the process of "blowing up a singularity," where the null 2-sphere is the "exceptional divisor" [279, for starters].

Not even a decade later, in 1972–3, Akira Tomimatsu and Humitaka Sato discovered a class of exact solutions [523, 524, 270] [☞ also [200] for a recent review and applications] that generalize the Kerr solution (with polar coordinates $\rho := \sqrt{x^2+y^2}$ and φ):

$$\textbf{Kerr–Tomimatsu–Sato}\qquad \mathrm{d}s^2 = -Fc^2\big[\mathrm{d}t-\omega\,\mathrm{d}\varphi\big]^2 + F^{-1}\big[E\,(\mathrm{d}\rho^2+\mathrm{d}z^2)+\rho^2\mathrm{d}\varphi^2\big], \tag{9.71a}$$

where the functions E, F and G are most easily expressed in terms of prolonged spheroidal coordinates (ξ,η,φ):

$$x = \rho_0\sqrt{(\xi^2-1)(1-\eta^2)}\cos\varphi,\quad y = \rho_0\sqrt{(\xi^2-1)(1-\eta^2)}\sin\varphi,\quad z = \rho_0\,\xi\eta, \tag{9.71b}$$

so $\rho = \rho_0\sqrt{(\xi^2-1)(1-\eta^2)}$:

$$E(\xi,\eta) := \frac{A(\xi,\eta)}{p^{2\delta}(\xi^2-\eta^2)^{\delta^2}},\quad F(\xi,\eta) := \frac{A(\xi,\eta)}{B(\xi,\eta)},\quad G(\xi,\eta) := \frac{2L/mc}{A(\xi,\eta)}(1-\eta^2)C(\xi,\eta), \tag{9.71c}$$

where $A(\xi,\eta)$, $B(\xi,\eta)$ and $C(\xi,\eta)$ are polynomials of degree $2\delta^2$, $2\delta^2$ and $(2\delta^2-1)$, respectively, and where the constants ρ_0 and p are algebraic functions of the mass m, angular momentum L, the integral parameter δ and the natural constants [523, 524]

$$\rho_0 := \frac{G_N}{c^2}\frac{p}{\delta}m \qquad\text{and}\qquad p = \sqrt{1-\frac{c^2}{G_N{}^2}\frac{L^2}{m^4}}. \tag{9.72}$$

The Tomimatsu–Sato solutions depend on the parameter δ, so that $\delta = 1$ gives the Kerr solution, but for $\delta \neq 1$ the Tomimatsu–Sato solutions contain naked singularities.

— ❦ —

It is important to understand that the very nontrivial solutions (9.55), (9.61), (9.63), (9.67) and (9.71) are but a few special – and physically very interesting – representatives of a general class of solutions of the Einstein equations without matter. In other words, solutions to the Einstein equations (9.39) include very nontrivial geometries that even contain locations (in the presented case, the so-called black holes) that have the appearance of a particle: they have a mass, and may have electric charge and intrinsic angular momentum.

Digression 9.5 It is reasonable then to inquire whether, e.g., the electron could be simply a charged black hole. However, with the mass and the charge of the electron, one easily obtains

$$r_S(e^-) = 1.353\times10^{-57}\,\text{m} \ll \ell_P \qquad \text{and} \qquad r_q(e^-) = 9.152\times10^{-37}\,\text{m} < \ell_P. \tag{9.73a}$$

Since $r_S(e^-) < r_q(e^-)$, this black hole has no event horizon, and represents a naked singularity. However, as both characteristic radii are smaller than the Planck length, Conclusion 1.5 on p. 30 indicates that this model is unverifiable. That is, because of Conclusion 1.5, it simply is not possible to determine any concretely verifiable consequence of representing the electron by a charged miniature (classical!) black hole.

Strictly speaking, the complete theory of quantum gravity does not exist,[19] so that only estimates exist that indicate that – contrary to the name – *quantum* black holes radiate. This radiation is named after Stephen Hawking, who in 1974 explained the quantum process that enables this radiation, and without violating the "one-way" nature of the event horizon. These estimates indicate that black holes lose mass via the Hawking radiation, and so have an "evaporation time" [403]:

$$t_{\text{evap.}} \approx 5{,}120\pi\frac{G_N{}^2}{\hbar c^4}M^3 \approx 8.407\times10^{-17}\big(M/\text{kg}\big)^3\,\text{s}. \tag{9.73b}$$

For charged leptons and quarks, electric charge conservation would have to obstruct their evaporation, but for neutrinos with a mass $m_\nu < 2\,\text{eV} \sim 3\times10^{-36}\,\text{kg}$ the evaporation time is of the order $< 4\times10^{-127}\,\text{s}$, which is some 83 orders of magnitude shorter than the Planck time. Conservation of angular momentum ($\frac{1}{2}\hbar$) of all fundamental fermions would also have to obstruct their evaporation – including neutrinos – when represented by a miniature black hole: Indeed, the Hawking radiation may consist only of particles that are lighter than the black hole that is evaporating by means of this radiation; only photons are lighter than neutrinos, but photons have integral spin.

In principle, therefore, miniature black hole models for quarks and leptons would have to be stable, but such models would seem to be essentially unverifiable owing to the result (9.73a); see however also Refs. [420, 421, 336, 17, 464, 57, 78, 79]. In particular, it has been known since 1968 [98] that a Kerr–Newmann black hole has no electric dipole moment, but does have a magnetic dipole moment with a gyromagnetic ratio equal to 2, just like the Dirac electron without the field theory $O(\alpha)$ corrections [☞ Digression 4.1 on p. 132].

Finally, the general solutions to the Einstein equations without matter (9.39), including the Schwarzschild, the Reissner–Nordstrøm, the Kerr and the Kerr–Newman geometry, define a class of macroscopic geometries of various possible vacua; i.e., empty spacetimes. In such models, the central objects such as black holes are not to be treated as matter, but as a geometric property (defect) of spacetime itself. Even qualitatively, this recalls the "topological" solutions discussed in Section 6.3.1, including also the Dirac magnetic monopole from Section 5.2.3, other similar solutions [☞ Conclusion 6.7 on p. 248], and the "glueball" solutions in non-abelian Yang–Mills gauge theories, discussed on p. 239.

[19] String theory is known to be a quantum theory that contains gravity; the technical development of this theory suffices to confirm these estimates but not yet to compute any corrections.

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

9.3.3 Other interesting solutions

This section will explore some known solutions to the Einstein equations. As in the previous section, the solutions are specified by providing the line element $\mathrm{d}s^2 = g_{\mu\nu}\mathrm{d}x^\mu \mathrm{d}x^\nu$. This determines the "background" spacetime geometry [☞ Conclusion 6.8 on p. 248] in which one may analyze the motion of particles, the presence of which one supposes is a small perturbation to the energy–momentum tensor density and so also the Einstein equations, so that the spacetime geometry is not significantly changed. Such solutions are often called "universes" or "worlds," understanding that this is an extremely simplified picture where this "world" consists only of the background spacetime geometry, the matter/energy required to stabilize this geometry, and the test particles the effect of which upon the geometry may be neglected. The interested Reader is directed to the catalogues [497, 372] for starters.

Standard geometries in cosmology

The definition of the geometry that is most often used in cosmology and is understood to be the standard was provided by Alexander Friedman, Georges Henri Joseph Édouard Lemaître, Howard Percy Robertson and Arthur Geoffrey Walker, and this we will refer to as the FLRW geometry. (Depending on the historical precision and socio-political accent, Authors in the research literature not infrequently omit one or more of these names and initials.) The metric tensor of the FLRW geometry is given in terms of the "reduced-circumference polar coordinates":

$$\textbf{FLRW}\quad \mathrm{d}s^2 = -c^2\mathrm{d}t^2 + a^2(t)\mathrm{d}\Sigma^2, \qquad \begin{cases} \mathrm{d}\Sigma^2 := \Big[\dfrac{\mathrm{d}r^2}{1-K\,r^2} + r^2\mathrm{d}\Omega^2\Big], \\ \mathrm{d}\Omega^2 := \mathrm{d}\theta^2 + \sin^2(\theta)\mathrm{d}\varphi^2, \end{cases} \tag{9.74}$$

where $a(t)$ is a dimensionless "scale function" of time, and K is the Gauss curvature of the space at the time when $a(t) = 1$. Alternatively, one writes $k := \frac{K}{|K|} = \pm 1$ a $k = 0$ when $K = 0$, whereby r is a dimensionless variable in the direction of the distance from the coordinate origin and $a(t)$ has the physical dimensions of length. In the case of positive curvature, space is a 3-sphere and the coordinates (r, θ, φ) cover only half of this space in a single-valued fashion, whereupon they are called the "reduced-circumference polar coordinates": in analogy with the cylindrical distance from the z-axis on the surface of a 2-sphere, the radial variable r grows from the north pole up to the equator but then decreases towards the south pole and this results in the two-valuedness of the coordinate system. Instead of (r, θ, φ), we may use the hyper-spherical coordinates:

$$\mathrm{d}\Sigma^2 = \mathrm{d}r^2 + S_K^2(r)\mathrm{d}\Omega^2, \qquad S_K(r) := \begin{cases} \frac{1}{\sqrt{K}}\sin(r\sqrt{K}) & K > 0, \\ r & K = 0, \\ \frac{1}{\sqrt{|K|}}\sinh(r\sqrt{|K|}) & K < 0, \end{cases} \tag{9.75}$$

which do not have this drawback.

This metric tensor solves the Einstein equations in the case when the matter has a homogeneous and isotropic energy–momentum tensor density, so that the Einstein equations reduce to the pair:

$$\Big(\frac{\dot a}{a}\Big)^2 + \frac{Kc^2}{a^2} - \frac{\Lambda c^2}{3} = \frac{8\pi G_N}{3}\,\varrho, \tag{9.76a}$$

$$2\frac{\ddot a}{a} + \Big(\frac{\dot a}{a}\Big)^2 + \frac{Kc^2}{a^2} - \Lambda c^2 = -\frac{8\pi G_N}{c^2}\,p, \tag{9.76b}$$

where ϱ and p denote the density and pressure of matter, and Λ is the cosmological constant. Since the redefinitions

$$\varrho \to \varrho - \frac{\Lambda c^2}{8\pi G_N} \quad \text{and} \quad p \to p + \frac{\Lambda c^4}{8\pi G_N} \tag{9.77}$$

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

effectively eliminate the cosmological constant, it follows that the presence of the cosmological constant may be simulated by *something* that (1) permeates the universe, (2) is homogeneous and isotropic, and (3) the pressure and the density of which satisfy the relation

$$p = -\varrho c^2. \tag{9.78}$$

Generally, *anything* that has a negative pressure ($p/\varrho < 0$) is called *dark energy* and its presence in the FLRW cosmology induces the universe to expand. For an accelerated expansion of the universe it would suffice were the dark energy to satisfy the relation

$$p < -\tfrac{1}{3}\varrho c^2. \tag{9.79}$$

A scalar field with this property is dubbed *quintessence*, and the ratio p/ϱ is then not necessarily a constant. Finally, one obtains an extremely accelerated expansion of the universe if

$$p < -\varrho c^2, \tag{9.80}$$

which is then referred to as phantom energy. Note that these are phenomenological definitions:

> ***Definition 9.5*** *Anything homogeneous and isotropic throughout the whole spacetime is called:*
>
> **dark energy** *if the pressure is negative:* $p/\varrho < 0$;
> **quintessence** *if the density and the pressure satisfy (9.79):* $p/\varrho < -c^2/3$;
> **cosmological constant** *if the density and the pressure satisfy (9.78):* $p/\varrho = -c^2$;
> **phantom energy** *if the density and the pressure satisfy (9.80):* $p/\varrho < -c^2$.

Dark energy is thus an umbrella term including its three more specific types. The demarcations are determined by the qualitative differences in the induced evolution of the universe: The cosmological constant causes the spacetime geometry to accelerate its expansion, while phantom energy causes this expansion to diverge in finite time. In turn, models of quintessence typically involve at least one dynamical field, which then varies over spacetime; moduli fields in superstring theory are natural and oft-tried candidates [☞ Footnote *34* on p. 443].

Of particular interest are the special cases of the FLRW geometry [367]:

$$ds^2 = \begin{cases} -c^2dt^2 + a_0^2\, e^{+2c\sqrt{\Lambda/3}\,t}\, d\vec{r}^{\,2}, & \textbf{de Sitter}, \\ -c^2dt^2 + d\vec{r}^{\,2}, & \textbf{Minkowski}, \\ -c^2dt^2 + a_0^2\, e^{-2c\sqrt{\Lambda/3}\,t}\, d\vec{r}^{\,2}, & \textbf{anti de Sitter}, \end{cases} \tag{9.81}$$

where $H := 2\sqrt{\Lambda/3} > 0$ is the so-called Hubble constant,[20] and $d\vec{r}^{\,2} = d\vec{r}\cdot d\vec{r}$ is the familiar Euclidean norm of the spatial differential $d\vec{r}$. Because of using the familiar (flat) Euclidean norm for the spatial part of the differential, the coordinates in equation (9.81) are also called the "flat coordinates." There also exists the "static" parametrization

$$ds^2 = -c^2\big(1 \mp \tfrac{1}{3}\Lambda\rho^2\big)d\tau^2 + \big(1 \mp \tfrac{1}{3}\Lambda\rho^2\big)^{-1}d\rho^2 + \rho^2\big(d\theta^2 + \sin^2(\theta)d\phi^2\big), \tag{9.82}$$

where ρ is a suitable "radial" coordinate; for a precise relation between equations (9.81) and (9.82), see Ref. [367]; the upper (negative) sign produces the metric for the de Sitter spacetime, and the lower (positive) sign for the anti de Sitter spacetime. Finally, there also exists the quotient parametrization

$$ds^2_{\text{AdS}} = \tfrac{L^2}{z^2}\big(-c^2dt^2 + dx^2 + dy^2 + dz^2\big). \tag{9.83}$$

[20] The proposal that the universe is expanding and with a rate now called the Hubble constant was made by Georges Lemaître in 1927, two years before Edwin Hubble confirmed and more precisely determined the expansion rate; see Refs. [550, 532, 68] and the references therein.

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

Expression (9.82) should make it clear that the de Sitter spacetime has a spherical horizon with the radius $\rho_H = \sqrt{3/\Lambda}$. In turn, the $z^2 \to 0$ limiting case of the expression (9.83) defines the flat metric $-\mathrm{d}z^2 = -c^2\mathrm{d}t^2 + \mathrm{d}x^2 + \mathrm{d}y^2$ on a (2+1)-dimensional space with the Minkowski metric, and that forms the "conformal limit" of the anti de Sitter spacetime.

Finally, the $(n{+}1)$-dimensional de Sitter spacetime may be defined also as the orthogonal group coset $O(1, n{+}1)/O(1, n)$, and the anti de Sitter spacetime equals $O(2, n)/O(1, n)$.

— ❦ —

Note that the $g^{\mu\nu}$-trace of the Einstein equations (9.44) produces $R = -\frac{8\pi G_N}{c^4} T$. Substituting this into equation (9.44) yields

$$R_{\mu\nu} = \frac{8\pi G_N}{c^4} \Big[T_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} T \Big]. \tag{9.84}$$

This makes it clear that every solution where the energy–momentum tensor density of matter vanishes, $T_{\mu\nu} = 0$, the Ricci tensor also must vanish. And the other way around, the vanishing of the Ricci tensor, via the Einstein equations (9.44), implies that $T_{\mu\nu} = 0$ also. The geometries (choices of the metric tensor) for which the Ricci tensor vanishes (and so $T_{\mu\nu} = 0$) are called Ricci-flat geometries. This of course includes the *flat* geometry, where the metric tensor $g_{\mu\nu} = -\eta_{\mu\nu}$ is a constant, and all components of both the Christoffel symbol and the Riemann tensor vanish.

In turn, neither the vanishing of the Ricci tensor – nor even of the entire Riemann tensor – implies that the metric is flat. For example, the Kasner geometry has the metric tensor defined as [367, generalized]

$$\textbf{Kasner} \quad \mathrm{d}s^2 = -c^2\mathrm{d}t^2 + \sum_{i=1}^{3} \big(\tfrac{t}{T_i}\big)^{2p_i} (\mathrm{d}x^i)^2, \tag{9.85}$$

where T_i are arbitrary constants with units of time. If the parameters p_i are chosen to satisfy the Kasner conditions

$$\sum_{i=1}^{3} p_i \;=\; 1 \;=\; \sum_{i=1}^{3} (p_i)^2, \tag{9.86}$$

the Einstein tensor ($G_{\mu\nu} := R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R$) and even the Ricci tensor vanish. If we further set any two of three parameters p_i to be zero and the third to be 1, then the entire Riemann tensor vanishes, although the metric tensor is not equal to $-\eta_{\mu\nu}$, the constant metric tensor of flat spacetime.

One of the unusual properties of the Kasner geometry inexorably follows from Kasner's conditions (9.86) themselves: one of the three parameters must be non-positive. That is, we have

$$\text{equation (9.86)} \quad \Rightarrow \quad \begin{cases} p_2^{\pm} &= \frac{1}{2}\Big(1 - p_1 \pm \sqrt{1 + 2p_1 - 3{p_1}^2}\Big), \\ p_3^{\pm} &= 1 - p_1 - \frac{1}{2}\Big(1 - p_1 \pm \sqrt{1 + 2p_1 - 3{p_1}^2}\Big), \end{cases} \tag{9.87}$$

where $-\frac{1}{3} \leqslant p_i \leqslant 1$ for $i = 1, 2, 3$. It is easy to verify that the only non-negative solutions are the permutations of the triple $\vec{p} = (0, 0, 1)$. In turn, if one of the parameters is maximally negative, we have permutations of the triple $\vec{p} = (-\frac{1}{3}, \frac{2}{3}, \frac{2}{3})$. A few examples with rational values are the permutations of $\vec{p} = (-\frac{2}{7}, \frac{3}{7}, \frac{6}{7})$, $(-\frac{3}{13}, \frac{4}{13}, \frac{12}{13})$, $(-\frac{6}{19}, \frac{10}{19}, \frac{15}{19})$, $(-\frac{4}{21}, \frac{5}{21}, \frac{20}{21})$, $(-\frac{5}{31}, \frac{6}{31}, \frac{30}{31})$, etc.

Since $\sqrt{-g} = ct/({T_1}^{p_1}{T_2}^{p_2}{T_3}^{p_3})$, the volume of Kasner geometry expands linearly in time. However, except for the class where the values (p_1, p_2, p_3) are permutations of the triple $(0, 0, 1)$ and where the Kasner geometry stagnates in two directions and expands in the third, the Kasner geometry expands in two spatial directions but shrinks in the third in all other cases.

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

Gödel's universe

One of the most unusual solutions to the Einstein equations was discovered in 1949 by Kurt Gödel; the metric tensor for the geometry of the so-called Gödel universe is specified as [372, 219]

$$\textbf{Gödel}\quad \mathrm{d}s^2 = -c^2\mathrm{d}t^2 + \frac{\mathrm{d}r^2}{1+\left(\frac{r}{r_g}\right)^2} + r^2\Big[1-\Big(\frac{r}{r_g}\Big)^2\Big]\mathrm{d}\phi^2 + \mathrm{d}z^2 - c\,\frac{2\sqrt{2}\,r^2}{r_g}\mathrm{d}t\,\mathrm{d}\phi, \tag{9.88}$$

where r_g is the Gödel radius. These cylindrical coordinates (t, r, ϕ, z) co-rotate with the entire universe, which results in the additional non-diagonal $\mathrm{d}t\,\mathrm{d}\phi$-term.

In this universe and with reference to the coordinate system (t, r, ϕ, z), a light ray that starts from the coordinate origin in the horizontal (r, ϕ)-plane follows an elliptical path that bends in the counter-clockwise direction. At the point where it reaches the distance r_g from the coordinate origin, the light ray is moving in the $+\hat{\mathrm{e}}_\phi$ direction and begins to return to the coordinate origin, where it closes the elliptic path. Thus, observers that are at rest in the coordinate origin cannot see outside the cylinder of the horizontal radius of r_g, which then defines an optical horizon for these observers. This curious property is a consequence of the fact that the light cones (generated by light-like vectors) at every point of the (x, y)-plane tilt in the $+\hat{\mathrm{e}}_\phi$ direction at an angle (away from the coordinate t-axis) that grows with the distance from the origin. At the distance $r_g\ \ln(1+\sqrt{2}) \approx 0.88\,r_g$, the light cone tilts over so much that one of the light-like vectors becomes parallel with $+\hat{\mathrm{e}}_\phi$ and generates a circular light-like path in the (x, y)-plane: a beam of light can be emitted so as to travel on a closed circle of radius $\approx 0.88\,r_g$ – without advancing in coordinate time, t [264].

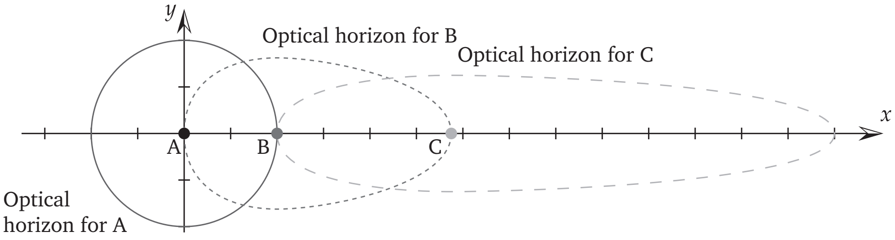

Figure 9.3 Optical horizons for observers A, B and C in the Gödel universe.

For an observer located outside the coordinate origin there exists a similar optical horizon, of an ovoid shape where the ovoid is narrower and longer in the region further away from the coordinate origin, as shown in Figure 9.3. Note that, by the definition of the optical horizon in the Gödel universe, light returns owing to the Doppler effect and the co-rotation of the entire universe and not owing to gravity. This optical horizon is thus of an essentially different nature from the event horizon in the Schwarzschild geometry.

In turn, a particle with a non-vanishing mass that is at some point at rest with respect to this coordinate system remains in that resting state, i.e., moves only in time. Thus, the coordinates (t, r, ϕ, z) are called co-rotating, and the radius r_g presents the effective optical horizon. For the Gödel universe it is convenient to define the angular speed

$$\Omega_g := \frac{\sqrt{2}\,c}{r_g} \tag{9.89}$$

with which the matter "at rest" and the entire Gödel universe rotate.

In spite of this rotation, the geometry (9.88) is homogeneous. From the form of the metric tensor, it should be clear that translations in the $\hat{\mathrm{e}}_t$ and $\hat{\mathrm{e}}_z$ directions as well as rotations in the $\hat{\mathrm{e}}_\phi$

directions are isometries (symmetries of the metric tensor) and that they are respectively generated by the differential operators

$$X_0 := \tfrac{1}{\Omega_g}\,\partial_t, \qquad X_3 := r_g\,\partial_z \qquad \text{and} \qquad X_2 := \partial_\phi. \tag{9.90}$$

As for the radial direction, ∂_r is clearly not a symmetry as this would translate $r \to r + r_0$, leaving a cylindrical "hole" of radius r_0, whereas a $r \to r - r_0$ translation would map points near the z-axis into a nonexistent domain with the absurd value $r < 0$. However, it turns out that there do exist *two* differential operators,

$$X_1 := \frac{1}{\sqrt{1+\big(\frac{r}{r_g}\big)^2}}\Big[\tfrac{r}{\sqrt{2}c}\cos\phi\,\partial_t + \tfrac{r_g}{2}\big[1+\big(\tfrac{r}{r_g}\big)^2\big]\sin\phi\,\partial_r + \tfrac{r_g}{2r}\big[1+2\big(\tfrac{r}{r_g}\big)^2\big]\cos\phi\,\partial_z\Big], \tag{9.91}$$

$$X_4 := \frac{1}{\sqrt{1+\big(\frac{r}{r_g}\big)^2}}\Big[\tfrac{r}{\sqrt{2}c}\cos\phi\,\partial_t - \tfrac{r_g}{2}\big[1+\big(\tfrac{r}{r_g}\big)^2\big]\cos\phi\,\partial_r + \tfrac{r_g}{2r}\big[1+2\big(\tfrac{r}{r_g}\big)^2\big]\sin\phi\,\partial_z\Big], \tag{9.92}$$

that do generate isometries. Gödel, in his original work in 1949, already used four of these five isometries to show that this geometry is homogeneous, and it was shown in 1992 [167] that the complete set of *five* isometries closes the $\mathfrak{so}(3) \oplus \mathfrak{tr}(\mathbb{R}^{1,1})$ algebra:

$$L_1 := X_4, \quad L_2 := X_1, \quad L_3 := -i(X_0{+}X_2), \qquad \begin{cases} [\,L_j\,,\,L_k\,] = i\varepsilon_{jk}{}^{\ell}L_\ell, \\ [\,L_j\,,\,X_0\,] = 0 = [\,L_j\,,\,X_3\,], \end{cases} \tag{9.93}$$

where $\mathfrak{tr}(\mathbb{R}^{1,1})$ is the abelian algebra of translations in the (t,z)-plane. These symmetries can then be used to map points, paths, vectors and other tensors from one point of the Gödel universe into another, so that it suffices to work out the geometric properties with reference to the given coordinate system (9.88) and with the origin of the spatial coordinates as the reference point.

The coordinate time t and the proper time τ are identical for the observer "at rest" at the coordinate origin. Very near the z-axis, so for $r \sim 0$, the Gödel geometry is approximately flat (in cylindrical coordinates). Using the homogeneity and the action of the algebra (9.93), this then holds locally for any observer.

In the co-rotating basis the Einstein tensor ($G_{\mu\nu} := R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R$) is given as

$$[G_{\mu\nu}] = \Omega_g^2\,\mathrm{diag}(-1,1,1,1) + 2\Omega_g^2\,\mathrm{diag}(1,0,0,0). \tag{9.94}$$

The Einstein equations then dictate that this geometry is maintained by a type of matter for which the energy–momentum tensor density has the same value. The first contribution describes the so-called lambda-vacuum, i.e., the solution with the cosmological constant [☞ relations (9.77)–(9.79)]. The second contribution describes a co-rotating perfect (and all-permeating) fluid, i.e., a co-rotating dust.[21] Note that the coefficients of the two contributions must be in the precise proportion as given in equation (9.94).

> ***Conclusion 9.8*** *The Gödel geometry of spacetime may be understood as the result of an even permeation of the whole spacetime with dark energy (cosmological constant) and a perfect fluid, and in the precise proportion provided in the expression (9.94).*

The Gödel geometry is a relatively rare example of a geodesically complete and non-singular geometry [☞ the lexicon entry, in Appendix B.1]: The coordinate system (9.88) covers the entire Gödel spacetime, and contains no singularity; it also has an unusually symmetric structure (9.93).

[21] The cylindrical solution with co-rotating dust was discovered in 1924 by Cornelius Lanczos [325], but the solutions is better known after Willem Jacob van Stockum, who analyzed it in 1937 [534].

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

Traveling through time

Of course, every particle incessantly travels – through time, in the direction of time flow. However, the Lanczos–Stockum solution of co-rotating dust and Gödel's co-rotating universe were amongst the first solutions to contain so-called closed time-like curves; the Kerr solution (9.63) also has such curves. Those are curves for which the tangent vector is always time-like [☞ Definitions 3.3 on p. 90], but which are closed in spacetime. An ordinary particle with nonzero mass may travel along such a curve, and so can return into its own past!

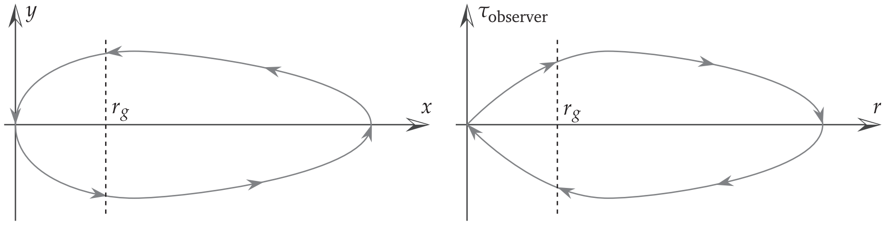

Figure 9.4 A time-like closed path in the Gödel universe.

The simplest such closed time-like curve in the Gödel universe is an ovoid path in the (x,y)-plane, e.g., with the x-axis as its symmetry axis, as shown (following the analysis of Ref. [219]) in Figure 9.4. Similar, but much more complicated closed time-like curves may then be found both in the Gödel universe, and in the Lanczos–Stockum solution with co-rotating dust, and also in many other exact solutions [☞ catalogues [497, 372] as well as the texts [367, 548]]. Note that, following the path in Figure 9.4, the particle moves backwards in time only outside the optical horizon of the observer at the coordinate origin. Also, Ref. [219] gives the necessary conditions: For a particle to move along such a closed time-like curve, it must be launched with the initial speed $v \gtrsim 0.98c$ (measured in the co-rotating coordinate system) and from a location $r \lesssim 1.7r_g$, as well as any other initial conditions that are obtained from these by isometry algebra transformations (9.93).

These concrete, exactly solved examples are particularly important to indicate the fact that many intuitively clear and acceptable characteristics of flat spacetime – including also the perhaps beguiling but precisely resolved situations in the special theory of relativity[22] – simply need not hold in the general theory of relativity. For details about closed time-like curves, the ambitious Reader should consult the books [519, 265].

Digression 9.6 Most typical scenarios of reversing the direction of traveling in time contradict energy conservation: Suppose an object X were to travel forward in time from $t < t_0$ to $t_1 > t_0$, then "turn around" to travel in time from t_1 back to $t_0 < t_1$, and then continue traveling in time forward as usual, through t_1 and beyond. Figure 9.5 depicts this process in two versions, to the left where the object X travels continuously backward in time, and to the right where it "jumps." So, in version (a), the

[22] The so-called paradoxes most often mentioned are the twin-paradox, and those of the ladder and the barn, the ruler and hole in the table, but there exist many others [512]. Not one of these puzzling situations is a real paradox and merely indicates that many of our notions acquired in everyday life are approximations that are really fit only for flat spacetime or at most locally, and so must be reconsidered and adapted beyond such local applications. For example, simultaneity becomes a relative notion, and the rigid body makes sense only at non-relativistically small speeds, since the action of the force cannot propagate through the body faster than the speed of light in vacuum, so that each body bends under the influence of non-simultaneous forces.

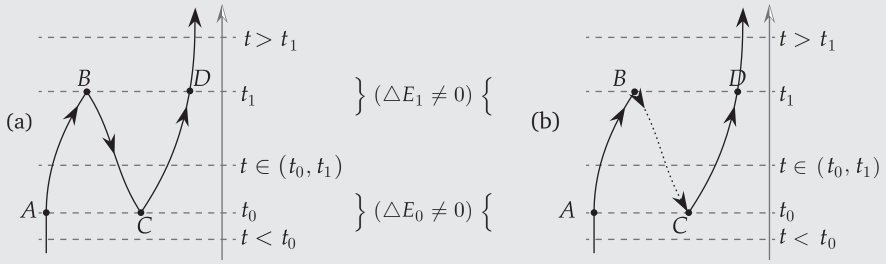

Figure 9.5 Two typical scenarios of time-travel: (a) with continuous backward travel, and (b) with "instantaneous" backward travel. Energy is measured by adding up all contributions at spacetime points simultaneous to a given observer and connected by the dashed lines. In both scenarios, energy fails to be conserved.

change $\triangle E_0 = (m_X + m_{\overline{X}})c^2$ occurs as time passes from before t_0 to after t_0, and then $\triangle E_1 = -(m_X + m_{\overline{X}})c^2$ as time passes through t_1. In version (b), $\triangle E_{0,1} = \pm m_X c^2$ at these two points in time. In diagrams with elementary particles similar to (a), another kind of particle is emitted from the point B and absorbed at the point C to balance 4-momentum conservation.

However, in the general, nontrivial geometries (and topologies) necessary to describe gravity in all generality, energy and 3-momentum are not globally well defined. These quantities are spatial 3-dimensional integrals of the $T_{\mu\nu}$ components of the energy–momentum density tensor, where the domain of integration is a 3-dimensional space-like hypersurface of simultaneous points in the 4-dimensional spacetime, as chosen by a specific class of observers. Most admissible 4-dimensional spacetime geometries admit a wide variety of such 3-dimensional space-like hypersurfaces, over which the required integrals produce widely differing results; the analysis is improved by restricting to coordinate systems satisfying the de Donder gauge condition, $\partial_\mu(\sqrt{-g}g^{\mu\nu}) = 0$ [2]. This exhibits the close relationship between energy conservation and time-travel, so the simple energy-conservation argument in Figure 9.5 need not hold. In fact, no general argument preventing time-travel can exist.

Counter-intuitively, and using the isometry algebra (9.93), it was shown [219] that the closed time-like curves in the Gödel universe nevertheless do not violate causality. In other cases, such as the closed time-like curves through the ergoregion of the Kerr geometry, paths that go through "wormholes" [☞ below] and many others [519, 265], where causality may be violated in principle, semi-classical arguments indicate that the quantum physics *probably* precludes violations of causality. However, based on such semi-classical arguments, Stephen Hawking hypothesized in 1992 that there exists a general chronology protection principle, except within the indeterminacy specified by Heisenberg's relations. Much milder is the hypothesis proposed by Igor Novikov back in 1975, whereby only self-consistent paths are permitted; this also includes traveling backwards in time if this does not cause a change in the existing history. A survey of these hypotheses and other practical, technical and conceptual questions related to closed time-like curves may be found in Refs. [542, 544]. Of course, as no complete theory of quantum gravity exists as yet, physical realizations of traveling along closed time-like curves and the physical realization of even chronology violation remain an open question☝.

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

9.3.4 Engineering spacetime, wormholes and topological bridges

Returning to the Einstein tensor (9.94) in the Gödel universe, which the Einstein equations equate with the energy–momentum tensor density of the matter/energy that maintains this geometry, points to an important property of the nonlinear system of Einstein equations (9.44):

> ***Conclusion 9.9*** *For each $i = 1, 2\ldots$, let $T^{(i)}_{\mu\nu}$ denote the energy–momentum tensor density for the ith matter/energy distribution, and $g^{(i)}_{\mu\nu}$ the corresponding solution of the Einstein equations (9.44). The joint matter distribution (if this is physically achievable) has the energy–momentum tensor density $\sum_i T^{(i)}_{\mu\nu}$ and the solution of the Einstein equations (9.44) $g^{(\Sigma)}_{\mu\nu}$. However, $g^{(\Sigma)}_{\mu\nu}$ is most often significantly different from either of the "partial" solutions $g^{(i)}_{\mu\nu}$, as well as from their sum.*

This property is intuitively acceptable: It should be the case that we can always freely combine different types of matter/energy (except that two macroscopic material objects, of course, cannot exist in the same place at the same time) and to add them to any initially given spacetime. The presence of additional matter/energy then must change the geometry of spacetime again in a way determined by the Einstein equations. However, the resulting metric tensor, in general, is not simply an analogous linear combination of metric tensors that follow from the presence of one or the other component energy–momentum tensor density. Succinctly,

> ***Conclusion 9.10*** *Energy–momentum density tensors of matter/energy distributions and their Einstein tensors are additive; the corresponding metric tensors are not.*

These conclusions rely on the usual interpretation of the Einstein equations as a differential system that determines the metric tensor as a function of a provided energy–momentum tensor density and initial and boundary data.

The converse approach partially follows from the logical sequence in Conclusion 9.8 on p. 345, and is sometimes referred to as the "engineering approach," wherein:

1. specify a desired geometry by specifying the corresponding metric tensor;
2. compute the Einstein tensor $G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R$ for this metric tensor;
3. this specifies the required energy–momentum tensor density $T_{\mu\nu}$ of the matter/energy distribution that produces/maintains the desired geometry by its presence.
4. Finally, explore:
   (a) What (physical/engineering) characteristics should this matter/energy distribution have, so as to have the required $T_{\mu\nu}$?
   (b) Is it (at least in principle) possible to construct a structure with the matter/energy distribution and the required $T_{\mu\nu}$?

For the purpose of classifying the types of matter/energy, the characterizing "energy conditions" were introduced. To define these conditions, we need:[23]

1. a time-like 4-vector field with components $\xi^\mu(\mathrm{x})$, i.e., $g_{\mu\nu}\xi^\mu\xi^\nu < 0$, $\forall \mathrm{x}$;
2. a light-like (or null) 4-vector field with components $k^\mu(\mathrm{x})$, i.e., $g_{\mu\nu}k^\mu k^\nu = 0$, $\forall \mathrm{x}$;
3. a causal 4-vector field with components $\zeta^\mu(\mathrm{x})$, i.e., $g_{\mu\nu}\zeta^\mu\zeta^\nu \leqslant 0$, $\forall \mathrm{x}$.

Since ξ^μ may be interpreted as a 4-vector that is tangential to the worldline of a massive particle, it follows that $\varrho := T_{\mu\nu}\xi^\mu\xi^\nu$ is the total mass–energy density (of the material particle as well as all

[23] Recall that the signature of the metric tensor $g_{\mu\nu}$ in the relativistic tradition followed in this chapter is the reverse of the signature of the metric tensor of flat spacetime, $\eta_{\mu\nu}$, used in the particle physics tradition; $g_{tt} < 0$ while $\eta_{tt} > 0$.

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

non-gravitational fields that act upon this particle in this spacetime point). Similarly, the quantity $\varrho_0 := T_{\mu\nu}k^\mu k^\nu$ is the limiting value of the mass–energy density ϱ for a massless particle/field.

The following "energy conditions" are used to typify matter/energy:

	Condition	**For all**
Dominant	$g^{\mu\nu}T_{\mu\rho}T_{\nu\sigma}\zeta^\rho\zeta^\sigma \leqslant 0$ and $g^{0\mu}T_{\mu\nu}\zeta^\nu < 0$	$g_{\mu\nu}\zeta^\mu\zeta^\nu \leqslant 0, \ (\zeta^0 > 0)$
Weak	$T_{\mu\nu}\xi^\mu\xi^\nu \leqslant 0$	$g_{\mu\nu}\xi^\mu\xi^\nu < 0$
Null*	$T_{\mu\nu}k^\mu k^\nu \leqslant 0$	$g_{\mu\nu}k^\mu k^\nu = 0$
Strong	$\big[T_{\mu\nu} - \frac{1}{2}g_{\mu\nu}T\big]\xi^\mu\xi^\nu \leqslant 0$	$g_{\mu\nu}\xi^\mu\xi^\nu < 0$

(9.95)

* The null condition is also often referred to as "light-like."

The relationship between these conditions is

$$\textbf{Dominant} \ \Rightarrow \ \textbf{Weak} \ \Rightarrow \ \textbf{Null} \ \Leftarrow \ \textbf{Strong}, \tag{9.96}$$

where it is important to note that, nomenclature to the contrary, the strong condition does not imply the weak, nor vice versa. These conditions also have their "averaged" version, where the integral of the condition over some spacetime region is satisfied although the condition is violated somewhere within the given region.

The Einstein–Rosen "bridge"

The Schwarzschild metric tensor (9.55) exhibits two pathological properties at the distance $r = r_S$:

1. the time component, $g_{00} = g_{tt} = -\big(1-\frac{r_S}{r}\big)c^2$ vanishes,
2. the radial component, $g_{rr} = -\big(1-\frac{r_S}{r}\big)^{-1}$ diverges.

In turn, as discussed in Section 9.3.1 on p. 334, the divergence or vanishing of an individual component of the metric tensor does not necessarily imply a real singularity in the geometry. Moreover, Lemaître's coordinates (9.60) prove that the location $r = r_S$ is not singular. This supports the nagging doubt that the familiar spherical coordinates (t, r, θ, φ) – and so maybe even Lemaître's – do not in fact describe the complete spacetime geometry in the vicinity of the black hole.

Also, a detailed analysis of the various trajectories of massive particles and light rays that pass through the event horizon [367] points to a very bizarre property, sketched in the diagram on the left-hand side of Figure 9.6: Particles directed towards the black hole follow spacetime paths that are seemingly disconnected when passing through the event horizon and require the coordinate time to diverge to $t \to +\infty$ (whereas the proper time remains finite), and the path segment within the event horizon to move backwards in coordinate time while computation proves that the proper time continues to pass forward for massless particles and to stagnate for light.

In Figure 9.6(a), follow a light ray directed at the black hole from an initial point A, as it passes through the point C in the coordinate time $t = 0$, passes through the horizon $(r = r_S)$ in coordinate time $t = +\infty$ at the "point" D, then *returns* in coordinate time, within the horizon, and falls into the $r = 0$ singularity in the spacetime point F. Namely,

$$\text{for} \ \ r < r_S, \qquad f_S(r) < 0 \quad \text{so} \ \ g_{tt} = -f_S(r) > 0 \ \ \text{and} \ \ g_{rr} = \big(f_S(r)\big)^{-1} < 0. \tag{9.97}$$

Thus, within the horizon, the coordinate t has a space-like character (particles may move in both directions of t) and the coordinate r has a time-like character, and particles may move only in the direction $r \to 0$, i.e., towards the singularity. Similarly to a light ray, a massive particle directed towards the black hole from the initial spacetime point B, passes through the point C in coordinate

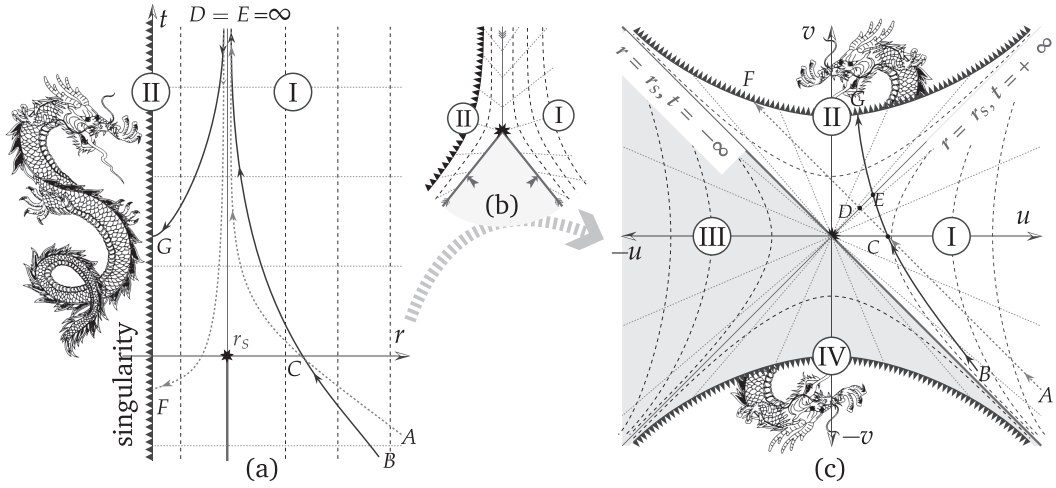

Figure 9.6 The Schwarzschild geometry (a) in the original (t, r) coordinates, (c) in Kruskal–Szekeres coordinates and (b) the transitionary stadium of the mapping from (a) to (c). A light ray directed toward the black hole follows the A-C-D-F path, while a particle with a non-vanishing mass directed toward the black hole follows the B-C-E-G path. The depiction (9.55) is spherically symmetric and angular coordinates θ, φ are not shown; every point in the figure lies on a sphere of the given radius. The diagram (b) shows how the diagram (a) “opens” in the mapping to the diagram (c).

time $t = 0$, passes through the horizon ($r = r_S$) in coordinate time $t = +\infty$ (at the “point” E), and then returns *retrograde* in coordinate time within the horizon, and falls into the $r = 0$ singularity (G). Throughout, the proper time of a massive particle passes forward, and remains finite.

Besides the appearance of a fictitious singularity at $r = r_S$, the discontinuity of the path – along which we know that the proper time is not discontinuous – also indicates that the Schwarzschild coordinates (t, r, θ, φ) are not appropriate. The Eddington–Lemaître coordinates (9.60) do remove the first but not also the second of these two problems. In 1950, John L. Synge discovered the incompleteness of the Schwarzschild coordinates, as well as a system of coordinates that is complete. Independently and unaware of Synge’s results, Christian Fronsdal again proved the incompleteness of Schwarzschild coordinates in 1959 (at CERN), and found a complete analytical description of the Schwarzschild geometry in the form of a higher-dimensional coordinate system with an algebraic constraint.[24] His solution turned out to be very similar to the solution that Martin Kruskal (at Princeton University) found a little earlier but did not publish, and of which D. Finkelstein and J. A. Wheeler (then professors at Princeton University) knew and to whom Fronsdal, in his original work [181], gave thanks for the communication. Independently from this group of explorers, the same solution was discovered also by Szekeres György, in Australia; the independent works by Kruskal and Szekeres were published in 1960 and this finite – and explicit – version of the description is today called the Kruskal–Szekeres diagram, and u and v in Figure 9.6(c), p. 350, are the Kruskal–Szekeres coordinates [367]. In turn, Fronsdal’s implicit description is today rarely mentioned.

[24] By definition, spaces of solutions of systems of algebraic equations are called *algebraic varieties* and form a major subject in the mathematical discipline of *algebraic geometry*. This connection between mathematics and physics will recur later, and much more vigorously, with the exploration of (super)strings.

The Schwarzschild and Kruskal–Szekeres coordinates are related as follows:

K–Sz		**Schwarzschild**	**K–Sz**		**Schwarzschild**	
$u_I,\ -u_{III}$	$=$	$\sqrt{\frac{r}{r_S}-1}\,e^{r/r_S}\cosh\left(\frac{ct}{2r_S}\right)$	$u_{II},\ -u_{IV}$	$=$	$\sqrt{1-\frac{r}{r_S}}\,e^{r/r_S}\sinh\left(\frac{ct}{2r_S}\right)$	(9.98a)
$v_I,\ -v_{III}$	$=$	$\sqrt{\frac{r}{r_S}-1}\,e^{r/r_S}\sinh\left(\frac{ct}{2r_S}\right)$	$v_{II},\ -v_{IV}$	$=$	$\sqrt{1-\frac{r}{r_S}}\,e^{r/r_S}\cosh\left(\frac{ct}{2r_S}\right)$	

$$\left(\frac{r}{r_S}-1\right)e^{r/r_S}=u^2-v^2,\qquad t=\begin{cases}\frac{2r_S}{c}\operatorname{arth}\left(\frac{v}{u}\right) & \text{in regions I and III;}\\ \frac{2r_S}{c}\operatorname{arth}\left(\frac{u}{v}\right) & \text{in regions II and IV;}\end{cases}\tag{9.98b}$$

where the subscript to Kruskal–Szekeres coordinates denotes the region in which the stated relation holds. By definition, $r \geqslant 0$, so the half-plane $(t,r)_{r<0}$ has no physical meaning. However, the half-plane $(t,r)_{r\geqslant 0}$ with the boundary $(r=0)$ is not geodesically complete – as was shown: paths that start outside the horizon, pass through the horizon and then fall into the singularity "pass" through the point at infinity and come back from it. In turn, the domain of Kruskal–Szekeres coordinates (shown in Figure 9.6(c), p. 350, as the part of the (u,v)-plane bounded by the singularity hyperbolas) is geodesically complete: All geodesic lines are either completely contained within this region or have a limiting point at infinity and outside the singularity hyperbolas. Also, every finite part of every geodesic path is entirely contained within the domain of Kruskal–Szekeres coordinates.

Figure 9.6(c), p. 350, is the Schwarzschild geometry presented in Kruskal–Szekeres coordinates (u,v): the half-plane $(t,r)_{r\geqslant 0}$ from Figure 9.6(a) is mapped into the region bounded by the "$r=r_S$, $t=-\infty$" diagonal and the upper singularity hyperbola. Figure 9.6(b) shows the "intermediate phase" between the Schwarzschild picture and the Kruskal–Szekeres picture, where one sees that:

1. the diagonal "$r=r_S$, $t=-\infty$" appears by "splitting" the lower Schwarzschild semi-axis $r=r_S$, $t\in(-\infty,0]$ into two semi-axes that then open into the "$r=r_S$, $t=-\infty$" diagonal;
2. the "splitting" of the lower Schwarzschild semi-axis $r=r_S$, $t\in(-\infty,0]$ provides the space of regions III and IV;
3. the upper Schwarzschild semi-axis $r=r_S$, $t\in[0,+\infty)$ becomes the semi-axis that divides the regions I and II, and its copy divides the regions III and IV.

The comparative examination of these two coordinates of the Schwarzschild geometry clearly demonstrates that the mapping $(t,r)_{r\geqslant 0}\xrightarrow{1-2}(u,v)$ is two-valued, i.e., that the Kruskal–Szekeres picture is a double covering of the Schwarzschild picture.

This double covering implies that every spacetime region with the Schwarzschild geometry there automatically must have an exact copy, and these two regions touch along the "$r=r_S$, $t=-\infty$" diagonal in the Kruskal–Szekeres picture. By means of Figure 9.6(b), p. 350, we see that in the Schwarzschild picture this means that the two copies of spacetime touch along the event horizon, but only up to the coordinate time $t=0$. As the coordinates may be changed by arbitrary general coordinate transformations [☞ Definition 9.1 on p. 319], the time $t=0$ of course has no invariant meaning and the moment when the two spacetime regions separate depends on the choice of the observer; the text [367] shows the detailed history of this process from the vantage point of two different observers.

Since $\operatorname{arth}(x)=\tanh^{-1}(x)=\sum_{k=0}^{\infty}\left(\frac{x}{k}\right)^{2k+1}$, in regions I and III and for sufficiently large but fixed u_*, we have that $t(u_*)\approx\frac{2r_S}{cu_*}v$, and the Kruskal–Szekeres coordinate v approximates the Schwarzschild time t. Thus, the Schwarzschild-simultaneous points in Kruskal–Szekeres Figure 9.6(c), p. 350, all lie on predominantly horizontal and approximately straight lines when "sufficiently deep" within the regions I and III;[25] in passing through the regions II and IV, these Schwarzschild-simultaneous points are depicted by the nonlinear curves in the Kruskal–Szekeres coordinates.

Figure 9.6 (c), p. 350, then clearly indicates that this depicts a dynamical process where, from the vantage point of a fixed observer outside the event horizon, a "bridge" (or tunnel) appears that connects the spacetime regions I and III. This process was discovered by Albert Einstein and Nathan Rosen in 1935, hence its name. However, only in 1962 did John A. Wheeler and Robert W. Fuller discover that this bridge is in fact an unstable configuration and that neither material objects nor light can pass through it. Because of this impassability and topological form $S^2 \times \mathbb{R}^1$ that is a 3-dimensional generalization of the cylinder ($S^1 \times \mathbb{R}^1$), these configurations became known as *wormholes*.

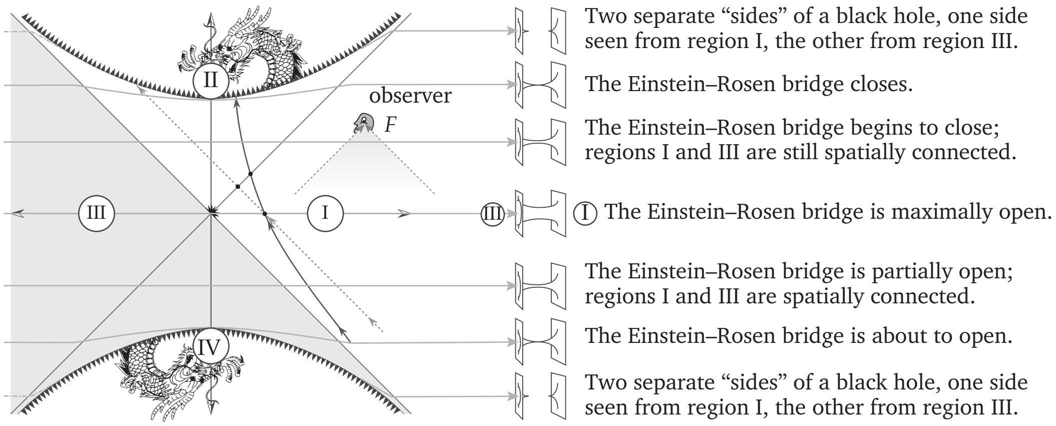
II
observer
F
III
I
IV
III
I
Two separate "sides" of a black hole, one side seen from region I, the other from region III.
The Einstein–Rosen bridge closes.
The Einstein–Rosen bridge begins to close; regions I and III are still spatially connected.
The Einstein–Rosen bridge is maximally open.
The Einstein–Rosen bridge is partially open; regions I and III are spatially connected.
The Einstein–Rosen bridge is about to open.
Two separate "sides" of a black hole, one side seen from region I, the other from region III.

Figure 9.7 The Einstein–Rosen "bridge" as a dynamical process. The ($\approx$ time) v coordinate distance between the lower (earlier) and upper (later) singularity has no physical meaning: Particles directed towards the "bridge" end up in the upper (future) singularity: massive particles follow the path depicted by the solid line, light follows the dashed one. The physically accessible regions I and III meet only at the Kruskal–Szekeres coordinate origin, usually thought of as the circumference of the "throat" of the bridge.

Figure 9.7 shows the Schwarzschild geometry in the Kruskal–Szekeres coordinates, where the Schwarzschild-simultaneous hypersurfaces are depicted as predominantly horizontal lines, which indicate to the right the status of the Einstein–Rosen bridge by a sketch of its cross-section. The lines $\mathscr{C}$ that connect the regions I and III through the Einstein–Rosen bridge always have a spatial character, i.e., tangent 4-vectors $\mathrm{V} \in T_{\mathrm{x}}(\mathscr{C})$ along these lines ($\mathrm{x} \in \mathscr{C}$) are space-like, $g_{\mu\nu}(\mathrm{x})V^{\mu}(\mathrm{x})V^{\nu}(\mathrm{x}) > 0$ for every $\mathrm{x} \in \mathscr{C}$. The diagram in Figure 9.7 shows that not even light rays – in the Kruskal–Szekeres coordinate system, light travels along straight 45° lines – can reach either from the inside of region I into the inside of region III, or the other way. The same is true of real, massive particles.

Only light rays that are entirely within the event horizon (diagonal lines that intersect in the center of the diagram in Figure 9.7) pass from the boundary of region I into the boundary of region

[25] Recall that the angular coordinates θ and φ are not depicted in the diagrams in Figure 9.6 on p. 350, so every point represents an entire sphere of indicated radius, and every line is then a 3-dimensional space of the $\mathbb{R}^1 \times S^2$ topology, where the radius of the sphere S^2 varies along the line $\mathbb{R}^1$, collapsing to a point only where this line $\mathbb{R}^1$ touches the singularity.

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

III and the other way around. However, these paths (of light-like character) are forever trapped in the event horizon.

> ***Conclusion 9.11*** *In spite of the existence of spatial connections (by paths to which all tangent vectors are of spatial character) between regions I and III, the Einstein–Rosen "bridge" is forever closed for real particles (which travel along paths of time-like character), including here light and all other gauge fields.*

> ***Comment 9.5*** *The Einstein–Rosen bridge, however, is* **not** *closed to virtual particles. This in principle permits an interference of wave-functions that permeate through the Einstein–Rosen bridge, and provides a form of Aharonov–Bohm effect: The spacetime for Feynman-esque integration over paths (histories) [☞ Procedure 11.1 on p. 416] is multiply connected and connects otherwise unreachable portions of the universe.*

Notice that the topology of spacetime is necessarily a dynamical concept since one of the dimensions is time-like. Abstractly, the 4-dimensional mathematical space of the physical spacetime is multiply connected, and the bridge is "always" present. However, the simultaneous points for any real physical observer, F, form a 3-dimensional subspace, $\mathscr{P}_{F,t}$, of space-like character, so that all tangent vectors $\mathrm{V} \in T_{\mathrm{x}}(\mathscr{P}_{F,t})$ to this subspace (for each $\mathrm{x} \in \mathscr{P}$) are space-like 4-vectors: $g_{\mu\nu}(\mathrm{x})V^{\mu}(\mathrm{x})V^{\nu}(\mathrm{x}) > 0$ for every $\mathrm{x} \in \mathscr{P}_{F,t}$. As the time t of the physical observer F passes, the topology of this subspace $\mathscr{P}_{F,t}$ varies, as sketched in the right-hand half of Figure 9.7 on p. 352. In the example of the Einstein–Rosen "bridge" in the Schwarzschild geometry, the two separated regions of space:

1. have a black hole each;
2. these two black holes connect in a moment;
3. the connection of these black holes opens into a space-like "bridge" (wormhole) of the $S^2 \times \mathbb{R}^1$ topology;
4. this "bridge" closes before even light can pass through it;
5. there remain two separated regions, with a black hole each.

It can, however, not be overstated that every real physical observer, F, can see only the events that can signal F from the interior of the "past" light cone $\mathscr{C}^{\wedge}_{F,t}$, the vertex ("here, now") of which is in the spacetime point $\mathrm{x}_{F,t}$. Figure 9.7 on p. 352 then makes it clear that no real physical observer can even see through an Einstein–Rosen bridge. Owing to the somewhat "instantaneous" nature of the Einstein–Rosen bridge, it vaguely recalls the instantons mentioned in Chapter 6 and the tunneling through them; see Footnote *16* on p. 248.

Stabilization of traversable wormholes

Recall that the Schwarzschild geometry solves the Einstein equations without an energy–momentum tensor density on the right-hand side. The above description of the Einstein–Rosen "bridge" shows that even the topology and geometry of otherwise empty spacetime may be highly nontrivial.

The geodesically complete picture of the Schwarzschild geometry [☞ Figures 9.6 on p. 350 and 9.7 on p. 352] indicates that the Einstein equations have solutions where the spacetime is topologically nontrivial. Namely, the regions I and III may be either regions in otherwise separate universes, or regions in the same universe, which are however arbitrarily far from one another as measured along any path that does not pass through the Einstein–Rosen "bridge." Concretely, suppose in a given moment one such "bridge" opens temporarily between a black hole near Earth and some black hole in this same spacetime, but in the Andromeda Galaxy. In this case our spacetime would become multiply connected, and the space would become momentarily multiply connected,

https://doi.org/10.1017/9781009291507.013 Published online by Cambridge University Press

as there exist closed paths that do pass through that “bridge” from Earth to Andromeda, and then return to Earth along a (much) longer way. In the moment when such a path (space-like, for the Einstein–Rosen “bridge” is impassable) exists, such a path cannot be continuously shrunk to a point. Alternatively, such a path is not the boundary of any surface that is entirely contained in the given spacetime.

This topological property is identical to the property of the surface of a torus, which contains closed paths that traverse the “big” or the “small” circle at least once, so they cannot be continuously deformed into a point. In contrast to such non-contractible paths, there also exist of course closed paths that are the boundaries of surfaces that are completely contained in the given space, and which then may be continuously contracted to a point. Thus “topologically” seen, the surface of a torus is equivalent to the surface of a 2-sphere to which was added a cylindrical “handle” (wormhole), as shown in Figure 9.8.

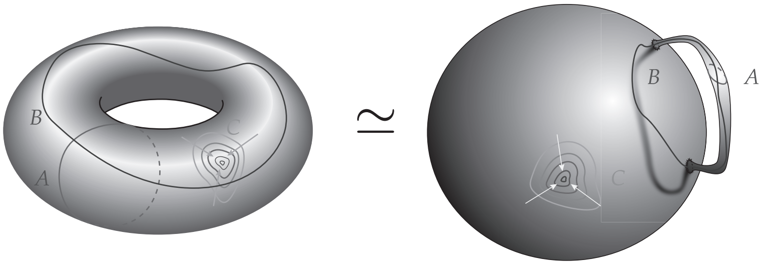

Figure 9.8 The torus surface with three topologically distinct closed paths: Neither *A* nor *B* can be continuously deformed into a point as can be done with the path *C*. Besides, the path *A* cannot be continuously deformed into the path *B*. The same holds for the “sphere with a handle” to the right, which is topologically equivalent to the torus.

In turn, that multiple connectedness – for real particles, fields and objects – has no practical meaning as the Einstein–Rosen “bridge” is impassable for them.

It is then reasonable to ask if there may exist some deformation of the Schwarzschild (or similar) geometry in which some such bridge between otherwise distant spacetime regions could exist and which would be traversable by real particles, fields and objects.

The metric tensor that exactly describes such a geometry evidently must have elements that are at least quadratic functions of at least some spatial coordinates, so that the spacetime solution would have two “branches,” i.e., “sheets,” which would then be connected by a tunnel, and so that in an adequate geodesically complete spacetime diagram (such as the Kruskal–Szekeres diagram for the Schwarzschild geometry) the otherwise separated regions of spacetime are connected through that tunnel by time-like paths. For solutions of this type the popular name “wormhole” was kept, but unlike the Einstein–Rosen space-like “bridge,” these time-like wormholes are named Lorentzian wormholes [541, 543].

The simplest example is provided by the metric tensor

$$\mathrm{d}s^2 = -c^2\mathrm{d}t^2 + \mathrm{d}\ell^2 + (k^2+\ell^2)\big(\mathrm{d}\theta^2 + \sin^2(\theta)\mathrm{d}\varphi^2\big), \tag{9.99}$$

where $r = \pm\sqrt{k^2+\ell^2}$ is the “true” radial coordinate, and $k > 0$ is a constant. For this metric tensor one computes the Einstein tensor, in spherical coordinates:

$$[G_{\mu\nu} = R_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}R] = \frac{k^2}{(k^2+\ell^2)^2}\mathrm{diag}\big[-c^2, -1, (k^2+\ell^2), (k^2+\ell^2)\sin^2(\theta)\big]. \tag{9.100}$$

The Einstein equations then equate this tensor with the energy–momentum tensor density of the matter/energy that is necessary at the connection of the two "branches" of the solution to maintain this geometry.

This use of the Einstein equations is identical to the use of the Gauss–Ampère equations in electrodynamics. There, the spherically symmetric electric field, for example, with a magnitude that decays as $1/r^2$ implies that there must exist an electric charge at the coordinate origin that maintains this field.

As the physical meaning of the T_{tt} component of the energy–momentum tensor density is the usual matter/energy density (including all non-gravitational fields), and T_{rr} is the radial pressure of this matter density, we see that the energy–momentum tensor density that is being equated with the result (9.100) must represent a very unusual matter/energy: both its density and its radial pressure are negative. However, in the original paper in 1989, Matt Visser [541] pointed out that there do exist physical systems that have been realized in laboratories, such as for the Casimir effect, and which exhibit at least some of these exotic properties. Later research in this respect discovered several other physical systems, the combinations of which could – in principle – be used to open and stabilize such Lorentz wormholes.

The fact that the matter/energy that maintains a traversable wormhole *must* have exotic properties follows from the simple insight [519]: When light enters a traversable wormhole, the rays are being focused towards a fictitious center, following the spacetime curvature caused by the gravitational effect of the energy/matter that maintains the wormhole traversable. The incoming rays therefore behave precisely as if they are gravitationally focused by the gravitational field of a massive object. In turn, when the light comes out on the "other side" of a traversable wormhole, the rays must be emanating as if they were welling from a center, following the spacetime curvature caused by the gravitational effects of the energy/matter that maintains the wormhole traversable. Effectively, these rays are then *refracted* by the gravitational field, indicating that the matter/energy density that maintains the wormhole traversable must be less than the density of empty, flat spacetime, i.e., must be negative.

The interested Reader should consult Refs. [546, 544] for additional examples and literature.

9.3.5 Exercises for Section 9.3

✎ ***9.3.1*** *Verify that the substitutions (9.77) eliminate the cosmological constant from the equations (9.76).*

✎ ***9.3.2*** *Adapting the relation (9.94), specify the proportion of cosmological constant and co-rotating perfect fluid that can emulate (**a**) dark energy, (**b**) quintessence, and (**c**) phantom energy.*

✎ ***9.3.3*** *Estimate the energy conditions (9.95) for (**a**) dark energy, (**b**) quintessence, (**c**) cosmological constant, and (**d**) phantom energy.*

✎ ***9.3.4*** *Determine which of the energy conditions (9.95) are violated by the matter/energy distribution required to support the Lorentzian wormhole (9.100).*