Relativity: The Special and General Theories

I include in these notes a course on the foundations of the general theory of relativity. The backbone is adapted from a pair of wonderful courses on special and general relativity given by Ian Benn at the University of Newcastle, Australia many years ago (of which I have very fond memories).

Further sources of which I have made use include [**1, 2, 3, 4, 5, 6, 7, 8**]. The books [**2, 6**] are excellent sources for special relativity. The book [**5**] is a highly regarded mathematics textbook on pseudo-Riemannian geometry (and relativity). The books [**3, 8**] are popular starting points for studies in general relativity, the former with more of a physics style, the latter more mathematically inclined. The book of Hawking and Ellis [**4**] deals with advanced material.

I note that basic concepts from pseudo-Riemannian geometry are taken for granted in these notes. Commonly, these concepts are developed alongside the development of the relativity theory. With the exception of the asterisked sections (§3.1, §8.2 and §8.3), this may be put off until §9.

> *"Come with us now on a journey through time and space..."*
> — The Mighty Boosh[1]

Mat Langford
Knoxville, December 2020

[1] 1998–2009.

Contents

1. Aristotelian spacetime

> *"Nature sets inanimate objects in motion until they reach their natural state of rest"* — Aristotle[2]

It is very intuitive to think of "events" as occurring at some point in "space" at some instant of "time". Since only differences in times are meaningful, it makes sense mathematically to model time as a *one dimensional affine space*, T. Since only relative positions are meaningful and since "space" seems to have three independent directions, it makes sense to model it with a *three dimensional affine space*, S.

Definition 1.1. *A (real) k-**dimensional affine space** is a triple $(A, \mathbb{A}, -)$, where A is a set of "points", $\mathbb{A}$ is a k-dimensional (real) linear space (the **space of displacements**), and*

$$\begin{aligned} - : A \times A &\to \mathbb{A} \\ (p, q) &\mapsto p - q \end{aligned}$$

is a map which satisfies

(1) for all $q \in A$, the map $p \mapsto p - q$ is a bijection; and

(2) for all $p, q, r \in A$,

$$(p - q) + (q - r) = p - r\,.$$

*This second condition is known as the **triangle rule**[3].*

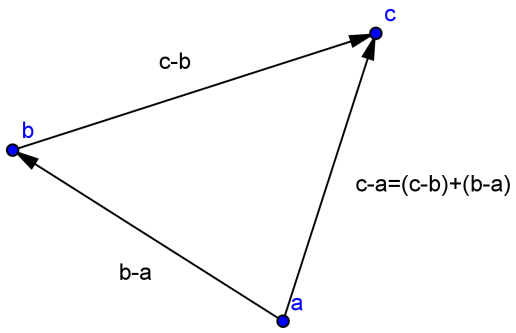

Figure 1.1. The triangle rule

[2]384–322 BCE. Aristotle's views have had a huge influence on later scientific and philosophical thought, as can be seen, for example, from the sheer volume of modern scientific terminology first coined in Aristotle's works. The text in quotation paraphrases Aristotle's views on motion from his work *Phusike akroasis* (*The Physics*). The model of "Aristotelian spacetime" expounded below is a modern mathematical charicature of this point of view.

[3]Of course, the rule dictates the notation, *not* the other way around: the affine structure "$-$" is not the additive inverse of any "$+$".

So we could model "events" as being points in the four dimensional affine space $A \doteqdot T \times S$. This is the cartesian product of the underlying sets with the "obvious" affine structure (modelled on $\mathbb{T} \times \mathbb{S}$)[4]

$$(t, p) - (s, q) \doteqdot (t - s, p - q)\,.$$

We may also equip T and S with Euclidean structures (i.e inner products on their spaces of displacements) corresponding to the existence of clocks and rulers respectively.

Physical objects such as electrons, galaxies, or apples may be modelled by curves

$$C\colon T \to S,$$

assigning to each time the[5] point at which the electron, galaxy, or apple lies in space. We call such curves (Aristotelian) **observers**, or **particles**. The graph $\{(t, C(t)) : t \in T\}$ is called the (Aristotelian) **worldline** of the observer.

The spacetime A inherits the following canonical projections:

$$\begin{aligned} \Pi_T\colon A &\to T \\ (t, p) &\mapsto t \end{aligned}$$

and

$$\begin{aligned} \Pi_S\colon A &\to S \\ (t, p) &\mapsto p\,. \end{aligned}$$

Given $t \in T$, the set $\Pi_T^{-1}(t) \subset A$, can be thought of as a set of **simultaneous** events (since each $p \in \Pi_T^{-1}(t)$ has the same "time coordinate", t). Since the set of simultaneous events corresponding to each time is determined naturally by the structure of A, we say that A has an **absolute simulteneity** or has **absolute time**.

Similarly, given $x \in S$, we can think of the set of events $\Pi_S^{-1}(x)$ as occurring at the same point, x, in space. We say that A is equipped with **absolute rest**, since an observer whose worldline coincides with $\Pi_S^{-1}(x)$ would be considered, by everybody, to be stationary.

Now, Aristotle taught that objects not acted on by "forces" remained at rest, whilst moving objects required external forces to keep them in motion. So we refer to A as **Aristotelian spacetime**. Since Aristotle also believed that the Earth is at rest at the centre of the universe, we could as well equip A with a preferred stationary observer (the worldline of the Earth). This corresponds to a preferred identification of S with its space of displacements, $\mathbb{R}^3$.

[4]The symbol "$\doteqdot$" means *equal to by definition*.

[5]Of course, this means that we're thinking of these objects as being "pointlike"

1.1. Problems with Aristotelian Spacetime. In his *Dialogue Concerning the Two Chief World Systems*, Galileo (via his protagonist Salviati) poses the following thought experiment:

> *Shut yourself up below decks on some large ship... have the ship proceed with any speed you like, so long as the motion is uniform and not fluctuating this way and that. You will discover not the least change in all the effects named, nor could you tell from any of them whether the ship was moving or standing still.*

This argument refutes the idea of absolute rest, and hence, if we are to accept Salviati's point of view, we must reject Aristotelian spacetime as a model for our world.

Galileo (and before Galileo, Copernicus) also refuted the idea that the Earth is at the centre of the universe. Notably, he used[6] *experimental evidence* to support his refutation: observations of the moons of Jupiter.

[6]In vain, as summarized in his famous response to the church *"E pur si muove"*.

2. Galilean spacetime

> *"He who attempts natural philosophy without geometry is lost."* — Galileo Galilei[7]

In order to incorporate the indistinguishability of rest and uniform motion into our spacetime model, we see that we need to banish the space projection Π_S from the Aristotelian construction. We may keep the time projection, however. So let's consider some "four-dimensional space" of events G equipped with an absolute time projection

$$\Pi_T \colon G \to T.$$

The presence of the time projection means that we retain absolute simultaneity: two events must be considered simultaneous if and only if they lie in the same $\Pi_T^{-1}(t)$.

Such a construction can be modelled mathematically by an *affine bundle*. We will require a little groundwork to define this object properly.

Definition 2.1. *A **fibre bundle** is a quadruple (E, B, π, F), where*

*(1) the **total space** E, the **base space** B, and the **typical fibre** F are topological spaces, and B is connected;*

*(2) the **bundle projection** $\pi : E \to B$ is a continuous surjection; and*

*(3) each point $p \in B$ admits a neighbourhood $U \subset B$ and a homeomorphism (called a **local trivialization**) $\phi : \pi^{-1}(U) \to U \times F$ for which $\pi_U(\phi(q)) = \pi(q)$, where π_U denotes projection onto the first factor.*

Note, in particular, that the **fibres** $E_p \doteqdot \pi^{-1}(\{p\})$ are homeomorphic to the typical fibre F.

We will often denote the fibre bundle (E, B, π, F) by $\pi : E \to F$, or simply by referring to the total space E.

Roughly speaking, a fibre bundle (E, B, π, F) is a space E which looks *locally* like a product $U \times F$ of a small "patch" U of B with the fibre F. "Globally", the structure may be different from the product. Indeed, the truncated cylinder $S^1 \times [-1, 1]$ and the Möbius band are both fibre bundles over S^1 with typical fibre the closed interval $[-1, 1]$. The latter is not globally a product due to its "twist".

It is important to note that, while the structure of a fibre bundle is that of a local product (via the local trivializations), there is in general no *canonical* choice of local trivialization.

[7] 1564–1642.

Modelling spacetime G as a fibre bundle $\Pi_T : G \to T$ over the one-dimensional affine space T removes the Aristotelian notion of absolute rest. However, we still want to be able to measure relative displacements and distances between simultaneous events. So, for each $t \in T$ (the base space), the **fibres** $G_t \doteq \Pi_T^{-1}(t)$ should be 3-dimensional affine spaces equipped with a Euclidean structure on their spaces of displacements, $\mathbb{G}_t$.

Definition 2.2. *A **vector bundle** is a fibre bundle for which*

(1) the typical fibre F is a linear space, and

(2) the map $v \mapsto \phi^{-1}(p, v)$ is a linear isomorhism between $\{p\} \times F$ and E_p for each $p \in E$ and each local trivialization map $\phi : U \mapsto U \times F$ about p.

By relaxing the linear structure to an affine one, we finally arrive at the concept of an *affine bundle*.

Definition 2.3. *Let $(\mathbb{E}, B, \pi_{\mathbb{E}}, \mathbb{F})$ be a vector bundle. An **affine bundle** modelled on $(\mathbb{E}, B, \pi_{\mathbb{E}}, \mathbb{F})$ is a fibre bundle (E, B, π_E, F) such that*

(1) the typical fibre F is an affine space modelled on $\mathbb{F}$;

(2) for each $p \in B$, the fibre E_p is an affine space modelled on $\mathbb{E}_p$; and

(3) the map $v \mapsto \phi(p, v)$ is an affine isomorhism between E_p and $\{p\} \times F$ for each $p \in E$ and each local trivialization map $\phi : U \mapsto U \times F$ about p.

We shall model **Galilean spacetime** by an affine bundle (G, T, Π_T, A) over the one-dimensional affine space T with fibre the three-dimensional affine space A. The affine structure on T allows us to measure temporal durations between events but in order to measure spatial displacements we must equip the space of displacements of A with a Euclidean structure. (This in turn induces a unique Euclidean structure on each of the fibres for which the local trivializations define isometries.)[8]

While each local trivialization provides a local product structure for G, and hence a class of observers "at rest", there are no *canonical* local splittings, and hence no canonical (or *absolute*) notion of rest in Galilean spacetime.

A **Galilean observer** will be modelled by a **section** of Galilean spacetime; that is, a continuous map $O\colon T \to G$ such that

$$\Pi_T(O(t)) = t\,. \tag{2.1}$$

The graph $\{(t, O(t)) \in G : t \in T\}$ is called the (Galilean) **worldline** of O.

[8] You shall be pleased to know that the mathematical structures of special and general relativity are actually simpler than that of Galilean spacetime!

The requirement $\Pi_T(O(t)) = t$ ensures that all observers have "synchronised their watches", in the sense that their "personal time" (the curve parameter) agrees with the "absolute time" singled out by the time projection.

Given two observers O and P, we can use the affine structure on the fibres (the 3-spaces of simultaneous events) to define the relative displacement $\mathbf{r}(t) \doteqdot P(t) - O(t)$ of P with respect to O. This is well-defined, since, by (2.1), $P(t)$ and $O(t)$ lie in the same fibre. The Euclidean structure then allows us to define their relative distance, $r(t) \doteqdot |\mathbf{r}(t)|$. We could try to define the relative velocity of P with respect to O in the obvious way:

$$\mathbf{v}(t_0) \doteqdot \lim_{t \to t_0} \frac{\mathbf{r}(t) - \mathbf{r}(t_0)}{t - t_0} \,.$$

However, whilst the denominator in the fraction on the right hand side of the definition makes sense, the numerator does not, since we are trying to take the difference of vectors in different spaces!

We need to add a little more structure to our model: an ability to compare displacement vectors at different times. Let's call such a structure a **parallelism** for G, since, conceptually, it tells us when directions at different times are "parallel". In order that "metre sticks remain metre sticks", we should also require that the parallelism consist of isometries (they should preserve the Euclidean structures of the fibres).

Definition 2.4. *A **parallelism** for G is a collection of isometries $\tau_{t_2,t_1} : \mathbb{G}_{t_2} \to \mathbb{G}_{t_1}$ for each pair $t_1, t_2 \in T$ such that*

$$\tau_{t_1,t_0} \circ \tau_{t_2,t_1} = \tau_{t_2,t_0}$$

for any three times $t_0, t_1, t_2 \in T$.

If G is equipped with a parallelism τ, then we can define the relative velocity of an observer P with respect to another observer O by

$$\mathbf{v}(t_0) \doteqdot \lim_{t \to t_0} \frac{\tau_{t,t_0}(\mathbf{r}(t)) - \mathbf{r}(t_0)}{t - t_0} \,.$$

If the velocity of P with respect to O is constant, in the sense that

$$\tau_{t,t_0}\mathbf{v}(t) = \mathbf{v}(t_0) \;\text{ for each }\; t, t_0 \in T \,,$$

then P is said to be in the same **inertial class** as O. If the relative velocity is zero, then we may say that O and P are *comoving*, since their relative displacement remains constant.

Conversely, one may obtain a parallelism by setting up a family of comoving observers. Indeed, let $\mathcal{O} = \{O_x\}_{x \in A}$ be a family of **co-moving** observers that **foliate** G. That is,

(1) each pair of observers O_x, O_y remains equidistant:

$$|O_x(t_2) - O_y(t_2)| = |O_x(t_1) - O_y(t_1)| \text{ for all } t_1, t_2 \in T$$

and

(2) every event in G lies on the worldline of exactly one observer from the family.

Then we may define a parallelism by "transporting along the foliation". That is,

$$\tau_{t_2,t_1}(O_x(t_2) - O_y(t_2)) \doteq O_x(t_1) - O_y(t_1)\,.$$

The parallelism τ identifies G with the affine space $T \times A$, every point $p = O_x(t)$ identified with (t, x). This affine space also has a space projection $O_x(t) \mapsto x$, and can be thought of as O's personal Aristotelian spacetime. Of course, *any* observer can set up their own personal point of view in this way.

The (strong) **Galilean principle of relativity** asserts that *there is a special inertial class of observers with respect to which physical laws hold good in their simplest form*[9]. Observers in this special inertial class are called **inertial observers**.

We interpret this as meaning that we should equip Galilean spacetime with a parallelism τ and a preferred inertial class[10] of observers (respect to which the Galileo–Newtonian laws of physics should be formulated).

2.1. Inertial coordinates. Any observer O can define coordinates on G adapted to his worldline. Choosing a time origin $t_0 \in T$, he defines a time coordinate $t : G \to \mathbb{R}$ using the projection map and the affine structure on T:

$$t(p) = \Pi_T(p) - t_0\,.$$

Choosing an orthonormal basis $\{e_1, e_2, e_3\}$ for $\mathbb{G}_{t_0}$ he defines three space coordinates $x^i : G \to \mathbb{R}$ using the Euclidean affine structure and the Galilean parallelism:

$$x^i(p) \doteq \big(p - O(\Pi_T(p))\big) \cdot e_i(t)\,,$$

where $e_i(t) \doteq \tau_{t_0,t}(e_i)$ is the parallel translate of e_i, and $\cdot$ is the inner product on the fibres.

[9] Albert Einstein, *The Foundation of the General Theory of Relativity*

[10] According to Mach, the inertial class is determined by the distribution of *all* the matter in the universe. Mach justifies this point of view with the following question: *If there is but a single particle in the universe, how can it distinguish rotation from non-rotation? Linear acceleration from rest?*

Of course, these coordinates depend on the choice of observer and his choices for t_0 and $\{e_1, e_2, e_3\}$. If $\hat{O}$ is a second observer, then she may define her own "hat" coordinates by choosing her own time origin $\hat{t}_0 \in T$ and orthonormal basis $\{\hat{e}_i\}_{i=1}^3$ for $\mathbb{G}_{\hat{t}_0}$:

$$\begin{aligned}\hat{t}(p) &= \Pi_T(p) - \hat{t}_0 \\ \hat{x}^i(p) &= \big(p - \hat{O}(\Pi_T(p))\big) \cdot \hat{e}_i(\Pi_T(p))\,.\end{aligned}$$

where $\hat{e}_i(t)$ is the parallel translate of $\hat{e}_i$.

Let's assume that the relative velocity of $\hat{O}$ with respect to O is uniform, in the sense that $\mathbf{v}(t) = \tau_{t,t_0}(\mathbf{v}_0)$ for some $\mathbf{v}_0 \in \mathbb{G}_{t_0}$ (this is the case, e.g., if both O and $\hat{O}$ are inertial observers).

We can relate the two time coordinates as follows:

$$\hat{t}(p) = \Pi_T(p) - \hat{t}_0 = \Pi_T(p) - t_0 + (t_0 - \hat{t}_0) = t(p) - a\,,$$

where we've defined $a \doteqdot \hat{t}_0 - t_0$.

Since any two orthonormal bases are related by an orthogonal transformation, we have $\hat{e}_i(t_0) = B(e_i(t_0))$ for some orthogonal transformation $B \in O(\mathbb{F}_{t_0})$.

Now, since the relative velocity of the two observers is uniform, we have,

$$\begin{aligned}(t_0 - t_1)\mathbf{v}_0 &= \tau_{t_0,t_1}\Big(\hat{O}(t_0) - O(t_0)\Big) - \Big(\hat{O}(t_1) - O(t_1)\Big) \\ \Rightarrow \Big(\hat{O}(t_1) - O(t_1)\Big) \cdot e_i(t_1) &= \Big(\hat{O}(t_0) - O(t_0)\Big) \cdot e_i(t_0) + (t_1 - t_0)\mathbf{v}_0 \cdot e_i\,,\end{aligned}$$

which gives the component equations

$$\Big(O(t_1) - \hat{O}(t_1)\Big)^i = \Big(\hat{O}(t_0) - O(t_0)\Big)^i + (t_1 - t_0)v_0^i\,.$$

Now, putting this together, and writing $\Pi_T(p) = \tau$, we obtain

$$\begin{aligned}\hat{x}^i(p) &= \sum_{j=1}^3 B_{ij}\left(x^j(\tau) - \Big(\hat{O}(t_0) - O(t_0)\Big)^i - (\tau - t_0)v_0^i\right) \\ &= \sum_{j=1}^3 B_{ij}\left(x^j(\tau) - A^j - t(p)v_0^j\right)\,,\end{aligned}$$

where we have defined $A \doteqdot \hat{O}(t_0) - O(t_0)$. Therefore, observers in the same inertial class may relate coordinates via a transformation of the form:

$$\begin{cases}\hat{t} = t - a \\ \hat{x}^i = \sum_{j=1}^3 B_{ij}(x^j - A^j - tv_0^j)\,.\end{cases}$$

These transformations form a 10-parameter group, which we call the **Galilean group**. The assertion that the laws of physics should be the same

for inertially related observers therefore requires, in particular, that any coordinate formulation of these laws should be invariant under the action of the Galilean group.

2.2. Galileo–Newtonian mechanics. Galilean velocities add in the obvious way: for any three Galilean observers, O, P, Q, the velocity of Q with respect to O is the velocity of Q with respect to P plus the velocity of P with respect to O. That is,

$$\dot{Q}_O = \dot{Q}_P + \dot{P}_Q \,,$$

where, for example,

$$\dot{Q}_O(t_0) \doteqdot \lim_{t\to t_0} \frac{\tau_{t,t_0}\left(Q_O(t)\right) - Q_O(t_0)}{t - t_0} \,,$$

and

$$Q_O(t) \doteqdot Q(t) - O(t)$$

are the relative velocity and displacement vectors of Q with respect to O.

In particular, (relative) velocities are observer dependent. On the other hand, within an inertial class, the acceleration (of any observer) does not depend on the choice of inertial observer: let O, P, and C be observers, such that O and P are in the same inertial class. Define the **relative acceleration** of C with respect to O by

$$\ddot{C}_O(t_0) = \lim_{t\to t_0} \frac{\tau_{t,t_0}\left(\dot{C}_O(t)\right) - \dot{C}_O(t_0)}{t - t_0} \,,$$

and similarly for P. Then $\ddot{C}_P = \ddot{C}_O$.

The **absolute acceleration** (or, simply, the **acceleration**) $\ddot{O}$ of an observer O is taken to be its acceleration with respect to the inertial observers.

Now, if we define a **mechanical particle** to be a pair (C, m) consisting of an observer C and a number $m > 0$ (called the **inertial mass** of C), then, according to Newton, we should attribute any non-inertial behaviour of the particle to the existence of a **force**, $\mathbf{f}$, via **Newton's second law of motion**:

$$m\ddot{C}(t) = \mathbf{f}(t) \,, \tag{2.2}$$

where $\ddot{C}$ is the absolute acceleration of C. We remark that this equation is really the definition of the concept of a *force* (acting on C). The picture is completed if we are able to develop an understanding of the nature of $\mathbf{f}$ (independently of C).

The preceding discussion should sound very familiar (albeit written, perhaps, in an unneccessarily complicated way!), and the Galileo–Newtonian

formalism we have set up has been a very successful description of mechanical phenomena. However, as it stands, it has two major problems. The first of these involves gravitation, and the second involves electromagnetism.

Exercises.

Exercise 2.1. *Let O, P, and C be observers, such that O and P are in the same inertial class. Define the* ***relative acceleration*** *of C with respect to O by*

$$\ddot{C}_O(t_0) = \lim_{t \to t_0} \frac{\tau_{t,t_0}\left(\dot{C}_O(t)\right) - \dot{C}_O(t_0)}{t - t_0}\,,$$

and similarly for P. Show that

$$\ddot{C}_P = \ddot{C}_O\,.$$

Exercise 2.2. *Prove that for any three Galilean observers, O, P, Q, the velocity of Q with respect to O is the velocity of Q with respect to P plus the velocity of P with respect to O. That is,*

$$\dot{Q}_O = \dot{Q}_P + \dot{P}_Q\,,$$

where, for example,

$$\dot{Q}_O(t_0) \doteqdot \lim_{t \to t_0} \frac{\tau_{t,t_0}\left(Q_O(t)\right) - Q_O(t_0)}{t - t_0}\,,$$

and

$$Q_O(t) \doteqdot Q(t) - O(t)$$

are the relative velocity and displacement vectors of Q with respect to O.

Exercise 2.3. *Show that, for an arbitrary observer x, and a particle (C, m),*

$$m\left(\ddot{C}_x(t) + \ddot{x}(t)\right) = \mathbf{f}(t)\,,$$

where $\ddot{C}_x$ is the acceleration of C with respect to x and $\ddot{x}$ is the absolute acceleration of x.

3. Newtonian gravity

> *"Amicus Plato – amicus Aristoteles – magis amica veritas"*
> — Isaac Newton[11], Quaestiones Quaedam Philosophicae [Certain Philosophical Questions] (c. 1664)

It can be argued that Newton's greatest achievement was his *universal law of gravitation*, which unified the celestial motions of stars with the terrestrial motions of apples by asserting that the force exerted on a particle, C say, with **gravitational mass** $\mu > 0$ by a second particle, D say, with gravitational mass $\Omega > 0$ is given by the famous law

$$\mathbf{f}(t) = -G\frac{\mu\Omega}{r(t)^3}\mathbf{r}(t)\,, \tag{3.1}$$

where $\mathbf{r} \doteqdot C - D$ is the relative displacement of C from D, $r(t) = |\mathbf{r}(t)|$ is their relative separation, and G is the gravitational constant, which may be set[12] to 1 by choosing appropriate units. We think of D as generating the **gravitational field** $g = -\Omega\mathbf{r}/r^3$ on spacetime, where $\mathbf{r}(p) \doteqdot p - D(\Pi_T(p))$ is the displacement vector of an event p from D and $r(p)$ its norm. So we have the more general form of (3.1):

$$\mathbf{f}(t) = \mathbf{F}(C(t)) \doteqdot -\mu\mathbf{g}(C(t))\,, \tag{3.2}$$

where $\mathbf{g}$ is some gravitational field on spacetime (generated by some massive body, say).

Suppose that a particle C is moving in a gravitational field as above. A passer-by, x, notes a relative acceleration $\ddot{C}_x$ of C with respect to herself. Assuming that gravity is the only force present, in the form of a gravitational field $\mathbf{g}$, she deduces that

$$m(\ddot{C}_x(t) + \ddot{x}(t)) = -\mu\mathbf{g}(C(t))\,,$$

where m is the inertial mass of C and $\ddot{x}$ is the acceleration of x with respect to the inertial class. Rearranging, we obtain

$$\ddot{C}_x(t) = -\left(\ddot{x}(t) + \frac{\mu}{m}\mathbf{g}(C(t))\right). \tag{3.3}$$

Now x might hope to use (3.3) to measure her own (instantaneous) acceleration $\ddot{x}$ with respect to the inertial class as follows: she takes two cannon balls, C and D, of respective inertial masses ℓ and m, and respective gravitational masses λ and μ. Then, by dropping the cannon balls (from a conveniently oblique local tower perhaps) at time 0, she may determine the gravitational

[11] 1643–1727.

[12] Later we will change it to some integer multiple of π.

field **g** by eliminating her acceleration from the two equations resulting from (3.3). She obtains

$$\ddot{C}_x(0) - \ddot{D}_x(0) = -\left(\frac{\lambda}{\ell} - \frac{\mu}{m}\right) g(x(0)) .$$

Thus, if x can measure the relative instantaneous acceleration of C and D with respect to herself, as well as the difference between their respective gravitational to inertial mass ratio, then she may determine **g**, and hence, by (3.3), her own absolute acceleration, *so long as the difference in mass ratios of C and D is non-zero.*

Unfortunately, her hopes are dashed by her observation that all cannon balls appear to experience identical acceleration, and hence apparently have the same mass ratios. The Eötvös experiments[13] verified that the mass ratio of any "ordinary" matter is always, in appropriate units, equal to unity (to one part in 20 million!)[14].

In summary, *the motions of freely falling*[15] *particles with respect to a uniformly accelerated frame are indistinguishable (by means of local experiments) from the motions of freely falling particles in a corresponding gravitational field.* This is called the **equivalence principle**.

The assertion of the equivalence principle suggests that the preferred class of observers should be the *freely falling observers* (rather than the inertial ones). That is, the laws of physics should look the same to all freely falling observers (henceforth **freefallers**).

The observer x could brush the problem aside by viewing her (unknown and indeterminable) acceleration $\ddot{x}$ with respect to the inertial class as part of the gravitational field $\mathbf{g}(x)$: since

$$\ddot{C}_x(t) = -\left(\ddot{x}(t) + \mathbf{g}(C(t))\right) , \tag{3.4}$$

we may rewrite (3.2) as the observer-dependent equation

$$\mathbf{f}_x(t) = -\mu \mathbf{g}_x(C(t)) \doteq -\mu(\ddot{x} + \mathbf{g}(C(t)))$$

and hence recover (assuming $\mu = m$)

$$m\ddot{C} = \mathbf{f}_x.$$

Now, since the inertial observers were precisely those curves satisfying $\ddot{C} = 0$ with respect to the chosen parallelism, the relative acceleration of two inertial observers is zero, and hence Newton's law of motion certainly looks the same to all inertial observers. This won't be the case for freefallers,

[13] Eötvös (1885, 1889).

[14] By the 1980s, the methods were improved to obtain an accuracy to 1 part in 100 billion!

[15] That is, not undergoing non-gravitational forces.

however, since the difference in acceleration of two freefallers is in general non-zero.

In the following section, we will see that it is possible to modify the Galilean parallelism so that all freely falling observers satisfy the modified second law $C'' = 0$ (with the dashes appropriately interpreted). Later, we shall see that Galilean relativity *cannot* be adjusted to accommodate the electromagnetic force.

3.1. The Newtonian connection*. We wish to reformulate the Galileo–Newtonian framework using the *tangent bundle* to G instead of its spaces of displacements.

Recall that any observer O can set up coordinates $\{t, x^i\}_{i=1}^3$ for Galilean spacetime adapted to his worldline using a parallel orthonormal frame $\{e_i\}_{i=1}^3$. These coordinates provide G with the structure of a *smooth manifold*. If we denote by $\{\partial_t, \partial_{x^i}\}_{i=1}^3$ the corresponding coordinate basis vectors and identify T with $\mathbb{R}$ using the time origin, then we find that the tangent vector C' to any other observer $C : \mathbb{R} \to G$ decomposes as

$$C'(t) = (\partial_t + \dot{C}_O^i \partial_{x^i})\big|_{C(t)} ,$$

where $\dot{C}_O^i$ are the components of $\dot{C}_O$ with respect to the frame $\{e_i\}_{i=1}^3$.

We need a way to differentiate tangent vector fields like C', and this boils down to equipping the tangent bundle TG with a connection ∇. Now, we want ∇ to be compatible with both the Galilean parallelism and the Euclidean connection on the fibres, and these two conditions require, respectively, that

$$\nabla_{\partial_t} \partial_{x^i} = \nabla_{\partial_{x^i}} \partial_t = 0$$

and

$$\nabla_{\partial_{x^i}} \partial_{x^j} = 0$$

for each $i, j = 1, 2, 3$. This leaves one coefficient, $\nabla_{\partial_t} \partial_t$, undetermined, and this is just what we need to accomodate the equivalence principle!

Using the compatibility conditions, the acceleration of C becomes

$$C'' \doteq \nabla_{C'} C' = \nabla_{\partial_t} \partial_t + \ddot{C}_O^i \partial_{x^i} ,$$

where $\ddot{C}_O^i$ are the components of $\ddot{C}_O$ with respect to the frame $\{e_i\}_{i=1}^3$. If we set

$$\nabla_t \partial_t = \Gamma^i \partial_{x^i} ,$$

where the components Γ^i are given by

$$\Gamma^i(p) = (\ddot{O}(t) + \mathbf{g}(O(t)))^i , \quad t = \Pi_T(p) ,$$

then, by (3.4), freely falling observers are precisely those observers which satisfy the equation

$$C'' = 0\,. \tag{3.5}$$

We note that, even though we used coordinates adapted to a particular freely falling observer to construct ∇, equation (3.5) is freefaller independent.

Now, by the equivalence principle, we may assume that $\mathbf{g}_O \doteqdot \ddot{O} + \mathbf{g}$ is a gravitational field. Thus, according to Poisson, we should rewrite

$$\nabla_{\partial_t}\partial_t = -\operatorname{grad}\phi_O$$

for some **gravitational potential** $\phi_O : G \to \mathbb{R}$, where grad is the fibrewise gradient operator, which differentiates in the direction of the fibres[16]. Defining the **mass distribution** $\rho \doteqdot -\operatorname{div}\mathbf{g}$, we obtain **Poisson's equation**

$$\rho = \Delta\phi_O\,, \tag{3.6}$$

where $\Delta \doteqdot \operatorname{div}\operatorname{grad}$ is the fibrewise Laplacian.

We remark that there is no post-Newtonian physics going on here: the Newtonian connection is simply a convenient (and elegant) way of repackaging Newtonian physics in such a way that Newton's second law looks the same to all *freely falling* observers.

Exercises.

Exercise 3.1. *Show that the Newtonian connection is independent of the choice of freely falling observer.*

Exercise 3.2.

(1) Let $\{t, x^i\}_{i=1}^3$ be coordinates adapted to the worldline of an observer O. Show that the only non-zero components of the curvature tensor of the Newtonian connection are

$$\operatorname{Rm}(\partial_t, \partial_{x^i})\partial_t = -\operatorname{Rm}(\partial_{x^i}, \partial_t)\partial_t = \operatorname{Hess}\phi_O(\partial_{x^i})\,,$$

where $\operatorname{Hess}\phi_O$ is the fibrewise Hessian of the gravitational potential ϕ_O determined by O.

(2) Deduce that the Newtonian connection is not compatible with any metric on TG.

This seems to rule out the possibility of constructing a sensible Newtonian connection in special relativity.

[16] And we identify $\operatorname{grad}\phi_O$ with a tangent vector field according to the above procedure.

4. Electromagnetism in Galilean spacetime

> "*It appears, from all that precedes, reasonably certain that if there be any relative motion between the Earth and the luminiferous ether, it must be small; quite small enough entirely to refute Fresnel's explanation of aberration.*" — Albert A. Michelson[17] and Edward W. Morley[18] (*On the Relative Motion of the Earth and the Luminiferous Ether.* American Journal of Science, 1887, **34** (203): 333–345).

In the 19th century, the nature of electromagnetic interaction came to be rather well understood, with the discovery and formulation of Maxwell's equations and the Lorentz force law.

We modeled a force acting on an observer by a "time-dependent vector field" (i.e. a map $\mathbf{f} : T \to \mathbb{G}$ satisfying $\mathbf{f}(t) \in \mathbb{G}_t$), guiding the observer's motion according to Newton's second law. The second law asserts that all non-inertial motion is to be attributed to a force, but says nothing about the nature or origin of the force. Newton's universal law of gravitation provided an explanation for the gravitational force, by asserting that the gravitational force acting on a massive particle arises from the "space-time dependent vector field" $\mathbf{g} \doteqdot M\mathbf{r}/r^3$ via the rule $\mathbf{f}(t) = -m\mathbf{g}(C(t))$.

Let us define a **vector field** on Galilean spacetime to be a map $V : G \to \mathbb{G}$ satisfying $V(p) \in \mathbb{G}_{\Pi_T(p)}$ for each $p \in G$ (i.e. at each point $p \in G$, we attach a vector $V(p)$ from the space of displacements corresponding to the space of events simultaneous with p).

In the 19th century, it was discovered that magnets affect the motion of charged particles, with the force exerted depending not just on position, but also on the particle's velocity. The culmination of these and other experiments was the formulation of Maxwell's equations and the Lorentz force law.

According to Maxwell, the electromagnetic force is determined by **electric** and **magnetic fields E** and **B** via the Lorentz force law (which we will come to below). These "fields" are vector fields which solve Maxwell's equations:

$$\begin{cases} \operatorname{div} \mathbf{B} = 0 \\ \dfrac{\partial \mathbf{B}}{\partial t} + \operatorname{curl} \mathbf{E} = 0 \\ \operatorname{div} \mathbf{E} = \rho \\ \dfrac{\partial \mathbf{E}}{\partial t} - \operatorname{curl} \mathbf{B} = -\mathbf{j}\,, \end{cases} \tag{4.1}$$

[17] 1852–1931.
[18] 1838–1923.

where $\rho : G \to \mathbb{R}$ is some specified function (called the **charge density**), and j is some specified vector field (called the **current density**). The operators div and curl are interpreted via the divergence and curl operators on the spaces of displacements, $\mathbb{G}_t$, at each time $t \in T$. In order to make sense of the time derivative ∂_t, we need to specify a parallelism for the fibres F_t at each time, and this amounts to choosing an inertial class of observers, or equivalently, some inertial coordinates (t, x^1, x^2, x^3) (with respect to some inertial basis $\{e_1, e_2, e_3\}$). We then define, for example,

$$\frac{\partial \mathbf{B}}{\partial t} \doteqdot \frac{\partial B^i}{\partial t} e_i(t)\,,$$

where B^i are the components of $\mathbf{B}$ with respect to the basis: $\mathbf{B} = B^i e_i$.

Observe that, in the absence of charges, ρ, and currents, $\mathbf{j}$, the electric and magnetic fields both satisfy the wave equation:

$$\frac{\partial^2 \mathbf{E}}{\partial t^2} - \Delta \mathbf{E} = 0 \;\text{ and }\; \frac{\partial^2 \mathbf{B}}{\partial t^2} - \Delta \mathbf{B} = 0\,.$$

Many experiments of the late 1800s provided evidence that light (and its recently discovered relatives) exhibits properties of such waves in the electromagnetic fields.

The **Lorentz force law** ties the electromagnetic fields to the Newtonian formalism by asserting the existence of **charged** (**massive**) **particles** (C, m, q), where (C, m) is a massive particle and $q \in \mathbb{R}$ is its **electric charge**, upon whom the **electromagnetic force** acts via the expression

$$\mathbf{f}(t) = q\big(\mathbf{E}(C(t)) + \dot{C}(t) \times \mathbf{B}(C(t))\big),$$

where $\times$ is the cross product on the fibres and $\dot{C}$ is the velocity of C with respect to the "laboratory frame".

When we equate the force to the acceleration via Newton's law, we find that

$$m\ddot{C} = q\big(\mathbf{E}(C(t)) + \dot{C}(t) \times \mathbf{B}(C(t))\big)\,, \tag{4.2}$$

where $\ddot{C}$ is the acceleration with respect to the inertial frame.

Something is amiss: the left hand side of (4.2), does not depend on the inertial class but the right hand side does! So the Lorentz force law cannot take the same form for all inertial observers. That's not all: it is readily verified that Maxwell's equations are not invariant under the Galilean transformations (and hence do not take the same form for all inertial observers either)! So the Maxwellian electromagnetic theory appears to violate the Galilean principle of relativity.

What's worse, the famous experiments of Michelson and Morley at the end of the 19th century found *no observer dependent variation in the velocity*

of light (which, as we have mentioned, exhibits properties of electromagnetic waves). This observation is a violation the Galilean law of addition of velocities, and is thus fundamentally incompatible with the Galilean structure of spacetime.

Exercises.

Exercise 4.1. *Show that Maxwell's equations are not invariant under Galilean transformations.*

Exercise 4.2. *Show that, in the absence of charges, ρ, and currents, $\mathbf{j}$, the electric and magnetic fields both satisfy the wave equation:*

$$\frac{\partial^2 \mathbf{E}}{\partial t^2} - \Delta \mathbf{E} = 0 \quad \textit{and} \quad \frac{\partial^2 \mathbf{B}}{\partial t^2} - \Delta \mathbf{B} = 0\,.$$

5. Minkowskian spacetime

> *The views of space and time which I wish to lay before you have sprung from the soil of experimental physics, and therein lies their strength. They are radical. Henceforth space by itself, and time by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality.* — Hermann Minkowski[19], 1908

To measure the speed of light, in units of metres per second, say, we would have to time light over a set distance. So we assume that we know how to measure time in seconds and distance in metres. The anomaly of the constancy of the speed of light must then correspond to some inconsistency in this presupposed method for measuring velocities.

Now suppose that you are working as a patent clerk in Geneva. Then you *know* there is nothing wrong with your measurement of seconds: you are wearing a Swiss watch! Supposing further that Maxwell and Michelson–Morley are correct in saying that light travels with constant speed c ms^{-1}, then there must be some problem with our measurement of distances. The resolution to our problem is now very simple: we must measure distances using the speed of light! That is, we define $1\text{m} = \frac{1}{c}$ ls, where one light second (ls) is the distance traversed by light in one second. In these units, the speed of light is equal to c ms^{-1}, or, in more natural units, one light second per second!

So what does it mean to measure distance in light seconds? In the 1970s, NASA sent Apollo missions to the moon to, among other things, place a large mirror on its surface. From Earth, a laser was fired at the moon, and the time taken for its return trip measured (with Swiss watches situated on the Earth's surface). Those watches counted approximately 2.5 seconds for the return journey, putting the moon 1.25 light seconds away[20].

There can be no mystery about the constancy of the speed of light if we measure distances this way. If we instead try to define distances using, say, rigid measuring rods, then are we measuring the same thing (albeit in different units)? Or is there some other sort of distance to be measured? One can interpret experiments such as the Michelson–Morley experiment by saying that *these two notions of distance (radar ranging and rigid measuring rods) are equivalent*[21]. Surely this is what we should have expected — for steel rules are really held together by electromagnetic interactions!

[19] 1864–1909.

[20] *On average* (we may really only deduce that the distance to the Moon *and back* is 2.5 light seconds).

[21] Perhaps we could call this the (**electromagnetic**) **equivalence principle**.

If we agree to measure distances by radar ranging, then we reduce distance measurements to time measurements. As we will see, it follows from the finiteness of the speed of light that these time measurements cannot accord with Galileo–Newtonian ideas. In particular, we are forced to give up the existence of absolute simultaneity measured by absolute time.

5.1. Simultaneity is relative. Suppose that we observe a duel, fought with pistols, and we wish to determine which duellist fired first. We could set up delicate instruments to exactly time the arrival of the muzzle flashes from the pistols. We could then determine who fired first according to which flash we observed first. But that would not be quite right. Since we know that light travels at a finite speed, the time of arrival of the photons from the muzzle flashes will depend upon the distance travelled. So if, for example, both flashes arrive together, we would deem the person furthest away from us to have fired first. Thus we cannot decide who fired first unless we know how to measure distances. We have agreed to do this by radar ranging.

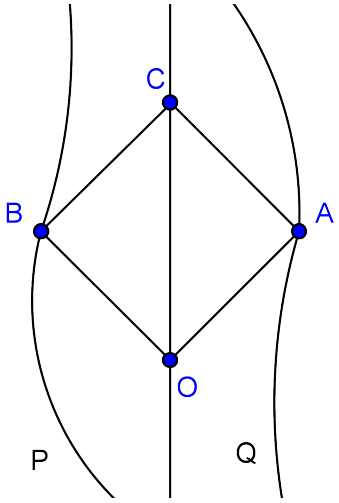

Figure 5.1. A fair duel.

Suppose that some observer, a marshal at the duel, fires a radar 'starting gun' pulse which travels to both duellists. The pulse triggers both weapons and the muzzle flashes are subsequently seen by the marshal at the same time. The marshal concludes that the duellists are the same distance apart and that both weapons were fired at the same instance. (The marshal might not be standing in a very clever place, but we assume he sacrifices personal safety to help simplify the discussion.) We can depict this sequence of events on a space-time diagram (see Figure 5.1). The marshal fires his radar pulse at the point O, a definite time and place. The electromagnetic pulse meets

one duellist at point A, and the other at point B. The instant at which the muzzle flashes are detected by the marshal is C. Thus points O and C lie on the marshal's worldline. The lines OA and AC represent the worldlines of photons, as do OB and BC. The marshal concludes that both weapons were fired simultaneously and so the duel was fair. Notice that we assume nothing about the motion of the duellists. They could be moving relative to each other. All we need to know is the point on their respective worldlines at which they fired their weapon. What they did before or after does not affect our determination of fairness (even if it might determine the outcome of the duel).

Now suppose (to ensure impartiality of the marshal) that each duellist has his second monitor the fight. Suppose that one of these seconds is present when the 'starting gun' is fired, but that he moves at uniform speed with respect to the marshal (see figure 5.2). We suppose that he receives one muzzle flash at point D with the other arriving later at E. Thus O, D and E lie on the worldline of the second. He must conclude that the pistol fired at A was closest, as the electromagnetic signal returns first, and therefore that duellist received the starting signal first and subsequently fired first. So he deems the duel unfair! Clearly both points of view are equally valid. We must conclude that the ordering of the events A and B depends upon your point of view. The marshal deems them to be simultaneous whereas our second deems A to precede B. If we measure distances by radar then the finiteness of the speed of light forces us to abandon absolute simultaneity.

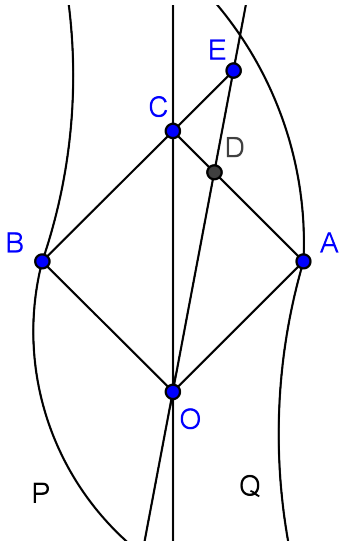

Figure 5.2. A fair duel?

We need to assume that each observer can measure time, but we are forced to conclude that this time is a personal matter. On our diagram, the ratio of the lengths of OD and OE represents the ratio of the times at which the second deems A and B to occur. However, as we will see, we cannot just take a ruler and measure on our diagram the lengths of OD and OC to measure the ratio of the times measured by the different observers. To see this, suppose that the second travels as fast as he can away from the marshal in the direction of A. The point D will become arbitrarily close to A as his velocity approaches the speed of light. So, since OA and AD are traversed by the flashes in the same period of time (as determined by the second), the time between firing the starting gun and receiving the muzzle flash will get arbitrarily small. Thus, as D gets closer to A, the length of OD represents a decreasing time. In the limit in which D approaches A this time must become zero. So a non-zero line segment represents a vanishingly small time. Then we cannot compare times measured by different observers by using a ruler!

Note that, in addition to assuming that the speed of light is constant, we are implicitly assuming some version of the Galilean assertion that *all inertial observers are created equal*.

5.2. Lorentzian Geometry. We shall see that the structure of space-time that we have just described may be modelled by a *four dimensional Lorentzian space*.

Definition 5.1. *A* ***pseudo-orthogonal*** *structure on a real-linear space V is a map $g : V \times V \to \mathbb{R}$ that is*

(1) symmetric,

(2) bilinear, and

(3) non-degenerate.

The final statement means the linear map $Y \mapsto g(X, Y)$ is the zero map only if $X = 0$ (i.e. the only vector "orthogonal" to everything is the zero vector).

A pair (V, g) is called a ***pseudo-orthogonal space****.*

You may recall that any orthogonal space admits an orthogonal basis. The usual proof of this employs the so-called Gram–Schmidt process. A similar procedure applies in the pseudo-orthogonal setting, however we need to be slightly more careful due to the presence of nontrivial **null** vectors: vectors $X \in V$ satisfying $g(X, X) = 0$.

Theorem 5.2. *Every n-dimensional pseudo-orthogonal space (V, g) admits a basis $\{E_i\}_{i=1}^n$ such that*

$$g(E_i, E_j) = \begin{cases} \pm 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \,. \end{cases}$$

*Such a basis is called a (**pseudo**)-**orthonormal** basis.*

For every pseudo-orthonormal basis for (V, g), the respective number (p, m) of plus and minus signs is the same.

Proof. We may assume that $n > 0$. In that case, there certainly exists some non-zero vector $E \in V$. In fact, we may arrange that E is not null. Indeed, if *every* $E \in V$ were null, then we would have

$$0 = g(X + Y, X + Y) = g(X, X) + 2g(X, Y) + g(Y, Y) = 2g(X, Y)$$

for all X and Y, which would violate non-degeneracy.

Set $E_1 \doteq E/g(E, E)$. If $V = \operatorname{span}\{E_1\}$, then we are done. Else, there exists $E \in V \setminus \operatorname{span}\{E_1\}$. In fact, we may arrange that E is in the kernel of the linear map $X \mapsto g(E, X)$. Indeed, if this is not the case, then we may choose $\lambda \in \mathbb{R}$ so that

$$0 = \lambda g(E_1, E_1) + g(E_1, E) = g(E_1, \lambda E_1 + E)$$

and thus replace E by $E + \lambda E_1$. Since $\ker g(E, \cdot)$ is a nontrivial subspace of V, we may further arrange, as before, that E is not null.

We now take $E_2 \doteq E/g(E, E)$. If $V \neq \operatorname{span}\{E_1, E_2\}$, then we may continue in this manner to find some E_3 satisfying $g(E_3, E_3) = \pm 1$ and $g(E_3, E_i) = 0$ for $i = 1, 2$. The process must eventually terminate since V is finite dimensional. This proves the first part of the theorem.

Now consider two pseudo-orthonormal bases $\{E_i\}_{i=1}^n$ and $\{E'_i\}_{i=1}^n$ for (V, g). We may reindex so that $g(E_i, E_i)$ is $+1$ for $i = 1, \ldots, p$ and -1 for $i = p + 1, \ldots, n$ and similarly for E'. If $p > p'$, then we can find some non-zero

$$X \in \operatorname{span}\{E_1, \ldots, E_p\} \cap \operatorname{span}\{F_{p'+1}, \ldots, F_n\} \,.$$

Now, we may write X equally well as $\sum_{i=1}^p X_i E_i$ or as $X = \sum_{i=p'+1}^n X'_i E'_i$. But then

$$0 < \sum_{i=1}^{p} (X_i)^2 = g(X, X) = -\sum_{i=p'+1}^{n} (X'_i)^2 < 0 \,,$$

which is absurd. The second claim follows. □

The pair (p, m) is called the **signature** of the pseudo-orthogonal structure g. A pseudo-orthogonal structure of signature $(n - 1, 1)$ is called a **Lorentzian** structure. A pseudo-orthogonal space equipped with a Lorentzian structure is called a **Lorentzian** space.

Example 5.3. ***Minkowski space****, $\mathbb{R}^{3,1}$, is the linear space $\mathbb{R}^4$ equipped with the Lorentzian structure η defined by*

$$\eta((X_0, X_1, X_2, X_3), (Y_0, Y_1, Y_2, Y_3)) = -X_0 Y_0 + \sum_{i=1}^{3} X_i Y_i \,.$$

It admits the "standard" pseudo-orthogonal basis $\{E_0, E_1, E_2, E_3\}$, where $E_0 \doteqdot (1,0,0,0)$, $E_1 \doteqdot (0,1,0,0)$, $E_2 \doteqdot (0,0,1,0)$, and $E_3 \doteqdot (0,0,0,1)$.

From now on, we work with Minkowski space, although all considerations hold for general Lorenzian spaces (cf. Exercise 5.1).

Definition 5.4. *A vector $X \in \mathbb{R}^{3,1}$ is called*

(1) ***timelike*** *if $\eta(X, X) < 0$,*

(2) ***spacelike*** *if $\eta(X, X) > 0$, and*

(3) ***lightlike****, or* ***null****, if $\eta(X, X) = 0$.*

Recall that a **cone** is a subset of a linear space that is closed under positive scalar multiplication.

Proposition 5.5. *The set $J \doteqdot \{X \in \mathbb{R}^{3,1} : \eta(X, X) < 0\}$ of timelike vectors in $\mathbb{R}^{3,1}$ has two connected components,*

$$J_\pm \doteqdot \{(X_0, X_1, X_2, X_3) \in J : \pm X_0 > 0\}$$

each of which is an open, convex cone. If $X \in J_+$ (resp. J_-) and $Y \in J$, then $Y \in J_+$ (resp. J_-) if and only if $\eta(X, Y) < 0$.

Proof. The set J is clearly a cone: if $\lambda > 0$ and $\eta(X, X) < 0$, then

$$\eta(\lambda X, \lambda X) = \lambda^2 \eta(X, X) < 0 \,.$$

It is open since the quadratic form $X \mapsto \frac{1}{2}\eta(X, X)$ is continuous. It is disconnected since the two subsets $J_\pm$ are separated by the hyperplane $\{X_0 = 0\}$. Each of these subsets is a convex (and hence connected) cone. Indeed, J_+ is the epigraph $\{X \in \mathbb{R}^{3,1} : X_0 > \sqrt{X_1^2 + X_2^2 + X_3^3}\}$ of the convex function $u_+(X_1, X_2, X_3) \doteqdot \sqrt{X_1^2 + X_2^2 + X_3^3}$ while J_- is the hypograph $\{X \in \mathbb{R}^{3,1} : X_0 < -\sqrt{X_1^2 + X_2^2 + X_3^3}\}$ of the concave function $u_-(X_1, X_2, X_3) \doteqdot -\sqrt{X_1^2 + X_2^2 + X_3^3}$.

Now consider any pair of timelike vectors $X, Y \in J_+$. We may write $X = X_0 E_0 + \vec{X}$ and $Y = Y_0 E_0 + \vec{Y}$ for some pair $\vec{X}$, $\vec{Y} \in \operatorname{span}\{E_1, E_2, E_3\} \cong \mathbb{R}^3$. If we denote by $\cdot$ and $|\cdot|$ the standard dot product and norm on $\mathbb{R}^3$, then

$X_0 > |\vec{X}|$ (and similarly for Y) and hence

$$\begin{aligned}\eta(X,Y) &= -X_0Y_0 + \vec{X}\cdot\vec{Y} \\ &< -|\vec{X}||\vec{Y}| + \vec{X}\cdot\vec{Y} \\ &\leq 0\end{aligned}$$

by the Cauchy–Schwarz inequality. On the other hand, if $Y \in J_-$, then $-Y \in J_+$ and hence $\eta(X,Y) = -\eta(X,-Y) > 0$. □

The set ∂J is called the **lightcone**. The set ∂J_+ is called the **future lightcone**, and ∂J_- the **past lightcone**. Timelike vectors in J_+ are referred to as **future pointing**, while those in J_- are referred to as **past pointing**.

We define the **length** of a timelike or null vector $X \in \mathbb{R}^{3,1}$ by $|X| \doteqdot \sqrt{|\eta(X,X)|}$.

Theorem 5.6 (Reverse Cauchy–Schwarz inequality)**.** *For all $X, Y \in J_+$,*

$$\frac{\eta(X,Y)}{|X||Y|} \leq -1\,.$$

Equality holds if and only if X and Y are colinear.

Proof. Since J_+ is a convex cone, $X + \lambda Y \in J_+$ for any $\lambda > 0$. Thus,

$$\begin{aligned}0 &> \eta(X+\lambda Y, X+\lambda Y) \\ &= \eta(X,X) + 2\lambda\eta(X,Y) + \lambda^2\eta(X,Y) \\ &= -|X|^2 - \lambda^2|Y|^2 + 2\lambda\eta(X,Y)\,.\end{aligned}$$

Now optimize the inequality with $\lambda = |X|/|Y|$. □

Corollary 5.7 (Reverse triangle inequality)**.** *For all $X, Y \in J_+$*

$$|X+Y| \geq |X| + |Y|\,,$$

with equality if and only if X and Y are colinear.

Proof. Since $X + Y \in J_+$, the reverse Cauchy–Schwarz inequality yields

$$\begin{aligned}|X+Y|^2 &= |X|^2 + |Y|^2 - 2\eta(X,Y) \\ &\geq |X|^2 + |Y|^2 + 2|X||Y| \\ &= (|X| + |Y|)^2\,.\end{aligned}$$

□

Now recall from high school that the **dot product** of two vectors $\mathbf{u}$, $\mathbf{v}$ in $\mathbb{R}^3$ is defined by

$$\mathbf{u}\cdot\mathbf{v} \doteqdot |\mathbf{u}||\mathbf{v}|\cos\theta\,,$$

where θ is the **angle** between $\mathbf{u}$ and $\mathbf{v}$, and $|\mathbf{u}| \doteqdot \sqrt{u_1^2 + u_2^2 + u_3^2}$ is the **length** of $\mathbf{u}$. Of course, as was often the case your adolescence, you were

lied to: rather, the angle is defined by the dot product, as allowed by the standard Cauchy–Schwarz inequality, and not the other way around. Since in the Lorentzian case the inequality is backwards, we can no longer use the cosine to define the angle between future pointing timelike vectors. Instead, a well-defined **hyperbolic angle** is determined by the **hyperbolic** cosine:

$$\cosh\theta \doteqdot \frac{|\eta(X,Y)|}{|X||Y|}\,,$$

where the sign of θ is taken to be positive if X and Y lie in the same connected component of J and negative otherwise. In Lorentzian geometry, hyperbolic trigonometry plays an analogous role to the "elliptic" trigonometry of Euclidean geometry.

5.3. Minkowskian Spacetime. Recall that a spacetime in which all inertial observers measure the same speed for light cannot support absolute simulteneity, and, therefore, cannot have a time projection Π_T. Since we are keeping the Galilean (inertial) postulate, we take "bare" spacetime to be a 4 dimensional affine space, M. Furthermore, we saw that the constancy of the speed of light imparts a natural cone structure on M, such that the length of the worldline of light is zero. We are led to equip the space of displacements $\mathbb{M}$ of M with a Lorentzian structure.

We define a **Minkowskian observer** as a curve[22] $C\colon T \to M$, where T is a one dimensional affine space, which is **future-pointing timelike**. That is, the tangent vector

$$C'(t) \doteqdot \lim_{h\to 0} \frac{C(t+h)-C(t)}{h} \in \mathbb{M} \tag{5.1}$$

of C exists and satisfies $C'(t) \in J_+$ for all $t \in \mathbb{R}$. For such a curve, the arc-length

$$L_{(a,b)}(C) \doteqdot \int_a^b |C'(t)|dt \tag{5.2}$$

of C is a well-defined geometric invariant: it does not depend on any choice of coordinates[23]. So it should also have a physical interpretation: we assign to $L_{(a,b)}(C)$ the concept of **personal time** as measured by C between the events $C(a)$ and $C(b)$. In order to "synchronise watches", we also assert that all observers are **arc-length parametrised** or **unit speed**. That is, $|C'(t)| = 1$ for all $t \in \mathbb{R}$. This then yields

$$L_{(a,b)}(C) = b - a\,.$$

[22]Note that we have sacrificed the freedom of choice of time-origin for clarity: Here, $\mathbb{R}$ should really be an interval in one-dimensional affine space.

[23]Note that the integral is well-defined even though t is an affine parameter. Roughly speaking, this is because dt is an infinitessimal *difference*.

So C measures the time $b-a$ between the events $C(b)$ and $C(a)$. (What could be simpler?)

We refer to an observer C whose velocity C' is uniform as an **inertial observer**. If we define the acceleration of C by

$$C''(t) = \lim_{h\to 0} \frac{C'(t+h) - C'(t)}{h}\,, \tag{5.3}$$

then it follows that inertial observers satisfy $C'' = 0$.

Now, at any moment $t \in \mathbb{R}$, the pseudo-orthogonal complement of $C'(t)$, $\Sigma_t \doteqdot \{X \in \mathcal{M} : \eta(C'(t), X) = 0\}$, is also a geometric invariant. We refer to Σ_t as the **instantaneous rest space of displacements**[24] of C at the moment t. The set $\sigma_t \doteqdot \{P \in M : P - C(t) \in \Sigma_t\}$ is called the **instantaneous rest space** of C at t. We interpret σ_t as the set of points which C deems simultaneous to the event $C(t)$. The motivation for this interpretation is clear if we consider an inertial observer measuring distances with radar as in the examples above.

Exercises.

Exercise 5.1. *We define the* ***standard pseudo-orthogonal*** *structure η of signature (p,m) on $\mathbb{R}^{p+m}$ by*

$$\eta(X,Y) = \sum_{i,j=1}^{n} X_i Y_j \eta_{ij}\,,$$

where X_i is the i-th component of X with respect to the standard basis $\{E_1,\ldots,E_n\}$ for $\mathbb{R}^{p+m}$ (and similarly for Y) and

$$\eta_{ij} \doteqdot \begin{cases} -1 & \text{if } i = 1\ldots,m \\ 1 & \text{if } i = m+1\ldots,n\,. \end{cases}$$

Show that every pseudo-orthogonal space (L,g) of signature (p,m) is ***isometric*** *to $\mathbb{R}^{p,m}$. That is, there exists a linear isomorphism $\phi : L \to \mathbb{R}^{p+m}$ such that*

$$\eta(\phi(X),\phi(Y)) = g(X,Y)\,.$$

The map ϕ is called an ***isometry****.*

Exercise 5.2. *Show that the acceleration C'' of an observer C lies in his instantaneous rest space. That is, $\eta(C', C'') = 0$.*

[24] I will refer to both the instantaneous rest space and the instantaneous rest space of displacements as the (instantaneous) rest space.

6. Consequences of the Lorentzian structure of spacetime

There once was a young fencer named Fisk,
whose speed was exceedingly brisk.
So fast was his action, the Lorentz–Fitzgerald contraction,
diminished his rapier to a disk.[25]

6.1. Rapidity and velocity. Consider two observers, O and P, such that $P(\tau)$ lies in the instantaneous rest space of O at time t. Denote $X = O'(t)$ and $Y = P'(\tau)$. How does O interpret Y? She splits it into components parallel and orthogonal to her own tangent vector:

$$Y = Y^{\parallel} + Y^{\perp}\,,$$

where

$$Y^{\parallel} \doteqdot \frac{\eta(X,Y)X}{\eta(X,X)} \;=\; -\eta(X,Y)X\,,$$

and

$$Y^{\perp} \doteqdot Y - Y^{\parallel}\,.$$

Since $Y^{\perp}$ lies in O's instantaneous rest space of displacements at t, $|Y^{\perp}|$ is interpreted by O as a distance. Furthermore, if we consider the inertial observer I such that $I(t) = O(t)$, and $I' = X$, then a point $I(|Y^{\parallel}|)$ is such that $I(|Y^{\parallel}|) - I(t) = Y^{\parallel}$. Since $|Y^{\parallel}| = L_{(t,t+|Y^{\parallel}|)}(I)$, $|Y^{\parallel}|$ is interpreted by O as a 'time'. Hence O views the ratio $|Y^{\perp}|/|Y^{\parallel}|$ as a **speed**. By the definition of θ, we have that

$$|Y^{\parallel}| = \cosh\theta$$
$$|Y^{\perp}| = \sinh\theta\,,$$

therefore $\tanh\theta$ is interpreted as the relative speed between the observers O and P, as observed by O at time t. The hyperbolic angle between two vectors often attains the name **rapidity**. Now, it's easy to see that P observes the same relative speed between P and O at his personal time τ. However, if we choose to write $\mathbf{v} \doteqdot \tanh\theta\mathbf{s}$, where $\mathbf{s}$ is a unit vector in the direction of $Y^{\perp}$, then $\mathbf{v}$ is P's velocity vector as observed by O, however, $-\mathbf{v}$ cannot (in general) be the velocity vector of O as observed by P, since $\mathbf{v}$ lies in the rest space of O, and not (in general) P.

6.2. Relativistic addition of speeds. In terms of the rapidity, the "relativistic addition rule for speeds" is simply hyperbolic trigonometry:

$$\tanh(\alpha+\beta) = \frac{\tanh\alpha + \tanh\beta}{1 + \tanh\alpha\tanh\beta}\,.$$

[25] An improvement, credited to George Gamow, is found by replacing two of the words.

6.3. The Doppler effect (frequency interpretation). Suppose that an inertial source emits two pulses. Let X be tangent to the worldline of the source such that $|X|$ is the time between emission of the pulses. The pulses are received by some inertial observer. Let Y be tangent to her worldline such that $|Y|$ is the time she measures between the pulses. Then $Y - X = N$ for some null vector N.

Since N is null (it is tangent to the worldline of a photon)

$$\begin{aligned} 0 = \eta(N,N) &= \eta(X,X) + \eta(Y,Y) - 2\eta(X,Y) \\ &= -|X|^2 - |Y|^2 + 2|X||Y|\cosh\theta \end{aligned}$$

where θ is the hyperbolic angle between X and Y. This is a quadratic equation for $|Y|$:

$$|Y| = |X|\cosh\theta \pm |X|\sqrt{\cosh^2\theta - 1}$$

where the first sign corresponds to the observer receding from the source. We can rearrange as follows

$$\begin{aligned} |Y| &= |X|\cosh\theta \pm |X|\sinh\theta \\ &= |X|\cosh\theta(1 \pm \tanh\theta)\,. \end{aligned}$$

We can write this in terms of the relative speed. For the observer receding from the source we have

$$|Y| = |X|\sqrt{\frac{1+v}{1-v}}\,.$$

If the source is periodically emitting pulses then the frequency of the pulse is inversely proportional to the period. So if ν is the frequency of emission and ν' the received frequency, we have

$$\nu' = \sqrt{\frac{1-v}{1+v}}\nu\,.$$

Therefore, for an observer receding from the source the frequency is lowered, that is, red-shifted.

6.4. The Doppler effect (wavelength interpretation). Suppose that an oberver $O : \mathbb{R} \to M$ 'detects' a 'photon' (null curve), $n : I \to M$ in an event $E \in M$. That is, $O(0) = n(0) = E$. Suppose O interprets this event as follows: Let $N = n'(0)$ be the tangent (null) vector to n at E, and $U = O'(0)$ the tangent vector to O at E. Then O may write $N = r(U + \mathbf{r})$, for some $r > 0$ and some unit vector $\mathbf{r}$ in O's instantaneous rest space (of displacements) at E. Then $-\mathbf{r}$ is interpreted as the direction of motion of n. We'll call r the 'colour' of n.

Suppose now that a second observer, $P : \mathbb{R} \to M$, also detects n at E, so that $P(0) = n(0) = E$. Let $V = P'(0)$, be P's tangent vector at E. Then

P writes $N = b(V + \mathbf{s})$, where $\mathbf{s}$ is a unit vector in P's instantaneous rest space (of displacements) at E. Suppose for simplicity that all of the relevant motion is planar, that is, U, V, N are coplanar vectors in $\mathbb{R}^{3,1}$. Then we may compare r and b as follows: First note that

$$r(U + \mathbf{r}) = b(V + \mathbf{s})$$

so that, taking the inner product of both sides with V,

$$\begin{aligned} r(\eta(U, V) + \eta(\mathbf{r}, V)) &= -b \\ \Rightarrow -\cosh\theta + \sinh\theta &= -\frac{b}{r}\,, \end{aligned}$$

where θ is the hyperbolic angle between U and V. Now, using the identity $\cosh^2\theta - \sinh^2\theta = 1$, we have (since $\cosh\theta \geq 1$),

$$\cosh\theta = \frac{1}{\sqrt{1 - \tanh^2\theta}}\,,$$

and

$$\sinh\theta = \frac{\pm\tanh\theta}{\sqrt{1 - \tanh^2\theta}}\,. \tag{6.1}$$

If we suppose $\eta(V, \mathbf{r}) \geq 0$, then $\sinh\theta \geq 0$ and $\tanh\theta > 0$ (since $\cosh\theta \geq 1$), and we get

$$-\cosh\theta + \sinh\theta = \frac{-1 + \tanh\theta}{\sqrt{1 - \tanh^2\theta}}\,.$$

Therefore,

$$\begin{aligned} \frac{b}{r} &= \frac{1 - \tanh\theta}{\sqrt{1 - \tanh^2\theta}}\,. \\ \Rightarrow b &= r\sqrt{\frac{(1 - \tanh\theta)^2}{(1 - \tanh\theta)(1 + \tanh\theta)}} \\ &= r\sqrt{\frac{1 - \tanh\theta}{1 + \tanh\theta}}\,, \end{aligned}$$

or, in terms of the relative speed $v = \tanh\theta$,

$$b = r\sqrt{\frac{1 - v}{1 + v}}\,.$$

That is, P measures a wavelength shorter by a factor of $\sqrt{(1 - v)/(1 + v)}$ than that measured by O.

6.5. Time dilation. Time dilation describes the descrepancy in time differences between events as measured by different observers in motion with respect to each other.

Consider two events P and Q connected by a forward-pointing timelike vector Y. (So P and Q lie on the worldline of some inertial observer.) How does another inertial observer O assign times and distances to the two events? He decomposes Y into components parallel and perpendicular to his own worldline. Then according to O, the time difference is $|Y^{\parallel}|$. But

$$\begin{aligned} |Y^{\parallel}| &= |Y| \cosh\theta \\ &= \frac{|Y|}{\sqrt{1-\tanh^2\theta}}\,. \end{aligned}$$

Of course, the proper time between these events (that is, the time measured by an inertial observer whose worldline joins them) is simply the length of Y. So the time difference between two events on an inertial observer's worldline as measured by another inertial observer is greater by a facter of $1/\sqrt{1-v^2}$ than the proper time, where $v = \tanh\theta$ is the relative speed between the two observers.

6.6. Lorentz–Fitzgerald contraction. Suppose now that some instantaneous observer $Y \in SJ_+$ (where SJ_+ is the set of unit length forward pointing timelike vectors) wishes to measure the length of a "rigid rod", with respect to which she is in motion. She deems the length of the rod to be the length of the vector K in her rest space that connects the worldlines of the ends of the rod.

So suppose that L is the vector in the rest space of the rod that joins the worldlines of its endpoints, and X is the vector tangent to the rod such that $\eta(X,X) = -|L|$.

Now $|L|Y$ is decomposed into components parallel and perpendicular to X:

$$|L|Y = \cosh\theta X + \sinh\theta L\,.$$

Then the vector $|L|\mathbf{r} \doteqdot \sinh\theta X + \cosh\theta L$ has length $|L|$ and lies in the rest space of Y ($\mathbf{r}$ is the reflection of Y about the light cone), so that K is parallel to $|L|\mathbf{r}$. That is,

$$\begin{aligned} K = \mu|L|\mathbf{r} &= \mu\left(\sinh\theta X + \cosh\theta L\right) \\ &= \mu\cosh\theta(\tanh\theta X + L)\,. \end{aligned} \tag{6.2}$$

But L is the projection of K onto the rest space of Y, so that $K = \lambda X + L$, and hence (from (6.2)) $\mu\cosh\theta = 1$. That is,

$$K = \tanh\theta X + L\,,$$

so that

$$|K| = \sqrt{|L|^2 - \tanh^2\theta |L|^2} = |L|\sqrt{1-v^2}\,,$$

where $v = \tanh\theta$ is the speed of Y relative to the rod. So Y observes the length of the rod to be shorter by a factor of $\sqrt{1-v^2}$.

The length of an object as determined by a co-moving observer is called its **proper length**.

6.7. Inertial coordinates and Lorentz transformations. Just as in Galilean spacetime, any Minkowskian observer may adapt coordinates to her worldline. Consider an observer $O : \mathbb{R} \to M$, with tangent vector $X_0 \in SJ_+$. Then O may complement X_0 with an orthonormal basis $\{X_1, X_2, X_3\}$ for her restspace $(\text{span}\{X_0\})^\perp$. Then, fixing an origin $o = O(0) \in M$, O defines coordinates

$$\begin{aligned} t(p) &= -\eta(X_0, p-o) \\ x^i(p) &= \eta(X_i, p-o) \end{aligned}$$

for each event $p \in M$.

Now suppose $\hat{O} : \mathbb{R} \to M$ is a second observer. Then $\hat{O}$ sets up his own coordinates for p:

$$\begin{aligned} \hat{t}(p) &= -\eta(\hat{X}_0, p-\hat{o}) \\ \hat{x}^i(p) &= \eta(\hat{X}_i, p-\hat{o})\,. \end{aligned}$$

Let's assume that $\hat{o} = o$. This can be achieved with a translation vector $A \doteqdot \hat{o} - o$. We also assume $\text{span}\{X_0, X_1\} = \text{span}\{\hat{X}_0, \hat{X}_1\}$, which can be achieved with an orthogonal transformation $B \in O(3)$ of $\text{span}\{\hat{X}_1, \hat{X}_2, \hat{X}_3\}$ (that is, by rotating and reflecting). Then, introducing the hyperbolic angle θ between X_0 and $\hat{X}_0$,

$$\hat{X}_0 = \cosh\theta X_0 + \sinh\theta X_1$$

and

$$\hat{X}_1 = \pm\sinh\theta X_0 + \cosh\theta X_1\,.$$

By re-choosing B (i.e by introducing a reflection if necessary), we may take the '+' sign in the second equation. We now have

$$\begin{aligned} \hat{t}(p) &= -\eta(\cosh\theta X_0 + \sinh\theta X_1, p-o) \\ &= t(p)\cosh\theta - x^1(p)\sinh\theta \\ &= \cosh\theta\left(t(p) - x^1(p)\tanh\theta\right) \\ &= \frac{t(p) - x^1(p)\tanh\theta}{\sqrt{1-\tanh^2\theta}}\,, \end{aligned}$$

and

$$\begin{aligned}\hat{x}^1(p) &= \eta(\sinh\theta X_0 + \cosh\theta X_1, p - o)\\ &= -t(p)\sinh\theta + x^1(p)\cosh\theta\\ &= \frac{x^1(p) - t(p)\tanh\theta}{\sqrt{1-\tanh^2\theta}}\,.\end{aligned}$$

Moreover (up to possibly more reflecting in $\mathrm{span}\{\hat{X}_1, \hat{X}_2, \hat{X}_3\}$), $\hat{x}^2(p) = x^2(p)$ and $\hat{x}^3(p) = x^3(p)$. We have derived the coordinate change formula for a 'Lorentz boost':

$$\begin{aligned}\hat{t} &= \frac{t - x^1 v}{\sqrt{1-v^2}}\\ \hat{x}^1 &= \frac{x^1 - tv}{\sqrt{1-v^2}}\\ \hat{x}^2 &= x^2\\ \hat{x}^3 &= x^3\,,\end{aligned}$$

where $v = \tanh\theta$. The remaining Lorentz transformations are given by the origin translation A and the orthogonal transformation B, resulting in the full *Poincaré group* of inertial coordinate transformations. The principle of relativity may now be rephrased as follows: *the laws of physics (in special relativity) should be invariant under action of the Poincaré group.*

Exercises.

Exercise 6.1. *Show that*

$$\tanh(\alpha+\beta) = \frac{\tanh\alpha + \tanh\beta}{1 + \tanh\alpha\tanh\beta}\,.$$

Interpret this as a relativistic addition rule for speeds.

Exercise 6.2. *Draw a spacetime diagram depicting the Doppler effect (both frequency and wavelength interpretations).*

Exercise 6.3. *Draw a spacetime diagram depicting the Lorentz–Fitzgerald contraction. Draw in the light cone and the 'Minkowski sphere'* $\{v \in \mathbb{R}^{3,1} : |v| = |L|\}$.

Exercise 6.4. *Resolve the "car-garage paradox"*

Exercise 6.5 (The car-garage paradox)**.** *Let's make the following adjustment to the Lorentz contraction discussion: Suppose that we replace the "rod" with the depth of a garage, and the observer O with the worldline of the front of a car, whose proper length is equal to the proper length of the garage. We wish to carry out the following experiment: We drive the car*

into the garage, at constant speed v 1, and, just as the rear of the car is inside, we slam the door shut.

Resolve the following apparent paradox: Since, from the car's perspective, the length of the garage is contracted, the car will not fit, and busts through the rear of the garage before the door shuts. On the other hand, from the garage's perspective, the car's length is contracted and it fits comfortably in the garage before the door shuts!

Exercise 6.6 (Xeno 2.0)**.** *Suppose we set up the following experiment: fix a mirror, mirror A, on one end of a straight track, and slide a second mirror, mirror B, uniformly along the track such that the two mirror faces are parallel. At some event prior to their collision, mirror A emits a photon toward B, which then bounces back and forth infinitely many times before the collision.*

(a) Draw a spacetime diagram of the situation.

(b) Calculate the distance travelled by the photon as determined by an observing physicist (i.e. co-moving with mirror A).

(c) Calculate the distance travelled by the photon as determined by a PhD student duct-taped to mirror B.

Exercise 6.7 (The twin paradox)**.** *Twin brothers Neil and Noel wave goodbye as Niel sets off for the moon. Upon reaching the moon, Niel realises he has forgotten his camera and immediately turns around to head home and get it. For simplicity, assume that Neil's motion on both legs of the trip is uniform. The "paradox" may be stated as follows:*

Since Noel moves uniformly with respect to Neil, Noel's personal time is dilated. Therefore Noel is younger upon Neil's return. On the other hand, since Neil moves uniformly with respect to Noel, Neil's personal time is dilated, and hence Neil is younger upon his return.

(a) Draw a spacetime diagram of the situation.

(b) Describe the relative lengths of the worldlines of the two observers between the start and end of Neil's trip.

(c) Deduce which twin is "really" older upon Neil's return.

Of course, the situation described is a little too simplistic, since Neil's worldline is not smooth. Physically speaking, he accelerates infinitelty at three points. The following exercise removes this issue, and makes the situation much clearer:

Exercise 6.8. *A class field trip to α-Centauri is planned. (You will be pleased to know that this is planned to be a return trip!) The rocket is to be designed to be capable of a constant acceleration g (to give a comfortable*

simulated gravity). The rocket will accelerate away from earth for a (proper) time T and then reverse the thrust to decelerate and arrive at α-Centauri after a further time T. Immediately the rocket will accelerate back towards earth for time T, at which point the thrust is again reversed so as to land the rocket after a total time of $4T$. Assuming α-Centauri to be 4.36 light years away find how long the trip will take. How much will earth-bound students not taking the trip age between the launching and the landing of the rocket?

Hints:

(i) The worldlines of the inertial and accelerating observers lie in a two-dimensional plane. Let X_0 be tangent to the inertial observer's worldline, and X_1 a unit spacelike normal, with $\{t, x\}$ the corresponding inertial coordinates. Let τ be the proper time of the accelerating observer. The (four) velocity of the rocket is given by

$$W = \frac{dt}{d\tau} X_0 + \frac{dx}{d\tau} X_1 \,.$$

Since proper time is just arc length, a proper-time-parametrised curve has a unit (four) velocity, and so

$$W = \cosh\theta X_0 + \sinh\theta X_1$$

where θ is a function of τ. The acceleration is defined by $A = \frac{dW}{d\tau}$, and we are told that $\eta(A, A) = g^2$. Hence determine θ as a function of τ, and then integrate to obtain the equation of the rocket's worldline in the inertial coordinates.

(ii) Choose units in which time is measured in years and distance in light years (so that $c = 1$). Now express g in these units.

Exercise 6.9. *Critically (and briefly) comment on the following: "Since α-Centauri is 4.36 light years away it would take light 8.72 years to make the return trip. As we cannot travel faster than light, we cannot make the trip in less than 8.72 years."*

Exercise 6.10. *Show (or argue that we have already shown) that the Poincaré group is the group of isometries of Minkowski space, that is, the group of transformations of M that preserve the (Lorentzian) length of displacement vectors.*

7. Mechanics in Minkowski space

Consider an inertial observer $C : \mathbb{R} \to M$, so that $C'(\tau) \in SJ_+$. Recall that inertial observers $C : \mathbb{R} \to M$ satisfy

$$C'' = 0\,. \tag{7.1}$$

Just as in Newtonian mechanics, we attribute any deviation from (7.1) to the existence of a '*force*' f acting on the world line of C:

$$mC''(\tau) = f(\tau)\,. \tag{7.2}$$

Just as in the Galilean setting, we'll call a pair (C, m), C an observer, $m > 0$, a (mechanical) *particle*. The constant m is called the *inertial mass* of (C, m). We'll discuss the interpretation of f in more detail in a moment.

7.1. Momentum and energy. Just as in Newtonian mechanics, we can rewrite Newton's law (7.1) as

$$p' = f\,,$$

where we have introduced the **momentum** $p = mC'$ of the particle (C, m).

Suppose the particle (C, m) is observed by an (instantaneous) inertial observer $X_0 \in SJ_+$. Then X_0 self-centredly decomposes P into components parallel and perpendicular to her world line:

$$p = EX_0 + \mathbf{p}\,.$$

It follows that

$$\begin{aligned} E &= |p| \cosh\theta \\ &= m \cosh\theta \\ &= \frac{m}{\sqrt{1 - v^2}} \\ &= m + \frac{1}{2}mv^2 + \dots\,, \end{aligned}$$

where the dots stand for terms of higher order in v. The second term in the expansion is the classical expression for the kinetic energy of the particle. So E might be interpreted by X_0 as an **energy**. If we are to accept this interpretation, then we must conclude that mass and energy are fundamentally intertwined.

Now, since $\mathbf{p} = m\mathbf{v}$, where

$$|\mathbf{v}| = \sinh\theta = \frac{v}{\sqrt{1 - v^2}}$$

is the speed accorded C by X_0, X_0 interprets $\mathbf{p}$ as a kind of momentum, which we call the **3-momentum** of (C, m) (as determined by X_0). Of course, this differs from the Newtonian expression[26] by the factor $1/\sqrt{1-v^2}$.

So energy and 3-momentum are no longer conserved quantaties: they are observer dependent. However, the momentum, p, is observer invariant, hence its length is an invariant. But this is just the mass of the particle. On the other hand, any observer can relate the mass to the observed energy and 3-momentum by Einstein's famous equation

$$\begin{aligned} -m^2 = \eta(p,p) = & -E^2 + |\mathbf{p}|^2 \\ \Rightarrow E^2 = & \, m^2 + \frac{m^2 v^2}{1-v^2}\,. \end{aligned}$$

7.2. Force and power. Now suppose X_0 introduces an inertial coordinate system for M with respect to an origin $O \in M$ and a basis $\{X_0, X_1, X_2, X_3\}$ for $\mathbb{R}^{3,1}$. That is, for any $P \in M$, we have

$$P - O = t(P)X_0 + x^i(P)X_i \qquad i = 1,2,3\,.$$

Then the tangent vector of C is

$$\begin{aligned} C'(\tau) &= \lim_{h\to 0} \frac{C(\tau+h) - C(\tau)}{h} \\ &= \lim_{h\to 0} \frac{(C(\tau+h) - O) - (C(\tau) - O)}{h} \\ &= \lim_{h\to 0} \frac{C^a(\tau+h) - C^a(\tau)}{h} X_a \qquad a = 0,\ldots,4 \\ &= \frac{dC^a}{d\tau}(\tau) X_a \qquad a = 0,\ldots,4\,, \end{aligned}$$

where we have defined $C^0 \doteqdot t \circ C$, and $C^i \doteqdot x^i \circ C$ for $i = 1,2,3$. Therefore, introducing the hyperbolic angle $\theta(\tau)$ between X_0 and $C'(\tau)$,

$$\begin{aligned} \frac{dC^0}{d\tau} &= \cosh\theta \\ \Rightarrow \frac{d}{d\tau} &= \cosh\theta \frac{d}{dt}\,. \end{aligned}$$

[26] Ian Benn quipped the following at this point: "If our observer were really perverse, he could try to make this expression look like the Newtonian one by introducing a "velocity dependent mass", $M(v) \doteqdot m/\sqrt{1-v^2}$. Whatever he does, he will not be able to escape the fact that the dynamics is non-Newtonian, so little will it profit him".

We'll use this coordinate change to see how X_0 interprets the force f. We have

$$\begin{aligned} f = \frac{d}{d\tau}p &= \cosh\theta \frac{d}{dt}(EX_0 + \mathbf{p}) \\ &= \cosh\theta \left(\frac{dE}{dt}X_0 + \frac{d\mathbf{p}}{dt}\right) \\ &= \cosh\theta \left(PX_0 + \mathbf{f}\right) , \end{aligned}$$

where we have defined $P \doteqdot \frac{dE}{dt}$ and $\mathbf{f} \doteqdot \frac{d\mathbf{p}}{dt}$. Since $\mathbf{p}$ is interpreted by X_0 as a **3-momentum**, $\mathbf{f}$ is interpreted as a **3-force**. Since E is interpreted by X_0 as an energy, P is interpreted as a **power**. The latter interpretation is also justified from the following consideration:

$$\begin{aligned} 0 = \frac{m}{2}\frac{d}{d\tau}\eta(C', C') &= \eta(C', mC'') \\ &= \eta(C', f) \\ &= \cosh^2\theta\, \eta(PX_0 + \mathbf{f}, X_0 + \mathbf{v}) \\ \implies P &= \mathbf{v}\cdot\mathbf{f} , \end{aligned}$$

which is another pre-relativity expression for the power. Note, however, the overall $\cosh\theta$ factors.

Exercises.

Exercise 7.1. *Show that the acceleration C'' of a particle is inertial observer independent. Deduce that Newton's law* (7.2) *is invariant under the Poincaré group.*

8. Electromagnetism in Minkowski space

Define a **charged particle** to be a triple (C, m, q), where C is an observer, $m > 0$ is its **mass** and $q \in \mathbb{R}$ is its **charge**.

8.1. The Lorentz force law. We shall assume that the force experienced by a charged particle in the presence of an electromagnetic field

(1) is proportional to its charge, and
(2) depends linearly on its tangent vector.

That is, we make the Ansatz

$$f(\tau) = qF(C'(\tau))\,,$$

where $F : \mathbb{R}^{3,1} \to \mathbb{R}^{3,1}$ is some linear map[27]. We can identify any such endomorphism with a bilinear form $F : \mathbb{R}^{3,1} \times \mathbb{R}^{3,1} \to \mathbb{R}$ using the metric:

$$F(X, Y) \doteqdot \eta(F(X), Y)\,.$$

Since $C'' \perp C'$, we have, by Newton's law,

$$qF(C', C') = \eta(qF(C'), C') = \eta(mC'', C') = 0\,.$$

Since the law is independent of q, we must have $F(C', C') = 0$. Since this must hold independent of the observer C, we must have $F(X, X) = 0$ for all forward-pointing timelike vectors X. This is actually sufficient to conclude the apparently stronger statement that F is skew-symmetric; that is,

$$F(X, Y) = -F(Y, X)$$

for all $X, Y \in \mathbb{R}^{3,1}$ (see Exercise 8.4).

Now, any instantaneous observer $U \in SJ_+$ interprets the field F by decomposing it into components parallel and perpendicular to his worldline. So define

$$\mathbf{E} \doteqdot F(U)\,. \tag{8.1}$$

Observe that

$$\eta(F(U), U) = F(U, U) = 0$$

and hence $\mathbf{E} \in U^\perp$. Thus, for any $\mathbf{v} \in U^\perp$,

$$\eta(F(\mathbf{v}), U) = -\eta(F(U), \mathbf{v}) = -\mathbf{E} \cdot \mathbf{v}\,.$$

We conclude that

$$F(\mathbf{v}) = (\mathbf{E} \cdot \mathbf{v})U + B(\mathbf{v})\,, \tag{8.2}$$

where B is some endomorphism of $U^\perp$ whose associated bilinear form is skew-symmetric.

[27] F is for Michael Faraday (1791-1867).

By Exercise 8.5,

$$B(\mathbf{v}) = \mathbf{v} \times \mathbf{B}$$

for some $\mathbf{B} \in U^\perp$ (uniquely determined by B), and hence

$$F(\mathbf{v}) = (\mathbf{E} \cdot \mathbf{v})U + \mathbf{v} \times \mathbf{B}\,.$$

So, decomposing $C' = \cosh\theta(U + \mathbf{v})$, we find that

$$F(C') = \cosh\theta F(U + \mathbf{v}) = \cosh\theta\big((\mathbf{E} \cdot \mathbf{v})U + (\mathbf{E} + \mathbf{v} \times \mathbf{B})\big)\,.$$

Thus, comparing with the expressions derived in §7.2, we find that

$$\begin{aligned} P &= q\mathbf{E} \cdot \mathbf{v} \\ \mathbf{f} &= q(\mathbf{E} + \mathbf{v} \times \mathbf{B})\,. \end{aligned}$$

Since the second equation is formally the Lorentz force law, we interpret $\mathbf{E}$ and $\mathbf{B}$ as the electric and magnetic fields observed by U.

8.2. Maxwell's Equations*. Up to a Lorentz transformation, we can arrange that $U = \partial_{x^0}$. Let $\{\frac{1}{2}F_{ij}\}_{i,j=1}^n$ be the components of F with respect to the basis $\{dx^i \otimes dx^j\}_{i,j=1}^n$ for the covariant two-tensors. Then

$$F_{ij} + F_{ji} = 0$$

so that

$$\begin{aligned} F &= \frac{1}{2}F_{ij}dx^i \otimes dx^j \\ &= \frac{1}{2}\sum_{i<j} F_{ij}(dx^i \otimes dx^j - dx^j \otimes dx^i) \\ &= \sum_{i<j} F_{ij}dx^i \wedge dx^j\,. \end{aligned}$$

So the exterior derivative of F is given by

$$\begin{aligned} dF &= \sum_{i<j}\sum_{k=1}^n \frac{\partial F_{ij}}{\partial x^k}dx^k \wedge dx^i \wedge dx^j \\ &= \sum_{i<j<k}\left(\frac{\partial F_{jk}}{\partial x^i} + \frac{\partial F_{ki}}{\partial x^j} + \frac{\partial F_{ij}}{\partial x^k}\right)dx^i \wedge dx^j \wedge dx^k\,. \end{aligned}$$

By (8.1),

$$\mathbf{E} = F(\partial_{x^0}) = \sum_{j=1}^3 F_{0j}\partial_{x^j}\,.$$

If we write

$$B = -B_1 dx^2 \wedge dx^3 + B_2 dx^1 \wedge dx^3 - B_3 dx^1 \wedge dx^2,$$

then, by (8.2), $B_1 = F_{23}$, $B_2 = F_{31}$ and $B_3 = F_{12}$, or, in matrix notation,

$$[F] = \begin{bmatrix} 0 & E_1 & E_2 & E_3 \\ -E_1 & 0 & B_3 & -B_2 \\ -E_2 & -B_3 & 0 & B_1 \\ -E_3 & B_2 & -B_1 & 0 \end{bmatrix}.$$

Moreover, for any $\mathbf{v} = \sum_{i=1}^{3} v_i \partial_{x^i} \in \partial_{x^0}^{\perp}$,

$$\begin{aligned} \mathbf{v} \times \mathbf{B} &= B(\mathbf{v}) \\ &= -B_1(v_2\partial_{x^3} - v_3\partial_{x^2}) + B_2(v_1\partial_{x^3} - v_3\partial_{x^1}) - B_3(v_1\partial_{x^2} - v_2\partial_{x^1}) \\ &= (v_2B_3 - B_2v_3)\partial_{x^1} - (v_1B_3 - B_1v_3)\partial_{x^2} + (v_1B_2 - B_1v_2)\partial_{x^3}\,. \end{aligned}$$

Thus,

$$\mathbf{B} = B_1\partial_{x^1} + B_2\partial_{x^2} + B_3\partial_{x^3}\,.$$

We note also that

$$B = -\iota_{\mathbf{B}}(dx^1 \wedge dx^2 \wedge dx^3) = - *_3 \mathbf{B}^{\sharp}\,,$$

where $*_3 : \Lambda(\mathbb{R}^3) \to \Lambda(\mathbb{R}^3)$ is the Hodge dual associated with the standard dot product on $\partial_{x^0}^{\perp} \cong \mathbb{R}^3$.

Since

$$\iota_{\partial_{x^0}} dF = \sum_{j<k} \left(\frac{\partial F_{jk}}{\partial x^0} + \frac{\partial F_{k0}}{\partial x^j} + \frac{\partial F_{0j}}{\partial x^k} \right) dx^j \wedge dx^k\,,$$

we obtain

$$(*_3 \iota_{dx^0} dF)^{\sharp} = \frac{\partial \mathbf{B}}{\partial x^0} + \operatorname{curl} \mathbf{E}\,,$$

where $* : \Lambda(\mathbb{R}^{3,1}) \to \Lambda(\mathbb{R}^{3,1})$ is the Hodge dual associated with the Minkowski metric on $\mathbb{R}^4$. Comparing with Maxwell's equations (4.1), we find that $\iota_{\partial_{x^0}} dF$ should vanish.

The remaining component of dF is

$$\begin{aligned} dF_{123} &= \frac{\partial F_{23}}{\partial x^1} + \frac{\partial F_{31}}{\partial x^2} + \frac{\partial F_{12}}{\partial x^3} \\ &= \frac{\partial B_1}{\partial x^1} + \frac{\partial B_2}{\partial x^2} + \frac{\partial B_3}{\partial x^3} \\ &= \operatorname{div} \mathbf{B}\,, \end{aligned}$$

which should also be zero. So the source-free Maxwell equations suggest that F should satisfy

$$dF = 0\,.$$

Now, since the Hodge duals of the basis vectors are determined by

$$\begin{aligned} *(dx^0 \wedge dx^1) &= \iota_{dx^1}\iota_{dx^0}(dx^0 \wedge dx^1 \wedge dx^2 \wedge dx^3) \\ &= \eta(dx^0, dx^0)\iota_{dx^1}(dx^1 \wedge dx^2 \wedge dx^3) \\ &= \eta(dx^0, dx^0)\eta(dx^1, dx^1)dx^2 \wedge dx^3 \\ &= -dx^2 \wedge dx^3 \end{aligned}$$

and

$$\begin{aligned} *(dx^1 \wedge dx^2) &= \iota_{dx^2}\iota_{dx^1}(dx^0 \wedge dx^1 \wedge dx^2 \wedge dx^3) \\ &= -\eta(dx^1, dx^1)\iota_{dx^2}dx^0 \wedge dx^2 \wedge dx^3) \\ &= \eta(dx^1, dx^1)\eta(dx^2, dx^2)dx^0 \wedge dx^2 \wedge dx^3 \\ &= dx^0 \wedge dx^3 \,, \end{aligned}$$

with the rest obtained by skew-symmetry and cyclic permutations of the indices $i = 1, 2, 3$, the Hodge dual of F is

$$*F = (*F)_{ij}dx^i \wedge dx^j \,,$$

where, in matrix notation,

$$[*F] = \begin{bmatrix} 0 & -B_1 & -B_2 & -B_3 \\ B_1 & 0 & E_3 & -E_2 \\ B_2 & -E_3 & 0 & E_1 \\ B_3 & E_2 & -E_1 & 0 \end{bmatrix} .$$

Since this simply interchanges $(\mathbf{E}, \mathbf{B}) \mapsto (-\mathbf{B}, \mathbf{E})$, we find that

$$(*_3\iota_{dx^0}dF)^\sharp = \frac{\partial \mathbf{E}}{\partial x^0} - \operatorname{curl} \mathbf{B} \,,$$

and

$$d * F_{123} = \operatorname{div} \mathbf{E} \,.$$

By the source Maxwell equations, we interpret $(*_3\iota_{\partial_{x^0}}dF)^\sharp$ as current density and $d*F_{123}$ as charge density. We arrive at the covariant Maxwell equations,

$$\begin{aligned} dF &= 0 \\ d * F &= J \,, \end{aligned} \tag{8.3}$$

where

$$J = \rho \, dx^1 \wedge dx^2 \wedge dx^3 - *\mathbf{j}^\flat \,.$$

8.3. Invariants*. To some extent, "one man's **E** is another man's **B**": the decomposition of F into **E** and **B** is observer dependent. Could it be then that every F can be interpreted by some observer as corresponding to a purely electric field? It turns out that this is not the case, and we can see this by finding certain invariant expressions formed from the electric and magnetic fields.

The simplest invariant is just the Minkowski product $\eta(F,F)$ of F with itself. Recall that the metric η can be defined on homogeneous tensors on $\mathbb{R}^{3,1}$ by asserting bilinearity and commutation with the tensor product, in the sense that

$$\eta(R\otimes S, T\otimes U) \doteqdot \eta(R,T)\eta(S,U)$$

for tensors R, T of the same type, and tensors S, U of the same type, and setting

$$\eta(u^\flat, v^\flat) \doteqdot \eta(u,v)$$

for any vectors $u, v \in \mathbb{R}^{3,1}$.

Using these rules, $\eta(F,F)$ is most easily calculated by introducing the pseudo-orthonormal basis $\{dx^i\}_{i=0}^3$, so that

$$\begin{aligned}
\eta(F,F) &= \eta(F_{ij}dx^i\otimes dx^j, F_{kl}dx^k\otimes dx^l)\\
&= F_{ij}F_{kl}\eta(dx^i\otimes dx^j, dx^k\otimes dx^l)\\
&= F_{ij}F_{kl}\eta^{ik}\eta^{jl}\\
&= F^{kl}F_{kl}\\
&= 2F^{0l}F_{0l} + 2\sum_{1\le k<l} F^{kl}F_{kl}\\
&= -2\mathbf{E}^2 + 2\mathbf{B}^2\\
&= 2(\mathbf{B}^2-\mathbf{E}^2)\,,
\end{aligned}$$

where

$$\mathbf{E}^2 \doteqdot \mathbf{E}\cdot\mathbf{E} \text{ and } \mathbf{B}^2 \doteqdot \mathbf{B}\cdot\mathbf{B}\,.$$

The next most obvious invariant is $\eta(*F,*F)$, but this is just $-\eta(F,F)$. Of course, we can also form the Minkowski product of F and $*F$:

$$\begin{aligned}
\eta(F,*F) &= F^{ij}*F_{ij}\\
&= 2F^{0j}*F_{0j} + 2\sum_{1\le i<j} F^{ij}*F_{ij}\\
&= -4\mathbf{E}\cdot\mathbf{B}\,.
\end{aligned}$$

Earlier we posed the question "is there always some observer who sees a purely electric field?". Certainly we now know that for this to be the case

it is necessary that $\eta(F,F)$ be negative, and that $\eta(F,*F)$ be zero. Is this sufficient? It turns out that it is, but we shall not show it.

Exercises.

Exercise 8.1. *Show that the dual space V^* of a linear space V is a linear space of the same dimension as V. Let $\{X_i\}_{i=1}^n$ be a basis for V. Show that the set $\{\alpha^i\}_{i=1}^n$ of linear transformations defined by $\alpha^i(X_j) = \delta^i_j$ forms basis for V^*.*

Exercise 8.2. *Let (V,g) be a pseudo-orthogonal space. Show that the map Φ from the linear space of endomorphisms of V (linear maps from V to itself) to the linear space of bilinear forms on V defined by*

$$\Phi(F)(X,Y) \doteqdot g(F(X),Y)$$

is an isomorphism.

Exercise 8.3. *Let V be a linear space. A **derivation at** $p \in V$ is a map $D : C(V) \to \mathbb{R}$ on smooth functions $f \in C(M)$ which*

(1) is linear,

$$D(f+\lambda g) = Df + \lambda Dg \ \text{ for all } \ \lambda \in \mathbb{R}\,,$$

and

(2) satisfies the Leibniz rule,

$$D(fg) = (Df)g(p) + f(p)(Dg)\,.$$

Note that the derivations form a linear space over $\mathbb{R}$.

Given any vector $X \in V$, we can define a derivation on functions $f \in C(M)$ by taking the directional derivative of f at p in the direction V:

$$D_X f = \sum_{i=1}^n X_i \left.\frac{\partial}{\partial x^i}\right|_p f\,.$$

Show that the map $X \mapsto D_X$ is a linear isomorphism from V to the space of derivations. Note that the basis vectors E_i map to the coordinate derivatives $\left.\frac{\partial}{\partial x^i}\right|_p$ under this identification.

Exercise 8.4. *Let F be a bilinear form on $\mathbb{R}^{3,1}$ satisfying $F(X,X) = 0$ for all $X \in SJ_+$.*

(1) Show that $F(X,X) = 0$ for all $X \in \mathbb{R}^{3,1}$.

(2) Deduce that F is skew-symmetric:

$$F(X,Y) = -F(Y,X)$$

for all $X, Y \in \mathbb{R}^{3,1}$.

Exercise 8.5. *For each vector* $\mathbf{B} \in \mathbb{R}^3$*, we may define a bilinear form* B *on* $\mathbb{R}^3$ *by taking the scalar triple product of two vectors with* $\mathbf{B}$*. That is,*

$$B(v, w) = (\mathbf{B} \times \mathbf{v}) \cdot \mathbf{w}\,.$$

Show that

(1) B *is skew-symmetric:* $B(\mathbf{v}, \mathbf{w}) = -B(\mathbf{w}, \mathbf{v})$ *for all* $\mathbf{v}, \mathbf{w} \in \mathbb{R}^3$*;*

(2) The linear space of skew-symmetric bilinear forms on $\mathbb{R}^3$ *is isomorphic to* $\mathbb{R}^3$ *via this identification. In particular, for each skew-symmetric bilinear form* B *on* $\mathbb{R}^3$*, there is a unique vector* $\mathbf{B} \in \mathbb{R}^3$ *such that* $B(\mathbf{v}, \mathbf{w}) = (\mathbf{B} \times \mathbf{v}) \cdot \mathbf{w}$ *for all* $\mathbf{v}, \mathbf{w} \in \mathbb{R}^3$*;*

(3) This is a fluke of 3-dimensions.

9. Gravity as curvature?

> *"...the essential achievement of general relativity, namely, to overcome 'rigid' space (i.e. the inertial frame), is only indirectly connected with the introduction of a Riemannian metric. The directly relevant conceptual element is the displacement field* $\Gamma_{ij}{}^{k}$*, which expresses the infinitesimal displacement of vectors. It is this which replaces the parallelism of spatially arbitrarily separated vectors fixed by the inertial frame (i.e. the equality of corresponding components) by an infinitesimal operation. This makes it possible to construct tensors by differentiation and hence to dispense with the introduction of 'rigid' space (the inertial frame). In the face of this, it seems to be of secondary importance in some sense that some particular* Γ *field can be deduced from a Riemannian metric..."* — Albert Einstein[28] (4 April 1955)

We have seen that any successful theory of gravity that includes a description of light needs to accommodate the following experimental observations:

(1) The speed of light is constant.

(2) Freefall is indistinguishable from uniform motion.

As we have seen, the first fact imparts a "lightcone structure" on spacetime, which may be modelled mathematically by a Lorenzian metric. The second fact suggests that freefallers could be described by the geodesics of some connection.

The following thought experiment (due to Einstein) suggests that the second of the above facts is incompatible with Minkowski spacetime.

Consider two freely falling observers: one located at the centre of the Earth, and an orbiting Yuri Gagarin. Since the laws of physics should look the same for both observers, both should agree on measurements of the invariants of those laws. Now, both observers certainly agree on the radius of Yuri's orbit, since there is no radial motion between them. Since spacetime is Minkowskian (and hence spacelike slices are Euclidean), according to a centre-of-the-Earth observer the length of Yuri's orbit is $2\pi r$. However, due to the effect of Lorentz contraction, the length of the orbit according to Yuri is $2\pi r\sqrt{1-v^2}$, where v is his orbital speed. This means that the flat geometry of Minkowski space implicitly discriminates between certain classes of freefallers. On the other hand, if the sectional curvature of

[28] 1879–1955. The quote is taken from F. Hehl, Y. Obukhov, *Élie Cartan's torsion in geometry and in field theory, an essay.*, Annales de la Fondation Louis de Broglie.

the 2-plane containing Yuri's orbit were positive, then the two observations could potentially be reconciled. This suggests that the equivalence principle could plausibly be reconciled with relativity if we use a non-flat Lorentzian geometry.[29]

In short, the problem is that Yuri's orbit — the worldline of a freefaller — is not a geodesic of the Minkowski connection (or any connection which is compatible with the Minkowski metric). So we are led to consider more general (non-flat) Lorentzian geometries. Indeed, remarkably, Maxwell's equations (in the form (8.3)) *require no alteration* if Minkowski spacetime is replaced by a general Lorentzian manifold!

To get an idea of how gravity might influence the structure of spacetime, consider a family of "nearby" freefallers. That is, a family $\omega_\epsilon(s) \doteqdot \omega(s,\epsilon)$ of timelike geodesics emanating from some initial point, p:

$$\begin{aligned} \omega_\epsilon(0) &= p \qquad \text{for all } \epsilon \\ \omega_\epsilon'' \doteqdot \nabla_s \partial_s \omega &= 0 \qquad \text{for all } \epsilon \\ \omega_0 &= \gamma\,. \end{aligned}$$

We assume that the variation field $J(s) \doteqdot \partial_\epsilon \omega(s,0)$ is a unit vector orthogonal to the world line of γ (i.e pointing into the instantaneous rest space of γ at each time s). Then, at least for small ϵ, we interpret $r_\epsilon \doteqdot \epsilon J$ as the position vector of the observer ω_ϵ with respect to the initial observer γ. So $r_\epsilon = \epsilon J''$ should be interpreted as the acceleration of the observer ω_ϵ with respect to γ. Now,

$$\begin{aligned} J'' \doteqdot \nabla_s \nabla_s J &= \nabla_s \nabla_s \partial_\epsilon \omega\big|_{\epsilon=0} \\ &= \nabla_s \nabla_\epsilon \partial_s \omega\big|_{\epsilon=0} \qquad (\nabla \text{ is torsion-free.}) \\ &= \nabla_\epsilon \nabla_s \partial_s \omega\big|_{\epsilon=0} - \mathrm{Rm}(\partial_s\omega, \partial_\epsilon\omega)\partial_s\omega\big|_{\epsilon=0} \\ &= -\mathrm{Rm}(\gamma', J)\gamma'\,. \qquad (\omega_\epsilon \text{ is geodesic.}) \end{aligned}$$

So we conclude that

$$r_\epsilon'' + \mathrm{Rm}(\gamma', r_\epsilon)\gamma' = 0\,.$$

The relative acceleration between two observers is interpreted as a "tidal force", caused by the presence of gravity. So let us define the **tidal force operator** $\mathrm{Rm}_\gamma : \Gamma(TM) \to \Gamma(TM)$ of γ by

$$\mathrm{Rm}_\gamma(X) \doteqdot \mathrm{Rm}(\gamma', X)\gamma'\,.$$

Returning to the Newtonian picture, we saw that

$$\mathrm{Rm}(\gamma', X)\gamma' = \mathrm{Hess}\,\phi_\gamma(X)\,.$$

[29] A four dimensional manifold equipped with a Lorenzian metric.

So the Newtonian tidal force operator (with respect to γ) is in this case $\mathrm{Rm}_\gamma \doteqdot \mathrm{Hess}\,\phi_\gamma$.

Recall that the Newtonian gravitational field is generated by a distribution of matter according to Poisson's equation:

$$\rho_\gamma = \Delta\phi_\gamma = \mathrm{tr}\,(\mathrm{Hess}\,\phi_\gamma) = \mathrm{tr}\,(\mathrm{Rm}_\gamma)\ .$$

This suggests a relativistic version of Poisson's equation:

$$\rho_\gamma = \mathrm{tr}\,(\mathrm{Rm}_\gamma) = \mathrm{Rc}(\gamma', \gamma')\,. \tag{9.1}$$

However, in contrast to the Newtonian situation, this is not sufficient to determine the connection (and hence determine the motion of freely falling particles). Besides, according to special relativity, energy and mass are but two faces of the same coin. Therefore not only mass distributions, but all forms of energy, should contribute to the gravitational field.

10. 2-tensors, 3-forms and conservation laws

> *"With the appearance of Einstein's general theory of relativity, Hilbert turned to that subject, which also occupied his colleague Felix Klein. Interestingly, the most lasting mathematical contribution out of this effort came from an algebraist who had recently engaged in studies of differential invariants. This was Emmy Noether*[30]*..., whom Hilbert and Klein brought to Göttingen to assist them in research.* — Carl B. Boyer, A History of Mathematics (1968, 1991).

For some (but not all) physical systems, concepts such as energy, momentum and mass can be made sense of. Their physical significance lies in their conservation.

10.1. 3-Forms. Consider Maxwell's equations:

$$dF = 0$$
$$d * F = * j \,,$$

where F is the electromagnetic field 2-form, and j is the **4-current** covector. We define the *total charge* of a spacelike[31] hypersurface Σ (with forward pointing timelike orientation) by

$$Q_\Sigma \doteqdot \int_\Sigma * j \,.$$

Recall **Stokes' theorem**.

Theorem 10.1 (Stokes' theorem)**.** *Let M be an n-dimensional oriented manifold with (possibly empty) boundary ∂M and let ω be a compactly supported*[32] *$(n-1)$-form on M. Then*

$$\int_M d\omega = \int_{\partial M} \omega \,.$$

Applying Stokes' Theorem to the definition of charge, we obtain

$$Q_\Sigma = \int_\Sigma * j = \int_\Sigma d * F = \int_{\partial \Sigma} * F$$

for any closed region Σ. Thus the charge in a region is the "total flux of the field through the boundary".

Now let γ be some curve which is normal to a family Σ_t of spacelike hypersurfaces (instantaneous rest spaces). Consider a region of spacetime

[30] 1882–1935.

[31] A hypersurface is a submanifold of one lower dimension than its ambient space. A hypersurface is said to be spacelike if its normal is everywhere timelike.

[32] That is, ω vanishes outside a compact subset of M.

Ω bounded by two of the spacelike hypersurfaces: Σ_2 and Σ_1. Since $d*j = dd*F = 0$ (equivalently $\operatorname{div} j^\sharp = *^{-1}d*j = 0$, where $j^\sharp$ is the vector field corresponding to the 1-form j), we have, assuming $*j$ has compact support,

$$0 = \int_\Omega d*j = \int_{\partial\Omega} *j = \int_{\Sigma_2} *j - \int_{\Sigma_1} *j = Q_2 - Q_1\,.$$

That is, the total charge of the 3-space Σ_2 is the same as the total charge of the 3-space Σ_1. This has an obvious interpretation as a conservation law.

The above discussion is not particular to electromagnetism. In fact, it is easy to see that *every* closed 3-form (divergence free vector field) leads to a "conservation law" in this way.

10.2. 2-Tensors. If our orientable manifold admits a pseudo-Riemannian metric, then Stokes' theorem implies the **divergence theorem**.

Corollary 10.2 (Divergence theorem)**.** *Let M be an n-dimensional oriented pseudo-Riemannian manifold with (possibly empty) boundary ∂M and let V be a vector field on M with compact support. If ∂M is nowhere null, then*

$$\int_M \operatorname{div} V d\mu = \int_{\partial M} g(\vec{n},\vec{n})g(\vec{n},V)d\sigma\,,$$

where μ is the Riemannian measure[33] *on M, σ the Riemannian measure induced on ∂M, and $\vec{n}$ is a choice of normal field to ∂M, normalized so that $|g(\vec{n},\vec{n})| \equiv 1$.*

Now let T be a symmetric, divergenceless covariant two tensor. We refer to such tensors as *stress-energy* tensors, since they arise in classical mechanics and special relativity in Euler–Lagrange equations of matter models, in which case their components are interpreted as "energies/momenta" and "stresses/strains". In the classical (and Minkowskian) setting, stress-energy tensors correspond to conservation laws. Roughly speaking, this is because $0 = (\operatorname{div} T)(e_i) = \operatorname{div}(T(e_i))$, so that $T(e_i)$ may be integrated as above to obtain a conserved quantity. However, in more general (i.e. non-flat) settings, this is not the case. On the other hand, if our spacetime admits isometries, then there is a procedure for generating conservation laws from stress-energy tensors (see Exercise 10.2).

The group of isometries of a pseudo-Riemannian manifold (M,g) is a Lie group. Its Lie algebra is called the **Killing algebra** of (M,g), since it consists of **Killing vector fields** (sometimes referred to as **infinitesimal isometries**). That is, vector fields K satisfying

$$\mathcal{L}_K g = 0\,.$$

[33] See section 11.1 for a definition of the Riemannian measure.

Minkowski spacetime, for example, has a Killing algebra of generated by 10 Killing vector fields: 3 generating spacelike translations, one generating a timelike translation, 3 generating rotations, and 3 generating the Lorentz boosts.

10.3. Noether's principle. Given a vector bundle with connection E over (M, g), suppose that the field $\phi \in \Gamma(E)$ is a critical point of the action

$$\mathcal{A}(\phi) \doteqdot \int_M \mathcal{L}(\cdot, \phi, \nabla\phi)\, d\mu$$

for some **Lagrangian** $\mathcal{L} : E \oplus T^*M \otimes E \to \mathbb{R}$. This means that, given any smooth, compactly supported[34] variation $s \mapsto \phi_s = \phi + sv \in \Gamma(E)$ with $\phi_0 = \phi$,

$$\left.\frac{d}{ds}\right|_{s=0} \mathcal{A}(\phi_s) = 0\,.$$

But then the divergence theorem yields

$$\begin{aligned} 0 &= \int_M \left.\frac{d}{ds}\right|_{s=0} \mathcal{L}(\cdot, \phi_s, \nabla\phi_s)\, d\mu \\ &= \int_M \left(\frac{\partial\mathcal{L}}{\partial\phi}(v) + \frac{\partial\mathcal{L}}{\partial\dot{\phi}}(\nabla v)\right) d\mu \\ &= \int_M \left(\frac{\partial\mathcal{L}}{\partial\phi} - \operatorname{div}\frac{\partial\mathcal{L}}{\partial\dot{\phi}}\right)(v)\, d\mu\,, \end{aligned} \tag{10.1}$$

where $\frac{\partial\mathcal{L}}{\partial\phi}|_{(p,\phi_p,(\nabla\phi)_p)} : E_p \to \mathbb{R}$ is the derivative of $\mathcal{L}$ with respect to the ϕ factor and $\frac{\partial\mathcal{L}}{\partial\dot{\phi}}|_{(p,\phi_p,(\nabla\phi)_p)} : (T^*M \otimes E)_p \to \mathbb{R}$ is the derivative of $\mathcal{L}$ with respect to the $\nabla\phi$ factor (after making the identifications $T_{\mathcal{L}(p,\phi_p,(\nabla\phi)_p)}\mathbb{R} \cong \mathbb{R}$ and[35] $T_{(p,\phi_p,(\nabla\phi)_p)}(E \oplus T^*M \otimes E) \cong T_pM \oplus (E \oplus T^*M \otimes E)_p$).

Now, since (10.1) holds for all $v \in E$, we conclude that the critical point ϕ satisfies the equation

$$\frac{\partial\mathcal{L}}{\partial\phi}(\cdot, \phi, \nabla\phi) = \operatorname{div}\frac{\partial\mathcal{L}}{\partial\dot{\phi}}(\cdot, \phi, \nabla\phi)\,.$$

This is called the **Euler–Lagrange** equation for the action $\mathcal{A}$.

Noether's principle asserts that "infinitesimal symmetries" of the Lagrangian give rise to conservation laws. Indeed, suppose that $\mathcal{L}$ is *infinitesimally invariant* under the action

$$\phi \mapsto \phi_s \doteqdot \alpha_s(\phi)$$

[34] I.e. ϕ_s differs from ϕ on at most a compact set (not depending on s). This ensures that the following steps make sense.

[35] Note that the latter identification is made using the connection ∇.

of some smooth family $\{\alpha_s\}_{s\in(-s_0,s_0)}$ of automorphisms of E. This just means that

$$0 = \left.\frac{d}{ds}\right|_{s=0} \mathcal{L}(\cdot, \phi_s, \nabla\phi_s)$$

for all $\phi \in \Gamma(E)$. If we set $A \doteqdot \left.\frac{d}{ds}\right|_{s=0} \alpha_s$, then the divergence theorem yields, for any $\Omega \subset M$ with nowhere-null boundary,

$$\begin{aligned} 0 &= \left.\frac{d}{ds}\right|_{s=0} \mathcal{A}|_\Omega(\phi_s) \\ &= \int_\Omega \left.\frac{d}{ds}\right|_{s=0} \mathcal{L}(\cdot, \phi_s, \nabla\phi_s)\, d\mu \\ &= \int_\Omega \left(\frac{\partial\mathcal{L}}{\partial\phi}\big(A(\phi)\big) + \frac{\partial\mathcal{L}}{\partial\dot\phi}\big(\nabla[A(\phi)]\big)\right) d\mu \\ &= \int_\Omega \left(\frac{\partial\mathcal{L}}{\partial\phi} - \operatorname{div}\frac{\partial\mathcal{L}}{\partial\dot\phi}\right)\big(A(\phi)\big)\, d\mu + \int_{\partial\Omega} g(\vec n, \vec n)\frac{\partial\mathcal{L}}{\partial\dot\phi}(\vec n, A(\phi))\, d\sigma\,, \end{aligned}$$

where $\vec n$ is a unit normal to $\partial\Omega$. Since ϕ satisfies the Euler–Lagrange equation, we conclude that

$$\int_{\partial\Omega} \frac{\partial\mathcal{L}}{\partial\dot\phi}(\vec n, A(\phi))\, d\sigma = 0\,.$$

We may once again interpret this as a conservation law (when Ω is taken to be the region bounded by two "time-slices" of M, say).

Exercises.

Exercise 10.1. *Show that a vector field is Killing if and only if its flow is a one-parameter family of isometries.*

Exercise 10.2. *Let T be a stress-energy tensor and K a Killing vector field on (M, g).*

(a) Show either that

$$d * j_K = 0\,,$$

where j_K is the 1-form defined by $j_K(X) \doteqdot T(K, X)$ for all vector fields X, or, equivalently, that

$$\operatorname{div} V_K = 0\,,$$

where V_K is the vector field corresponding to j_K; that is, $g(V_K, X) = T(K, X)$ for all vector fields X.

(b) Deduce a conservation law.

*If K generates open timelike integral curves, then the conserved integral of j_K (or V_k) over a spacelike hypersurface Σ is interpreted as the **energy** contained in Σ. If K generates open spacelike integral curves, then the conserved*

integral of j_K (or V_K) over a spacelike hypersurface Σ is interpreted as the ***momentum*** *of Σ in the direction of K. If K generates closed, spacelike integral curves, then the corresponding conserved quantity is interpreted as an* ***angular momentum*** *of Σ about the axis defined by K.*

Exercise 10.3. *Given a covector field j on Minkowski spacetime $\mathbb{R}^{3,1}$, show (directly) that the Euler–Lagrange equation for the action functional*

$$\mathcal{A}(\phi) \doteqdot \int \left(\frac{1}{2} \eta(d\phi, d\phi) + \eta(\phi, j) \right) * 1$$

(defined on one-forms ϕ) is the ***wave equation***

$$*^{-1} d * d\phi = j \,.$$

Hint: recall that

$$\alpha \wedge *\beta = \eta(\alpha, \beta) * 1 \,.$$

11. Einstein's equation

> *"It is analogous to a building, one wing of which is built from fine marble (left hand side of the equation), the other of cheap wood (right hand side of the equation)."* — Albert Einstein (Journal of the Franklin Institute, 1936)

Recall equation (9.1). On more than one occasion between 1905 and 1914, Einstein and Grossmann toyed with the following field equation for relativity:

$$\mathrm{Rc} = T\,, \tag{11.1}$$

where T is some specified stress-energy tensor functioning as the source of the gravitational field. On each occasion, Einstein rejected the equation as unphysical: the reason being that, by the second Bianchi identity, the divergence of the Ricci tensor is given by

$$\operatorname{div} \mathrm{Rc} \doteqdot \frac{1}{2} d\,\mathrm{R}\,,$$

where $\mathrm{R} \doteqdot \mathrm{tr}_g(\mathrm{Rc})$ is the scalar curvature, whereas the divergence of the stress-energy tensor should be zero. Of course, with hindsight, there is an obvious modification we can make to equation (11.1): we can make the left hand side divergenceless by subtracting a tensor whose divergence is $\frac{1}{2} d\,\mathrm{R}$. The most obvious choice is $\frac{1}{2}\,\mathrm{R}\,g$. We arrive at the famous equation posed by Einstein in 1914:

$$\mathrm{Rc} - \frac{1}{2}\,\mathrm{R}\,g = T\,. \tag{11.2}$$

In fact, David Hilbert[36] arrived at (11.2) independently, at more or less the same time (after the fierce, but productive[37], and mostly friendly, competition of November 1914) via a very different approach: the principle of least action.

11.1. The Hilbert action.

> *"I assure you that with respect to the quantum I have nothing new to say ... I am now exclusively occupied with the problem of gravitation and I hope to master all difficulties with the help of a friendly mathematician here. But one thing is certain: in all my life I have labored not nearly as hard, and I have become imbued with great respect for mathematics, the subtler part of which I had in my simple-mindedness regarded as pure luxury until now. Compared with this problem*

[36] 1862–1943.

[37] See Ivan T. Todorov, *Einstein and Hilbert: The Creation of General Relativity.*

the original relativity is a child's play." — Albert Einstein
(Letter to Arnold Sommerfeld, October 1912)

Hilbert's approach was to obtain a field equation from an action functional. The simplest choice for an action is simply the total volume of the manifold, but the Euler–Lagrange equation obtained is not very useful: $g = 0$ (this reflects the fact that volume can always be decreased by simply "scaling-down" the metric). The next most natural choice[38] would arguably be the **total scalar curvature**:

$$\mathcal{A}(g) \doteqdot \int_M \mathrm{R}\, d\mu\,.$$

Here, μ denotes the Riemannian measure induced on M by the metric and Ω is some subset of M. If (U, ϕ) is a coordinate chart, then the integral over U is defined by

$$\int_U f d\mu \doteqdot \int_{\phi(U)} f \circ \phi^{-1} \sqrt{|\det g^\phi|} d\mathcal{L}\,,$$

where g^ϕ is the matrix of the components of the metric g in the ϕ-coordinates, and $\mathcal{L}$ is the Lebesgue measure. To integrate over all of M, we choose a locally finite atlas $\{\phi_\alpha : U_\alpha \to \mathbb{R}^4\}_\alpha$ and a subordinate partition of unity $\{\rho_\alpha\}_\alpha$, and set

$$\int_M f d\mu \doteqdot \sum_\alpha \int_{\phi_\alpha(U_\alpha)} (\rho_\alpha f) \circ \phi_\alpha^{-1} \sqrt{|\det g^{\phi_\alpha}|} d\mathcal{L}\,.$$

This is well-defined by the change-of-variables formula for integration in $\mathbb{R}^4$.

We assume that the actual universe is a (local) extremiser of the Hilbert action. So consider the change in $\mathcal{A}$ produced by changing the metric $g \mapsto g_\varepsilon \doteqdot g + \varepsilon h$, where h is a compactly supported symmetric two-tensor:

$$\mathcal{A}(g_\varepsilon) \doteqdot \int_\Omega \mathrm{R}_\varepsilon\, d\mu_\varepsilon\,.$$

By assumption, we must have,

$$\begin{aligned} 0 &= \left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} \int_\Omega \mathrm{R}_\varepsilon\, d\mu_\varepsilon \\ 0 &= \left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} \int_\Omega g_\varepsilon^{ij} \mathrm{Rc}_{\varepsilon ij} d\mu_\varepsilon \\ &= \int_\Omega \left(\partial_\varepsilon g_\varepsilon^{ij} \mathrm{Rc}_{\varepsilon ij} + g_\varepsilon^{ij} \partial_\varepsilon \mathrm{Rc}_{\varepsilon ij}\right) d\mu_\varepsilon \Big|_{\varepsilon=0} + \int_\Omega \mathrm{R}_\varepsilon\, \partial_\varepsilon d\mu_\varepsilon \Big|_{\varepsilon=0}\,. \end{aligned}$$

[38] It is easy to come up with other scalars to integrate — an infinite number arise by contracting tensor products of Rm and its derivatives. It is the scalar curvature, however, that leads to general relativity.

Let's first consider the derivative of the inverse metric:

$$0 = \partial_\varepsilon \left(g_\varepsilon^{ik} g_{\varepsilon kj} \right) = \left(\partial_\varepsilon g_\varepsilon^{ik} \right) g_{\varepsilon kj} + g_\varepsilon^{ik} \partial_\varepsilon g_{\varepsilon kj} \,,$$

so that

$$\left(\partial_\varepsilon g_\varepsilon^{ik} \right) g_{\varepsilon kj} = \, - g_\varepsilon^{ik} h_{kj} \,.$$

We arrive at

$$\begin{aligned} \partial_\varepsilon g_\varepsilon^{ij} &= \, - g_\varepsilon^{ik} g_\varepsilon^{jl} h_{kl} \\ \Rightarrow \partial_\varepsilon g_\varepsilon^{ij} \big|_{\varepsilon=0} &= \, - g^{ik} g^{jl} h_{kl} \doteqdot -h^{ij} \,. \end{aligned}$$

Therefore,

$$\int_\Omega \partial_\varepsilon g_\varepsilon^{ij} \mathrm{Rc}_{\varepsilon ij} d\mu_\varepsilon \Big|_{\varepsilon=0} = \, - \int_\Omega h^{ij} \, \mathrm{Rc}_{ij} \, d\mu \,. \tag{11.3}$$

Now consider the derivative of the measure. In local coordinates,

$$\begin{aligned} \partial_\varepsilon d\mu_\varepsilon &= \partial_\varepsilon \sqrt{-\det g_\varepsilon^\phi} d\phi \\ &= \, - \frac{1}{2\sqrt{-\det g_\varepsilon^\phi}} \partial_\varepsilon \det g_\varepsilon^\phi d\phi \\ &= \frac{1}{2} \sqrt{-\det g_\varepsilon^\phi} g_\varepsilon^{ij} h_{ij} d\phi \\ &= \frac{1}{2} g_\varepsilon^{ij} h_{ij} d\mu_\varepsilon \,, \end{aligned}$$

where we used the following formula for the derivative of a determinant:

$$\partial_\varepsilon \det A(\varepsilon) = A^{ij}(\varepsilon) \det A(\varepsilon) \partial_\varepsilon A_{ij}(\varepsilon) \,.$$

We arrive at

$$\partial_\varepsilon d\mu_\varepsilon \big|_{\varepsilon=0} = \frac{1}{2} g^{ij} h_{ij} d\mu = \frac{1}{2} h^{ij} g_{ij} d\mu \,.$$

Therefore,

$$\int_\Omega \mathrm{R}_\varepsilon \, \partial_\varepsilon d\mu_\varepsilon \Big|_{\varepsilon=0} = \int_\Omega \frac{1}{2} h^{ij} \, \mathrm{R} \, g_{ij} d\mu \,. \tag{11.4}$$

The remaining integrand, $g^{ij} \partial_\varepsilon \mathrm{Rc}_{\varepsilon ij} |_{\varepsilon=0}$, turns out to be of divergence form, and therefore integrates to zero for all variations h having compact support in Ω. To see this, recall that the difference of two connections defines a tensor, and hence so too does the assignment

$$(U, V) \mapsto \frac{d}{d\varepsilon} \Big|_{\varepsilon=0} \nabla_U^\varepsilon V \,,$$

where ∇^ε is the Levi-Civita connection of g_ε. Denote this tensor by $\dot\Gamma \in \Gamma(T^*M \otimes T^*M \otimes TM)$. Note that its covariant derivative satisfies

$$\nabla_i \dot\Gamma_{jk}{}^l = \partial_i \dot\Gamma_{jk}{}^l - \dot\Gamma_{pk}{}^l \Gamma_{ij}{}^p - \dot\Gamma_{jp}{}^l \Gamma_{ik}{}^p + \dot\Gamma_{ik}{}^p \Gamma_{ip}{}^l \,.$$

(This is just the usual formula for the covariant derivative of a tensor field with respect to a coordinate chart.) Differentiating the formula for the components of Rm_ε at $\varepsilon = 0$, we then find that

$$\left(\tfrac{d}{d\varepsilon}\big|_{\varepsilon=0} \mathrm{Rm}_\varepsilon\right)_{ijk}{}^l = \nabla_k \dot\Gamma_{ij}{}^l - \nabla_j \dot\Gamma_{ik}{}^l \,.$$

Taking the trace yields

$$\left(\tfrac{d}{d\varepsilon}\big|_{\varepsilon=0} \mathrm{Rc}_\varepsilon\right)_{ij} = \nabla_j \dot\Gamma_{il}{}^l - \nabla_l \dot\Gamma_{ij}{}^l \,,$$

and hence

$$g^{ij} \left(\tfrac{d}{d\varepsilon}\big|_{\varepsilon=0} \mathrm{Rc}_\varepsilon\right)_{ij} = \operatorname{div} X \,,$$

where X is a vector field obtained by taking the difference of certain traces of $\dot\Gamma$. We conclude that

$$\int_\Omega g_\varepsilon^{ij} \partial_\varepsilon \mathrm{Rc}_{\varepsilon ij} d\mu_\varepsilon \Big|_{\varepsilon=0} = 0 \,.$$

Thus, recalling (11.3) and (11.4), we obtain

$$0 = \int_\Omega h^{ij} \left(\mathrm{Rc}_{ij} - \frac{1}{2} \mathrm{R}\, g_{ij}\right) d\mu$$

for all h with compact support in Ω. Since h and Ω were otherwise arbitrary, we must in fact have

$$\mathrm{Rc} - \frac{1}{2} \mathrm{R}\, g = 0 \,,$$

which is Einstein's vacuum equation.

We note at this point that we could just as well argue for the inclusion of some amount $\Lambda \in \mathbb{R}$ of the lower order volume term in our functional. That is, we may take

$$\mathcal{A}(\Omega) \doteqdot \int_\Omega \mathrm{R}\; d\mu - 2\Lambda \int_\Omega d\mu \,.$$

It is not hard to see that this will result in the **vacuum Einstein equation with cosmological constant** Λ:

$$\mathrm{Rc} - \frac{1}{2} \mathrm{R}\, g + \Lambda g = 0 \,.$$

If we include suitable "Lagrangian" densities $\mathcal{L}_k(g) \doteqdot \mathcal{L}(\phi_k, g)$ to our action corresponding to "matter fields" ϕ_k (sections of some vector bundle over M), then we obtain the most general form of the **Einstein equation**:

$$\mathrm{Rc} - \frac{1}{2} \mathrm{R}\, g + \Lambda g = T \,, \tag{11.5}$$

where

$$T_{ij} \doteqdot -\sum_k \left(\frac{\partial \mathcal{L}_k}{\partial g^{ij}} - \frac{1}{2}\mathcal{L}_k g_{ij} \right) . \tag{11.6}$$

Exercises.

Exercise 11.1. *Let (M, g) be a Lorentzian manifold and let Λ be a smooth function. Show that the tensor Λg is divergenceless if and only if* $\operatorname{grad} \Lambda = 0$*; that is, Λ is locally constant.*

Exercise 11.2.

(a) Extremise the action with matter Lagrangian to obtain (11.5)–(11.6)*.*

A natural example of a matter Lagrangian is given by

$$\mathcal{L}_{\text{scalar}}(\phi, g) \doteqdot \frac{1}{2}|\nabla\phi|^2 + V ,$$

where ϕ is a "scalar field" (a smooth function) and V is a "potential energy" (also just a smooth function). The term $\frac{1}{2}|\nabla\phi|^2$ is the "kinetic energy".

(b) Show that the stress-energy tensor corresponding to $\mathcal{L}_{\text{scalar}}$ is

$$T_{\text{scalar}} = -d\phi \otimes d\phi + \frac{1}{2}\left(\frac{1}{2}|\nabla\phi|^2 + V\right) g .$$

12. Schwarzschild's solution

> *"If the semi-diameter of a sphere of the same density as the Sun in the proportion of five hundred to one, and by supposing light to be attracted by the same force in proportion to its mass with other bodies, all light emitted from such a body would be made to return towards it, by its own proper gravity."* — John Michell[39], (Letter to Henry Cavendish, 1788).

> *"It is therefore possible that the greatest luminous bodies in the universe are on this account invisible."* — Pierre-Simon Laplace[40], (Exposition du Système du Monde, 1796).

Einstein's equation, although delightfully elegant, is an intimidating beast: In any local coordinate system it becomes a system of $n^2 = 16$ coupled non-linear hyperbolic PDE (although the symmetry of the Einstein and stress tensors reduces this to 10 coupled equations) to be solved for the 10 independent components of the metric. Of course, *every* Lorentzian manifold solves the Einstein equation for *some* T (namely, $T = G$); however, finding meaningful solutions involves prescribing T.

The simplest stress-energy tensor is $T = 0$. Our physical motivation for the Einstein equation suggests that solutions should model spacetime in regions absent of matter and energy (but not necessarily their influence!). So the resulting equation is called the **vacuum** (**Einstein**) **equation**:

$$\mathrm{Rc} - \frac{1}{2}\,\mathrm{R}\,g = 0\,.$$

Taking the trace (with respect to g) yields $\mathrm{R} = 0$, so that, in fact,

$$\mathrm{Rc} = 0\,. \tag{12.1}$$

That is (in dimensions $n \geq 3$), the Einstein tensor vanishes (if and) only if the Ricci tensor vanishes. At first glance, this appears to imply the trivial solution, i.e. flat Minkowski space. A moment's thought however suggests that other solutions should be possible, since, after all, gravity extends its influence rather effectively through the vast tracts of vacuum between the galaxies, where the vacuum equation should be satisfied. Luckily, there do indeed exist non-flat, Ricci-flat metrics.

In normal coordinates, the Ricci curvature is, to highest order, the coordinate Laplacian of the metric. Since the metric is Lorentzian, this means that (12.1) is, when written in normal coordinates, a coupled system of 10

[39] 1724–1793.
[40] 1749–1827.

nonlinear hyperbolic PDE. In order to solve such a system, one needs to impose appropriate boundary conditions. It is not immediately obvious how to do this, since even the underlying manifold on which the metric lives is unspecified.

We will not yet concern ourselves with the general existence theory for (12.1). Let us instead try to find some special solutions. One way to do this is to reduce the nonlinear system of PDE to a system of ODE by imposing certain symmetry assumptions. We will seek a a metric that is spherically symmetric in space and unchanging in time. Physically, this would be a sensible model for a stable, isolated gravitational system.

A naïve way to do this is to assume that there are coordinates (t, r, θ, ϕ) such that the metric is of the form

$$g = -f(r)dt \otimes dt + h(r)dr \otimes dr + r^2 g_{S^2}(\theta, \phi)$$

for some positive functions f and h of only the r-coordinate, where

$$g_{S^2} \doteqdot d\theta \otimes d\theta + \sin^2\theta \, d\phi \otimes d\phi$$

is the metric of the 2-sphere. We can then stick this into the Einstein equation and find out what f and h have to be (or if such a metric cannot solve Einstein's equation). This approach is fine if we just want to find a solution of Einstein's equation, however, by being more careful, we can actually learn a little more.

Suppose only that our metric is spherically symmetric; that is, it has a set of Killing fields generating $SO(3)$, the Killing algebra of S^2. This algebra is generated by vector fields R, S, T having the cyclic commutation relations:

$$\begin{aligned} [R, S] &= T \\ [S, T] &= R \\ [T, R] &= S \, . \end{aligned}$$

Now, S^2 has a standard chart (ϕ, θ) on which these Killing fields take the form:

$$\begin{aligned} R &= \partial_\phi \\ S &= \cos\phi\partial_\theta - \cot\theta \sin\phi\partial_\phi \\ T &= -\sin\phi\partial_\theta - \cot\theta\cos\phi\partial_\phi \, . \end{aligned}$$

So we are tempted to assume that there is a coordinate patch $(\{t, r, \theta, \phi\}, U)$ for which our metric takes the form

$$g = f(t, r)dt \otimes dt + h(t, r)dr \otimes dr + r^2 g_{S^2}(\theta, \phi) \, , \tag{12.2}$$

however, we don't need to assume this at all. Here's a rough argument for why it holds automatically: since the Killing algebra generated by R, S, and T is involutive (i.e. closes under the Lie bracket), the Frobenius Theorem

implies that it is tangent to a foliation of our manifold. From the structure of the Killing algebra, the leaves of the foliation are rigid 2-spheres (that is, the metric restricted to each leaf, a topological 2-sphere, is the standard 2-sphere metric). We can coordinate an "initial" 2-sphere using polar coordinates ϕ, θ. These coordinates can then be extended (at least locally) to the other 2-spheres using Fermi normal coordinates (a, b, ϕ, θ) based on the initial 2-sphere (the idea is to shoot off geodesics from the initial surface in the two normal directions, and use the distance along them to define "radial" coordinates a and b). In these coordinates, ∂_a and ∂_b are orthogonal to the 2-spheres, so that g takes the form

$$g = g_{aa} da \otimes da + g_{ab}(da \otimes db + db \otimes da) + g_{bb} db \otimes db + r^2 g_{S^2} \,,$$

where g_{aa}, g_{ab}, g_{bb} and r only depend on the coordinates a and b. By a straightforward coordinate change $(a, b) \to (t, r)$, we can diagonalise the first part of the metric, so that

$$g = g_{tt} dt \otimes dt + g_{rr} dr \otimes dr + r^2 g_{S^2} \,,$$

where g_{tt} and g_{rr} only depend on the coordinates t and r. So spherical symmetry really means that the metric looks like (12.2) in appropriate coordinates. We note that we have proved a general geometric statement, independent of Einstein's equation: any spherically symmetric (n-dimensional) Lorentzian metric admits local coordinates about each point for which it is of the form

$$g = f(t, r) dt \otimes dt + h(t, r) dr \otimes dr + r^2 g_{S^{n-2}} \,.$$

We make one further simplification: that $f = g_{tt}$ and $h = g_{rr}$ do not depend on the t-coordinate. This is equivalent to assuming that ∂_t is a Killing vector field. If g_{tt} is negative, then ∂_t is a time-like Killing field. A metric possessing a timelike Killing field is said to be *stationary*. If $g_{tt} < 0$, we must have $g_{rr} > 0$. So let's write $g_{tt} = -e^{-2\alpha}$ and $g_{rr} = e^{-2\beta}$ for some functions α and β of the r-coordinate, so that

$$g = -e^{-2\alpha} dt \otimes dt + e^{-2\beta} dr \otimes dr + r^2 \sin^2\theta d\phi \otimes d\phi + r^2 d\theta \otimes d\theta \,,$$

Since g is diagonal in these coordinates, it is easy to compute the dual metric:

$$g = -e^{2\alpha} \partial_t \otimes \partial_t + e^{2\beta} \partial_r \otimes \partial_r + r^{-2} \sin^{-2}\theta \partial_\phi \otimes \partial_\phi + r^{-2} \partial_\theta \otimes \partial_\theta \,.$$

We[41] are now ready to attack Einstein's equation.

Choose your own adventure exercise. *Determine the connection using one of the following methods:*

[41]Rather *you*, as I have lazily left the rest of it as an exercise. Feel free to send an email if you're stuck, or, worse, if my formulae are incorrect.

(1) *Calculate the (algebraically independent) connection coefficients of g with respect to the coordinate frame $\{\partial_i\}_{i=0}^3 = \{\partial_t, \partial_r, \partial_\theta, \partial_\phi\}$ and its dual $\{dx^i\}_{i=0}^3 = \{dt, dr, d\theta, d\phi\}$ using the formula*

$$2dx^k(\nabla_i \partial_j) \doteq 2\Gamma_{ij}{}^k = g^{kl}\left(\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij}\right) .$$

Or

(2) *Calculate the (algebraically independent) connection coefficients of g with respect to the orthonormal frame*[42] $\{e_a\}_{a=0}^3 = \{F^{-1}\partial_t, H^{-1}\partial_r, r^{-1}\partial_\theta, (r\sin\theta)^{-1}\partial_\phi\}$ *and the dual frame $\{E^a\}_{a=0}^3 \doteq \{F dt, H dr, r d\theta, r\sin\theta d\phi\}$, where $F = e^{-\alpha}$ and $H = e^{-\beta}$, using the formula*

$$2\Gamma_{ab}{}^c = g^{cd}\left(C_{abd} - C_{bda} + C_{dab}\right) ,$$

where $C_{abc} \doteq g([e_a, e_b], e_c)$ are the structure coefficients *of the frame. Or*

(3) *Calculate the (algebraically independent) connection 1-forms $\{\omega_a{}^b\}_{a,b=0}^3$, defined by*

$$\omega_a{}^b(e_c) = -(\nabla_c E^b)(e_a) ,$$

using the formula

$$g_{bc}\omega_a{}^c = \omega_{ab} = (\iota_a\iota_b dE_d)E^d + \iota_b dE_a - \iota_a dE_b ,$$

where $E_b = g_{ab}E^a$.

HINT*: By the symmetry $\Gamma_{ijk} = \Gamma_{jik}$ in case (1), and the skew-symmetry $\Gamma_{ijk} + \Gamma_{ikj} = 0$ in cases (2)-(3), there are a priori $\frac{n^2(n+1)}{2} = 40$ independent connection coefficients in case (1), $\frac{n^2(n-1)}{2} = 24$ independent connection coefficients in case (2), and $\frac{n(n-1)}{2} = 6$ independent connection one-forms in case (3). The number of coefficients reduces further due to the diagonal structure of the metric, and many of the remaining components vanish due to the simple dependence of the metric coefficients on the coordinates.*

Choose your own adventure exercise. *Calculate the Riemann tensor using one of the following methods:*

(1) *Calculate the components of the Riemann tensor with respect to the coordinate frames using the formula*

$$\mathrm{Rm}_{ijk}{}^l = \partial_j\Gamma_{ik}{}^l - \partial_i\Gamma_{jk}{}^l + \Gamma_{ik}{}^p\Gamma_{jp}{}^l - \Gamma_{jk}{}^p\Gamma_{ip}{}^l .$$

Or

[42]The orthonormal frame approach has two advantages: the metric components are $\mathrm{diag}(-1, 1, 1, 1)$, so that raising and lowering an index is just multiplication by ± 1, and there are fewer independent connection components (1-forms) to calculate, since C_{abc} (ω_{ab}) is antisymmetric in a, b, whereas $\partial_k g_{ij}$ is symmetric in i, j. But be careful with factors of -1 when raising and lowering indices.

(2) *Calculate the components of the Riemann tensor with respect to the orthonormal frames using the formula*

$$\mathrm{Rm}_{abc}{}^{d} = e_b\Gamma_{ac}{}^{d} - e_a\Gamma_{bc}{}^{d} + \Gamma_{ac}{}^{p}\Gamma_{bp}{}^{d} - \Gamma_{bc}{}^{p}\Gamma_{ap}{}^{d} - C_{ba}{}^{p}\Gamma_{pc}{}^{d}\,,$$

where $C_{ab}{}^{c} = g^{cd}C_{abd} = g^{cc}C_{abc}$. *Or*

(3) *Compute the Riemann curvature 2-forms* $\{\mathrm{Rm}_a{}^{b}\}_{a,b=0}^{3}$ *using the formula*

$$\mathrm{Rm}_a{}^{b} = d\omega_a{}^{b} + \omega_a{}^{c} \wedge \omega_c{}^{b}\,.$$

HINT*: Due to the symmetries*

$$\begin{aligned} \mathrm{Rm}_{ijkl} &= -\mathrm{Rm}_{jikl}\,, \\ \mathrm{Rm}_{ijkl} &= -\mathrm{Rm}_{ijlk}\,, \\ \mathrm{Rm}_{ijkl} &= \mathrm{Rm}_{klij}\,, \quad \textit{and} \\ \mathrm{Rm}_{ijkl} + \mathrm{Rm}_{jkil} + \mathrm{Rm}_{kijl} &= 0 \end{aligned}$$

of the Riemann tensor, there are only $\frac{n^2(n^2-1)}{12} = 20$ *algebraically independent components in cases (1)-(2) and only* $\frac{n(n-1)}{2} = 6$ *independent curvature two-forms in case (3).*

Choose your own adventure exercise. *Compute the Ricci tensor using one of the following methods:*

(1) *Contract the Riemann tensor in the coordinate frame to obtain*

$$\mathrm{Rc}_{ij} = \mathrm{Rm}_{ikj}{}^{k}\,.$$

Or

(2) *Contract the Riemann tensor in the orthonormal frame to obtain*

$$\mathrm{Rc}_{ab} = \mathrm{Rm}_{acb}{}^{c}\,.$$

Or

(3) *Contract the Riemann 2-forms in the orthonormal frame to obtain the Ricci 1-forms*

$$\mathrm{Rc}_a = \iota_b\mathrm{Rm}_a{}^{b}\,.$$

HINT*: Since* Rc *is symmetric, the number of independent components in cases (1)-(2) is* $\frac{n(n+1)}{2} = 10$.

Choose your own adventure exercise. *Set the Ricci tensor equal to zero to obtain three differential equations (plus one redundant one) for the functions* α *and* β *(or* F *and* H*). Deduce that*

$$e^{-2\alpha} = e^{2\beta} = \left(1 - \frac{M}{r}\right)\,,$$

where M is a constant that appears in solving the ODE. *If you don't get something like this, go back to step one and check your signs, otherwise...*

Congratulations, you have successfully completed your quest! Your prize: the **Schwarzschild metric**,

(12.3)
$$g = -\left(1 - \frac{M}{r}\right) dt \otimes dt + \left(1 - \frac{M}{r}\right)^{-1} dr \otimes dr + r^2 \sin^2\theta d\phi \otimes d\phi + r^2 d\theta \otimes d\theta.$$

The constant M (by comparing with the "weak-field limit") is interpreted as the mass of the source of the field. We make the following remarks:

(1) As $r \to M$, we see that the g_{rr} component of the metric blows-up. But we assumed that ∂_t is a timelike Killing field (equivalently, $g_{tt} < 0$), which is not true when $r = M$. Therefore our chart is not defined here. In fact, the "singularity" at $r = M$ may be "removed" by embedding the Schwarzschild solution inside a larger spacetime satisfying the Einstein equation, for which the hypersurface $r = M$ simply demarcates the point where the timelike Killing field ∂_t becomes spacelike (and ∂_r becomes timelike). One way to achieve this is to making use of a better choice of coordinate ansatz (Kruskal–Szekeres coordinates, say).

(2) Since ∂_r is a timelike Killing field for $r < M$, we have really discovered two different regions (distinguished by $r > M$ and $0 < r < M$) for which (12.3) solves Einstein's equation. (In Kruskal–Szekeres coordinates, both of these regions lie in a single chart.)

(3) In the inner chart, we can investigate the limit $r \to 0$. We find that there is a ridgy-didge curvature singularity at $r = 0$: the normed curvature
$$|\mathrm{Rm}|^2 \doteqdot g(\mathrm{Rm}, \mathrm{Rm}) = \mathrm{Rm}_{ijkl}\,\mathrm{Rm}^{ijkl} = \frac{12M^2}{r^6}$$
blows-up as $r \to 0$. Note, however, that there's no good way of figuring out when coordinate singularities are dinky-di singularities: for example, the scalar curvature of the Schwarzschild metric is identically zero, and we need to look at the more complicated invariant $|\mathrm{Rm}|$ to see that something nasty occurs. More generally, $|\mathrm{Rm}|$ could remain bounded at some point while some other geometric quantity (e.g $|\nabla \mathrm{Rm}|$, or $|\nabla^2 \mathrm{Rm}|$,...) blows-up there. Each of these singularities may have a different physical significance (or none at all).

(4) Since the trace part (the Ricci curvature) of the curvature tensor of the Schwarzschild metric is bounded (it is zero), it must be the

trace-free part which blows-up. The trace-free part, $\mathring{\mathrm{Rm}}$, of Rm is called the WEYL TENSOR. It is given explicitly by

$$\mathring{\mathrm{Rm}} \doteqdot \mathrm{Rm} - \frac{1}{2}\left(\mathrm{Rc} - \frac{\mathrm{R}}{4}g\right) \owedge g - \frac{\mathrm{R}}{24} g \owedge g \,,$$

where $\owedge$ is the **Kulkarni–Nomizu product**, defined on a pair of covariant 2-tensors S and T by

$$\begin{aligned}(S \owedge T)(X,Y,Z,W) \doteqdot\ & S(X,Z)T(Y,W) - S(X,W)T(Y,Z) \\ & + S(Y,W)T(X,Z) - S(Y,Z)T(X,W)\,.\end{aligned}$$

The Weyl tensor does not convey information on how the volume of the body changes (this is encoded in the Ricci tensor), but rather only how the shape of the body is distorted by the tidal force. It turns out that the Weyl tensor is invariant under conformal changes of the metric, $g \mapsto \mathrm{e}^{2u}g$, $u \in C(M)$. So the Schwarzschild singularity is also a singularity of the conformal structure of spacetime.

Exercises.

Exercise 12.1. *Obtain the Schwarzschild metric by completing the* choose your own adventure *exercises in this section.*

Exercise 12.2. *Show that the "stationary" assumption was not necessary. Deduce that any spherically symmetric vacuum solution of Einstein's equation is automatically Schwarzschild, and hence stationary. In particular,* a spherically symmetric, uncharged mass cannot emit gravitational waves. *This result is known as* Birkhoff's Theorem*, which is the simplest of the democratic "no hair" theorems for singularities.*

13. Geodesy of the Schwarzschild solution

> *"It is always pleasant to have exact solutions in simple form at your disposal."* — Karl Schwarzschild[43], 1916.

13.1. Constants of the motion. Recall that a **Killing field** is a vector field K whose integral curves generate isometries; that is,

$$g_{\phi(p,s)}\big(\phi(\cdot,s)_* u, \phi(\cdot,s)_* v\big) = g_p(u,v)$$

for all $u, v \in T_pM$, $p \in M$, and $s \in (-s_0, s_0)$, where $\phi : M \times [0, s_0) \to M$ is the **flow** of K:

$$\begin{cases} \dfrac{d\phi}{ds}(p,s) = K(\phi(p,s))\,, \\ \phi(p,0) = p\,. \end{cases}$$

It follows that

$$\mathcal{L}_K g = 0\,, \tag{13.1}$$

where $\mathcal{L}$ denotes the Lie derivative. Now, the Lie derivative "commutes with contractions", so that

$$0 = \mathcal{L}_K g(X,Y) = K\left[g(X,Y)\right] - g(\mathcal{L}_K X, Y) + g(X, \mathcal{L}_K Y)\,.$$

Since the Lie derivative of a vector field Y with respect to X is the commutator $[X,Y]$, we obtain

$$0 = \mathcal{L}_K g(X,Y) = K\left[g(X,Y)\right] - g([K,X],Y) + g(X,[K,Y])\,.$$

Using the metric compatibility and symmetry of ∇, we conclude that

$$\begin{aligned} 0 &= \mathcal{L}_K g(X,Y) \\ &= g(\nabla_K X, Y) + g(X, \nabla_K Y) - g([K,X],Y) + g(X,[K,Y]) \\ &= g(\nabla_X K, Y) + g(X, \nabla_Y K)\,. \end{aligned} \tag{13.2}$$

In other words, the bilinear form related to ∇K by the metric is skew-symmetric. Equation (13.2) is called **Killing's equation**. Killing's equation is equivalent to (13.1); solving it is one way to find Killing fields.

Now suppose that γ is a geodesic of g (with parameter s). Then

$$\begin{aligned} \partial_s g(\gamma', K) &= g(\nabla_s \gamma', K) + g(\gamma', \nabla_s K) \\ &= g(\gamma', \nabla_s K) && \text{(Since } \gamma \text{ is geodesic)} \\ &= 0 && \text{(By Killing's equation (13.2))}\,. \end{aligned} \tag{13.3}$$

That is, if K is a Killing field, and γ is a geodesic, then the scalar $g(\gamma', K)$ is constant along γ. We will call $g(\gamma', K)$ a **constant of the motion of** γ.

[43] 1873–1916.

Recall that (outside the Schwarzschild radius) the Schwarzschild metric

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

has a Killing algebra generated by four Killing fields, one generating time translations, $K \doteqdot \partial_t$, and three generating rotations,

$$\begin{aligned} R &= \partial_\phi \\ S &= \cos\phi\ \partial_\theta - \cot\theta \sin\phi\ \partial_\phi \\ T &= \sin\phi\ \partial_\theta - \cot\theta \cos\phi\ \partial_\phi\,. \end{aligned}$$

Let γ be a timelike, future pointing, unit speed geodesic of the Schwarzschild metric. We interpret γ as a "test particle" of negligible mass, m. In the (t, r, θ, ϕ) coordinates, we can write

$$\gamma' = t'\partial_t + r'\partial_r + \theta'\partial_\theta + \phi'\partial_\phi\,,$$

where $\cdot'$ denotes differentiation with respect to the proper time, s, and we conflate the coordinates t, r, θ, ϕ with their values along γ, $t\circ\gamma, r\circ\gamma, \theta\circ\gamma, \phi\circ\gamma$.

By rotating the θ, ϕ coordinates, we may assume that γ satisfies $\theta \equiv \pi/2$. Then

$$\gamma' = t'\partial_t + r'\partial_r + \phi'\partial_\phi\,.$$

Now, since K is Killing, the conservation law (13.3) implies that

$$E \doteqdot g(m\gamma', K) = -mt'\left(1 - \frac{M}{r}\right)$$

is a constant of the motion of γ. Comparing with the physics of special relativity (which is particularly justified where r is large compared to M), we interpret E as the energy of γ, as determined by a comoving observer.

Similarly,

$$L \doteqdot g(m\gamma', R) = m\phi' r^2$$

is a constant of the motion. We interpret L as the angular momentum of γ about the "axis" defined by the integral curves of R, as determined by a comoving observer.

Now consider

$$\begin{aligned} -m^2 = g(m\gamma', m\gamma') &= mg(m\gamma', t'K + r'\partial_r + \phi' R) \\ &= mt'E + m^2(r')^2\left(1 - \frac{M}{r}\right)^{-1} + m\phi' L \\ &= -E^2\left(1 - \frac{M}{r}\right)^{-1} + m^2(r')^2\left(1 - \frac{M}{r}\right)^{-1} + \frac{L^2}{r^2}\,. \end{aligned}$$

Rearranging, we obtain

$$E^2 = m^2 + m^2 \left(r'\right)^2 + \frac{L^2}{r^2} - m \cdot \frac{mM}{r} - \frac{ML^2}{r^3} \,.$$

So we interpret each of the terms on the right as a form of energy. The first term we are already familiar with as a form of energy from special relativity, and the second and third terms correspond with classical kinetic energies. The term $-\frac{mM}{r}$ corresponds to the Newtonian gravitational potential, which justifies the identification of the constant M with the mass of the system. The final term $\frac{ML^2}{r^3}$ is not familiar from Newtonian physics or special relativity. It turns out that it is precisely what was required to explain the "perihelion anomoly" of the planet Mercury (see [**3**]).

13.2. Radial geodesics. Consider a test particle with unit mass and no angular momentum. From above, the energy of the particle must satisfy

$$E^2 = \left(\frac{dr}{ds}\right)^2 + \left(1 - \frac{M}{r}\right)$$

$$\Rightarrow \left(\frac{dr}{ds}\right)^2 = E^2 - \left(1 - \frac{M}{r}\right) \,. \tag{13.4}$$

Let us synchronize clocks so that $r_0 \doteq r(0)$ satisfies

$$0 = E^2 - \left(1 - \frac{M}{r_0}\right)$$

$$\Rightarrow r_0 = \frac{M}{1 - E^2} \,. \tag{13.5}$$

Then $r'|_{r=r_0} = 0$ (so we are dropping the test particle from rest at time zero). If $E^2 < 1$, then $r_0 > M$, so that r_0 is part of the exterior Schwarzschild coordinate chart. Combining (13.4) and (13.5), and separating variables, we obtain

$$\sqrt{\frac{r}{r_0 - r}} dr = -\sqrt{\frac{M}{r_0}} ds \,.$$

This can be solved by reparametrising r so that

$$r = \frac{r_0}{2}(1 + \cos\xi).$$

The solution is then given by

$$2r = r_0(1 + \cos\xi)$$

$$2s = r_0\sqrt{\frac{r_0}{M}}(\xi + \sin\xi) \tag{13.6}$$

for $\xi \in [0, \xi_M)$, where ξ_M is the point at which the geodesic reaches the edge of the exterior chart. On the other hand, since the metric takes the same form on the interior chart, for $\xi \in (\xi_M, \pi)$ the equations (13.6) define a

radial geodesic that lies in the interior. Notice that r and s remain smooth with respect to ξ as $\xi \to \xi_M$. By including this smooth limit point in our manifold and identifying the outer and inner geodesics at that point, we can smoothly "glue" the interior and exterior solutions together to obtain a single, connected solution described (modulo the horizon $r = M$) by our two coordinate charts[44].

On the other hand, we saw that the norm $|\mathrm{Rm}|$ of the curvature tensor blows up as $r \to 0$, which implies that the metric is not smooth as $r \to 0$. That is, our geodesic cannot be extended (smoothly) through the curvature singularity $r = 0$. Since $s \to \sqrt{r_0^3\pi^2/M}$ as $r \to 0$, *the singularity is reached in finite proper time*. Something terrible must happen to our observer at proper time $s = \sqrt{r_0^3\pi^2/M}$!

We have thrown the word "singularity" around a lot here, all the while remaining very vague about its meaning. It was used to refer to the bad behaviour of the metric, in Schwarzschild coordinates, as $r \to M$, and to the behaviour of the curvature, $|\mathrm{Rm}| \to \infty$, as $r \to 0$. There are many other things that could potentially go wrong, for example other invariants such as derivatives of the curvature could blow-up, or the injectivity radius could approach zero. The situation above, in which an observer ceases to exist, seems worthy also of the monacle "singularity", even if it is perhaps not coupled with something like the blow-up of curvature. This latter kind of singularity certainly seems physically more objectionable than the other kinds.

Exercises.

Exercise 13.1. *Show that the angular momenta of a radial Schwarzschild geodesic γ about the axes determined by the Killing fields S and T are zero.*

[44]The result can also be described in a single coordinate chart, known as *Kruszkal–Szekeres coordinates* (see [**3**]).

14. The Friedmann universe

> *Among the authorities it is generally agreed that the Earth is at rest in the middle of the universe, and they regard it as inconceivable and even ridiculous to hold the opposite opinion. However, if we consider it more closely the question will be seen to be still unsettled, and so decidedly not to be despised. For every apparent change in respect of position is due to motion of the object observed, or of the observer, or indeed to an unequal change of both.* — Nicolaus Copernicus[45] (On the Revolutions of the Heavenly Spheres).

There is experimental evidence (namely, the temperature distribution of the cosmic microwave background radiation and the present distribution of galaxies) to suggest that the universe is *spatially isotropic* at large scales, at least from the point of view of an Earth-bound astronomer. That is, at each moment on our astronomer's watch, the universe looks, to him, approximately the same, at a large enough scale, in all directions. The Copernican principle suggests that this should also be the case for astronomers on other planets in other epochs.[46]

Definition 14.1. *A Riemannian manifold (Σ, γ) is called **isotropic at** $p \in \Sigma$ if, given any $u, v \in S_p\Sigma$ (the unit tangent space at p), there is an isometry ϕ of Σ that fixes p and rotates the direction u to the direction v. That is,*

$$\phi(p) = p \ \text{ and } \ \phi_* u = v \,.$$

*A Riemannian manifold (Σ, γ) is called **isotropic** (or **maximally symmetric**) if it is isotropic at every $p \in \Sigma$.*

Exercise 14.1. *A Riemannian manifold (Σ, γ) is called **homogeneous** if given $p, q \in \Sigma$ there is an isometry ϕ of Σ such that $\phi(p) = q$. Show that every complete isotropic Riemannian manifold is homogeneous.*

We will seek a solution (M, g) to Einstein's equation admitting a smooth function $t : M \to \mathbb{R}$ with nonvanishing gradient such that

$$g = -dt \otimes dt + \alpha^2 \gamma \,,$$

where α is constant on each level hypersurface $\Sigma_{t_0} \doteq \{x \in M : t(x) = t_0\}$, $\mathcal{L}_{\partial_t} \gamma = 0$, and, denoting by $\iota_{\Sigma_t} : \Sigma_t \hookrightarrow M$ the inclusion map, $\iota_{\Sigma_t}^* \gamma$ is an isotropic Riemmanian metric on Σ_t for each t.

[45] 1473–1543

[46] The model we shall develop, based on these assumptions, is also known as the *Friedmann* or *Friedmann–Robertson–Walker* or *Robertson–Walker* or *Friedmann–Lemaître* model. We have given Friedmann all the credit since he was the earliest to develop the model and since *Friedmann model* it is shorter than *Friedmann–Lemaître–Robertson–Walker model*.

Proposition 14.2. *If (Σ, γ) is isotropic, then γ is **Einstein**:*

$$\mathrm{Rc} = f\gamma$$

for some $f \in C^\infty(\Sigma)$.

Proof. Recall that the **sectional curvature** of a two-plane $\Pi \subset T_p\Sigma$ is defined by

$$K(\Pi) \doteqdot \mathrm{Rm}(e_1, e_2, e_1, e_2)$$

for some (and, in fact, any) orthonormal basis $\{e_1, e_2\}$ for Π.

Thus, for any $u \in T_p\Sigma$,

$$\mathrm{Rc}(u, u) = \sum_{a=1}^{n} \mathrm{Rm}(e_a, u, e_a, u)\,,$$

where $\{e_a\}_1^n$ is any orthonormal basis for $T_p\Sigma$. Using the Gramm–Schmidt procedure, we can assume that $e_1 = u/|u|$, so that

$$\mathrm{Rc}(u, u) = \sum_{a=2}^{n} |u|^2\, \mathrm{Rm}(e_1, e_a, e_1, e_a) = |u|^2 \sum_{a=2}^{n} K(\Pi_a)\,,$$

where $\Pi_a \doteqdot \mathrm{span}\{e_1, e_a\}$.

Now, since the curvature tensor is invariant under isometries, the isotropy of γ implies that $K(\Pi_a) = K(\Pi_b) = k_p$ for all $a, b = 2, \ldots, n$. It follows that

$$\mathrm{Rc}(u, u) = (n-1)k_p|u|^2 = (n-1)k_p\gamma(u, u)\,.$$

By polarisation, we conclude that

$$\mathrm{Rc}(u, v) = (n-1)k_p\gamma(u, v)$$

for all $u, v \in T_p\Sigma$. Since p was arbitrary, the claim follows. □

Theorem 14.3 (Schur's Theorem)**.** *If γ is a Riemannian Einstein metric on Σ^n, and $n \geq 3$, then γ has locally constant scalar curvature, and*

$$\mathrm{Rc} = \frac{\mathrm{R}}{n}\gamma\,.$$

Proof. If γ is Einstein, $\mathrm{Rc} = f\gamma$, then, taking the trace, $\mathrm{R} = nf$, so that $\mathrm{Rc} = \frac{\mathrm{R}}{n}\gamma$. It follows that

$$\mathrm{div}\,\mathrm{Rc} = \frac{1}{n} d\,\mathrm{R}\ .$$

However, we have seen that the second Bianchi identity yields

$$\mathrm{div}\,\mathrm{Rc} = \frac{1}{2} d\,\mathrm{R}\,.$$

It follows that $(n-2)d\,\mathrm{R} = 0$. □

Corollary 14.4. *If (Σ^n, γ), $n \geq 3$, is isotropic, then γ has constant sectional curvature k, and* $\mathrm{Rc} = (n-1)k\gamma$.

Now (the argument is similar to the one for the Schwarzschild metric) isotropy of γ ensures that there exist local coordinates $\{t, r, \theta, \phi\}$ about each $p \in \Sigma$ such that

$$\gamma = e^{2\beta} dr \otimes dr + r^2 \left(d\theta \otimes d\theta + \sin^2 \theta d\phi \otimes d\phi \right),$$

where β depends only on r (that is, $\partial_\theta \beta = \partial_\phi \beta = \partial_t \beta = 0$). Therefore, the Lorentzian metric g on M locally takes the form

$$g = -dt \otimes dt + \alpha^2 \left[e^{2\beta} dr \otimes dr + r^2 \left(d\theta \otimes d\theta + \sin^2 \theta d\phi \otimes d\phi \right) \right],$$

The non-vanishing components of the Ricci tensor of γ are given (in $\{r, \theta, \phi\}$ coordinates) by

$$\begin{aligned} \mathrm{Rc}_{rr} &= \frac{2}{r} \beta_r \\ \mathrm{Rc}_{\theta\theta} &= e^{-2\beta}(r\beta_r - 1) + 1 \\ \mathrm{Rc}_{\phi\phi} &= \sin^2 \theta \ \mathrm{Rc}_{\theta\theta}, \end{aligned}$$

where $\beta_r \doteqdot \partial_r \beta$ (see Exercise 14.2).

Equating $\mathrm{Rc} = 2k\gamma$, we obtain the ODE

$$\begin{aligned} \frac{1}{r} \beta_r &= k e^{2\beta} \\ e^{-2\beta}(r\beta_r - 1) + 1 &= 2kr^2. \end{aligned}$$

Eliminating β_r, we obtain

$$e^{-2\beta} = 1 - kr^2,$$

and hence

$$\gamma = \frac{1}{1 - kr^2} dr \otimes dr + r^2 \left(d\theta \otimes d\theta + \sin^2 \theta d\phi \otimes d\phi \right).$$

When $k = 0$, this simply the Euclidean metric on $\mathbb{R}^3$ in polar coordinates.

We conclude the following:

(1) If $k > 0$, then (Σ, γ) is locally isometric to $S^n_{1/\sqrt{k}}$, the n-sphere of radius $1/\sqrt{k}$, and in particular has finite diameter. If (Σ, γ) is

geodesically complete[47], then it is necessarily compact[48], and its universal cover is $S^n_{1/\sqrt{k}}$.

(2) If $k = 0$, then (Σ, γ) is locally isometric to $\mathbb{R}^n$. However (Σ, γ) could be either unbounded (e.g. $\mathbb{R}^n$) or bounded (e.g the cubic torus $T^n \doteqdot S^1 \times \cdots \times S^1$). If (Σ, γ) is geodesically complete, then its universal cover is $\mathbb{R}^n$.

(3) If $k < 0$, then (Σ, γ) is locally isometric to $H^n_{1/\sqrt{-k}}$, n-dimensional hyperbolic space of curvature k, and could be either unbounded (e.g. $H^n_{1/\sqrt{-k}}$) or bounded (e.g. quotients of H^n by lattices). If (Σ, γ) is geodesically complete, then its universal cover is $H^n_{1/\sqrt{-k}}$.

We next consider the components of the Ricci tensor of g. The non-vanishing components are

$$\begin{aligned}
\mathrm{Rc}_{tt} &= -3\frac{\ddot{\alpha}}{\alpha}\,, \\
\mathrm{Rc}_{rr} &= g_{rr}\left(\frac{\ddot{\alpha}}{\alpha} + 2\left(\frac{\dot{\alpha}}{\alpha}\right)^2 + \frac{2k}{\alpha^2}\right), \\
\mathrm{Rc}_{\theta\theta} &= g_{\theta\theta}\left(\frac{\ddot{\alpha}}{\alpha} + 2\left(\frac{\dot{\alpha}}{\alpha}\right)^2 + \frac{2k}{\alpha^2}\right), \\
\mathrm{Rc}_{\phi\phi} &= g_{\phi\phi}\left(\frac{\ddot{\alpha}}{\alpha} + 2\left(\frac{\dot{\alpha}}{\alpha}\right)^2 + \frac{2k}{\alpha^2}\right),
\end{aligned}$$

where $\dot{\alpha} \doteqdot \partial_t \alpha$ and $\ddot{\alpha} \doteqdot \partial_t^2 \alpha$. It follows that

$$\mathrm{R} = 6\left(\frac{\ddot{\alpha}}{\alpha} + \left(\frac{\dot{\alpha}}{\alpha}\right)^2 + \frac{k}{\alpha^2}\right).$$

(See Exercise 14.4.)

Before solving the Einstein equation, we need to specify a stress tensor. One possibility is an empty universe, $T = 0$. In this scenerio, we have

$$\begin{aligned}
\ddot{\alpha} &= 0 \\
\dot{\alpha}^2 + k &= 0\,.
\end{aligned}$$

[47]Geodesic completeness refers to the infinite (smooth) extendability of all geodesics. By the Hopf–Rinow theorem, geodesic completeness of a Riemannian manifold (i.e. positive definite metric) is equivalent to completeness of the manifold as a metric space, with distance defined by the length of shortest joining curves. We will discuss geodesic completeness further in the next lecture.

[48]If (Σ, γ) satisfies $\mathrm{Rc} \geq (n-1)k\gamma$ for some $k > 0$, then the diameter of Σ is bounded above by $\pi/\sqrt{k}$. This is known as Meyers' Theorem.

In particular, $k \leq 0$. Solving the second equation, we get $\alpha(t) = \pm t\sqrt{-k} + \alpha_0$. Translating the time coordinate, we can set $\alpha_0 = 0$. It appears that the universe either expands uniformly out of a single point at $t = 0$, existing for all $t > 0$, or exists for all $t < 0$, collapsing into a single point at $t = 0$. However, it can be checked that the curvature tensor vanishes. It follows that (M, g) is locally isometric to Minkowski spacetime, so that our chart is just a subset of Minkowksi spacetime with some funny coordinates.

By Einstein's equation, the stress-energy tensor must be of the form

$$T = (\rho + p)dt \otimes dt + pg\,, \tag{14.1}$$

where ρ and p depend only on t (i.e. they are constant on the level-sets Σ_t). Given $x \in M$, $\rho(x)$ is interpreted as the energy density at x as seen by an observer at x with tangent vector ∂_t, and $p(x)$ is interpreted as the pressure (which is the same in all directions) at x as felt by an observer at x with tangent vector ∂_t. Since ρ and p depend only on t, the pressure and energy-density of this universe are uniform across each space slice Σ_t.

Einstein's equation yields the ODE

$$\frac{\rho}{3} = \left(\frac{\dot{\alpha}}{\alpha}\right)^2 + \frac{k}{\alpha^2} \tag{14.2}$$

$$p = -\left(2\frac{\ddot{\alpha}}{\alpha} + \left(\frac{\dot{\alpha}}{\alpha}\right)^2 + \frac{k}{\alpha^2}\right). \tag{14.3}$$

These equations are known as the **Friedmann equations**. Adding the two yields

$$\frac{\ddot{\alpha}}{\alpha} = -\frac{1}{6}(3p + \rho)\,,$$

which we can write as the pair of first order equations

$$H = \frac{\dot{\alpha}}{\alpha} \tag{14.4}$$

$$\dot{H} + H^2 = -\frac{1}{6}(3p + \rho)\,. \tag{14.5}$$

We gain an equation from the requirement that T be divergence free:

$$\dot{\rho} = -3H(p + \rho)\,,$$

but this can also be obtained from differentiating (14.2).

We would like to solve (14.4)–(14.5), however this will not be possible without specifying $3p + \rho$. On the other hand, if $3p + \rho$ has a sign, we can determine a very important property of (14.5):

Proposition 14.5. *Let $H : (t_0, 0] \to \mathbb{R}$ be a solution to (14.5) with $H_0 \doteqdot H(0) > 0$ which does not extend farther backwards in time. If $3p + \rho \geq 0$, then $t_0 > -\infty$ and $H \nearrow \infty$ as $t \searrow t_0$.*

Proof. This is a simple comparison property of the ODE $\dot{\phi}+\phi^2 = 0$. We will prove the statement in case $3p + \rho > 0$ (a simple perturbation argument[49] shows that it also holds for $3p + \rho \geq 0$). Let u be the solution to

$$\begin{cases} \dot{u} + u^2 = 0 \\ u(0) = H_0 \,. \end{cases}$$

That is,

$$u(t) = \frac{H_0}{1 + H_0 t} \,.$$

We will prove that $H(t) > u(t)$ for all $t \in (-H_0^{-1}, 0) \cap (t_0, 0)$.

Set $w \doteqdot \log(H/u)$. Then w satisfies the equation

$$\dot{w} = -(H - u) - \frac{3p + \rho}{6H} \,. \tag{14.6}$$

In particular, $\dot{w}(0) < 0$. It follows that $w(t) > 0$ for t less than but close to 0.

Suppose, contrary to the claim, that there is some $t_1 < 0$ at which $w(t_1) = 0$. From above, we may assume without loss of generality that $w(t) > 0$ for all $t \in (t_1, 0)$. Equivalently, $H > u$ for $t \in (t_1, 0)$. By the mean value theorem, there must be some $t_2 \in (t_1, 0)$ such that $\dot{w}(t_2) = 0$. But this contradicts (14.6). Therefore w (and hence $H - u$) must be positive for all $t \in (t_0, 0)$ such that $H(t)$ is defined. The claim now follows. □

By (14.2) and (14.3), the energy density and pressure both become infinite as $t \searrow t_0$.

The assumption that $3p + \rho$ is non-negative is a consequence of the **Strong Energy Condition** (SEC):

$$\mathrm{Rc}(u, u) \geq 0 \;\text{ for all timelike }\; u \in TM \,.$$

Roughly speaking, the SEC says that gravity is always attractive. For a perfect fluid, it can only be violated in the presence of negative energy density or pressure (see Hawking and Ellis [**4**, §4.3] for a discussion of the common "energy conditions" in general relativity). In any case, it was observed by Hubble (confirming a prediction of Lemaître motivated by the calculations we have just done) in 1929 that the current value of H is positive. We conclude, based on our model, that *the (spatial) universe has been expanding since some finite cosmological time in the past, at which it emerged, somehow, from a region of infinite energy density and pressure.*

[49] Precisely, consider $w_\varepsilon = w - \varepsilon t$, with w as below, to obtain strict inequality $\dot{w}_\varepsilon(0) < 0$, and continue as below. The claim follows by taking $\varepsilon \searrow 0$.

Now recall the equation (14.5). Since $\alpha > 0$, and we have assumed $\rho + 3p > 0$, we find that $\ddot{\alpha} < 0$. That is, *the expansion of the universe is slowing down*. This suggests the following question: Will the stress energy term $3p + \rho$ be large enough to cause $\dot{\alpha}$ eventually to fall to zero, and become negative, resulting in a recollapse of the universe, or will $3p + \rho$ be sufficiently weak that $\dot{\alpha}$ only asymptotes to zero, resulting in a universe that expands forever? Saul Perlmutter, Adam Riess, and Brian Schmidt recently shared the 2011 Nobel prize in physics for observing, much to their mutual surprise, that neither of these scenarios is true of our universe, since the present value of $\ddot{\alpha}$ is actually positive: *The expansion of the universe is accelerating*! What is this ubiquitous but unseen "dark energy" which violates the SEC, and causes the accelerated expansion of the cosmos?

14.1. The cosmological constant. Recall that the modified Einstein equation in the presence of the "cosmological term" is

$$\mathrm{Rc} - \frac{1}{2}\,\mathrm{R}\,g + \Lambda g = T\,, \tag{14.7}$$

where Λ is a constant. The cosmological term arises naturally in Hilbert's derivation of the Einstein equation, and was used early on by Einstein in an attempt to square the theory with the belief that the universe is static.

In order to see how this might work, we rewrite (14.7) as:

$$\mathrm{Rc} - \frac{1}{2}\,\mathrm{R}\,g = T - \Lambda g \doteqdot T_\Lambda\,.$$

With T as in (14.1), this becomes[50]

$$\mathrm{Rc} - \frac{1}{2}\,\mathrm{R}\,g = (\rho + p)dt^2 + (p - \Lambda)g\,.$$

So "cosmological dark energy" allows us to violate the SEC, satsfied by "vanilla" stress-energy tensors, without requiring the existence of an exotic, unseen energy field of negative pressure: Λ is simply part of the local geometry of the universe[51].

The Friedmann equations now take the form

$$\frac{\ddot{\alpha}}{\alpha} = \dot{H} + H^2 = -\frac{1}{6}(3p + \rho - 2\Lambda)\,, \tag{14.8}$$

$$\dot{\rho} = -3H(p + \rho)\,. \tag{14.9}$$

Einstein now sets $p = 0$ and $\rho = 2\Lambda$ to obtain a static universe (with constant α and ρ). However, once it became clear, from the observations of

[50]We can think of this as reinterpreting the cosmological constant as being caused by a form of ubiquitous but unseen "vacuum energy" (or "dark energy") of constant energy density, $\rho_{\text{vacuum}} = \Lambda$, and constant pressure, $p_{\text{vacuum}} = -\Lambda$. More on this in a moment.

[51]On the other hand, it would be nice to have an explanation of *why* the universe would choose a certain value for Λ and no other.

Slipher and others, that the universe was not static, but rather expanding, Einstein quickly disposed of the cosmological constant, asserting: "If there is no quasi-static world, then away with the cosmological term!"

The cosmological term was resurrected with attempts to combine general relativity and quantum field theory (and observations of the Casimir effect), which produced estimates for the value of the energy of the vacuum. Calculating this energy leads to the "worst[52] prediction in the history of physics" with the predicted value coming out something like 100 orders of magnitude larger the experimental value. The cosmological term is now reinstated to cancel out the vacuum energy. However, these calculations rely on ad hoc arguments, and in any case are far beyond the scope of my knowledge, so I will not discuss them further.

On the other hand (ignoring quantum behaviour) the cosmological constant may be used to account for the accelerated expansion observed by Perlmutter, Riess and Schmidt: returning to the modified Friedmann equations (14.8) and (14.9), observe that, even if normal matter obeys the SEC, we can still get the experimentally determined value for $\ddot{\alpha}_0 > 0$ with the right choice of $\Lambda > 0$. Now, Λ is constant, whilst ρ and p are dynamic. Thus, the dynamic state of the universe will become time dependent, since $\ddot{\alpha}$ depends on the sign of $3p + \rho - 2\Lambda > 0$. However, in the case that $\Lambda < 0$, (14.8) implies that $\ddot{\alpha}/\alpha$ is uniformly negative, and it follows that such a universe will inevitably collapse after some finite universal time. If instead $\Lambda > 0$ the situation is more complicated, with perpetual expansion (accelerated or decelerated) and eventual collapse possible. Moreover, since the energy density decreases as α increases, a universe whose expansion is initially "slowing down" (i.e. with $3p + \rho - 2\Lambda > 0$) can reach a critical time at which the expansion begins to accelerate (i.e. where $3p + \rho - 2\Lambda$ changes sign). The current data for the radiation, mass and cosmological energy densities suggests an eternal accelerated expansion [**3**].

Finally, we note that the Friedmann singularity is very different from that of Schwarzschild: although both metrics experience curvature singularities and geodesic incompleteness, the Schwarzschild metric, on the one hand, is Ricci flat, and hence its curvature singularity occurs at the level of the Weyl curvature. In particular, the curvature singularity cannot be "cured" by conformal transformations of the metric. The Friedmann metric, on the other hand, has controlled Weyl curvature. Indeed, multiplying it by α^{-2} and setting $\tau \doteqdot \int_0^t \alpha(s)ds$, we find that it is conformally equivalent to

[52]On the other hand, one might argue that this is only the *second* worst prediction in the history of physics, since some well-known theorists predict that there are at least 10^{500} universes, out by 500 orders of magnitude from the observed value.

the nonsingular metric

$$\tilde{g} = -d\tau \otimes d\tau + \gamma\,.$$

This observation is central to Roger Penrose's "conformal cyclic cosmology".

Exercises.

Exercise 14.2. *Show that the non-vanishing components of the Ricci tensor of γ are given (in $\{r, \theta, \phi\}$ coordinates) by*

$$\begin{aligned} \mathrm{Rc}_{rr} &= \frac{2}{r}\beta_r \\ \mathrm{Rc}_{\theta\theta} &= e^{-2\beta}(r\beta_r - 1) + 1 \\ \mathrm{Rc}_{\phi\phi} &= \sin^2\theta\ \mathrm{Rc}_{\theta\theta}\,, \end{aligned}$$

where $\beta_r \doteqdot \partial_r \beta$.

Exercise 14.3.

(a) Consider $\mathbb{R}^4$ with cylindrical coordinates (t, r, θ, ϕ), where (r, θ, ϕ) are the standard polar coordinates on $\mathbb{R}^3$, in which the Euclidean metric takes the form

$$\delta = dt \otimes dt + dr \otimes dr + r^2\left(d\theta \otimes d\theta + \sin^2\theta d\phi \otimes d\phi\right).$$

Given $k > 0$, define new coordinates T and R on $\mathbb{R} \setminus \{0\}$ by

$$t = T\sqrt{1 - kR^2} \ \text{ and } \ r = \sqrt{k}RT\,.$$

Show that

$$\delta = dT \otimes dT + kT^2\left(\frac{1}{1 - kR^2}dR \otimes dR + R^2\left(d\theta \otimes d\theta + \sin^2\theta d\phi \otimes d\phi\right)\right).$$

Deduce that the region $(R, \theta, \phi) \in (0, \frac{1}{\sqrt{k}}) \times (0, 2\pi) \times (0, \pi)$ of $\mathbb{R}^3$ equipped with the metric

$$\gamma \doteqdot \frac{1}{1 - kR^2}dR \otimes dR + R^2\left(d\theta \otimes d\theta + \sin^2\theta d\phi \otimes d\phi\right)$$

is locally isometric to the sphere $S^3_{1/\sqrt{k}}$.

(b) Denote by t the standard "time" coordinate on Minkowski space $\mathbb{R}^{3,1}$ and by (r, θ, ϕ) the standard polar coordinates on the pseudo-orthogonal compliment, $\mathbb{R}^3$, in which the Minkowski metric becomes

$$\eta = -dt \otimes dt + dr \otimes dr + r^2\left(d\theta \otimes d\theta + \sin^2\theta d\phi \otimes d\phi\right).$$

Given $k < 0$, define new coordinates T and R on the region

$$\mathcal{J} \doteqdot \{x \in \mathbb{R}^4 : \eta(x, x) < 0\}$$

by

$$t = T\sqrt{1 - kR^2} \ \text{ and } \ r = \sqrt{-k}RT\,.$$

Show that

$$\eta|_{\mathcal{J}} = -dT \otimes dT - kT^2 \left(\frac{1}{1-kR^2} dR \otimes dR + r^2 \left(d\theta \otimes d\theta + \sin^2\theta d\phi \otimes d\phi \right) \right).$$

Deduce that the region $(R, \theta, \phi) \in (0, \infty) \times (0, 2\pi) \times (0, \pi)$ *of* $\mathbb{R}^3$ *equipped with the metric*

$$\gamma \doteqdot \frac{1}{1-kR^2} dR \otimes dR + R^2 \left(d\theta \otimes d\theta + \sin^2\theta d\phi \otimes d\phi \right)$$

is locally isometric to the hyperbolic space $H^3_{1/\sqrt{-k}}$.

Exercise 14.4. *Show that the non-vanishing components of the Ricci tensor of* g *are*

$$\begin{aligned}
\mathrm{Rc}_{tt} &= -3\frac{\ddot{\alpha}}{\alpha}, \\
\mathrm{Rc}_{rr} &= g_{rr} \left(\frac{\ddot{\alpha}}{\alpha} + 2 \left(\frac{\dot{\alpha}}{\alpha} \right)^2 + \frac{2k}{\alpha^2} \right), \\
\mathrm{Rc}_{\theta\theta} &= g_{\theta\theta} \left(\frac{\ddot{\alpha}}{\alpha} + 2 \left(\frac{\dot{\alpha}}{\alpha} \right)^2 + \frac{2k}{\alpha^2} \right), \\
\mathrm{Rc}_{\phi\phi} &= g_{\phi\phi} \left(\frac{\ddot{\alpha}}{\alpha} + 2 \left(\frac{\dot{\alpha}}{\alpha} \right)^2 + \frac{2k}{\alpha^2} \right),
\end{aligned}$$

where $\dot{\alpha} \doteqdot \frac{d}{dt}\alpha$ *and* $\ddot{\alpha} \doteqdot \frac{d^2}{dt^2}\alpha$. *It follows that*

$$\mathrm{R} = 6 \left(\frac{\ddot{\alpha}}{\alpha} + \left(\frac{\dot{\alpha}}{\alpha} \right)^2 + \frac{k}{\alpha^2} \right).$$

Exercise 14.5. *Show that* T *satisfies the* SEC *if and only if*

$$3p + \rho \geq 0 \;\; \textit{and} \;\; p + \rho \geq 0 \,.$$

15. The initial value formulation of Einstein's equation

> *"In so far as a scientific statement speaks about reality, it must be falsifiable: and in so far as it is not falsifiable, it does not speak about reality."* — Karl Popper[53] (The Logic of Scientific Discovery)

In order for general relativity to be a fully predictive theory, it should admit an *initial value formulation* satisfying some version of the following properties:

(1) There is a class of initial conditions (restricted perhaps by certain reasonable constraints) which always provides solutions.

(2) Solutions are determined uniquely by those initial conditions.

(3) Solutions depend continuously on the initial conditions.

A theory satisfying these conditions is called "well-posed".

In classical physics, the physical laws are generally phrased as second order partial differential equations (perhaps the Euler–Lagrange equation of some action functional), whose solutions govern the "state" of the system (whose values are usually observable quantities), depending usually on time and space coordinates. The problem is then solvable by specifying the state of the system and its first derivatives at some initial time and/or at the spatial boundary (the exact nature of the boundary conditions will depend on the PDE at hand).

In general relativity, the problem is not so straightforward, since space and time themselves are part of the system for which we are trying to solve. In order to see how we might escape this conundrum, let us turn the problem around: suppose we are given a solution to Einstein's equation; the solution consists of a four dimensional manifold M equipped with a Lorentzian metric g satisfying the Einstein equation:

$$\mathrm{Rc} - \frac{1}{2}\,\mathrm{R}\,g = T,$$

or, equivalently,

$$\mathrm{Rc} = T - \frac{1}{2}\,\mathrm{tr}(T)g\,.$$

We have seen that this tensor equation can be viewed as a partial differential equation on M. Indeed, in local coordinates $x^i : U \to \mathbb{R}$, it becomes a second

[53] 1902–1994

order system for the unknown components g_{ij} of the metric:

$$\begin{aligned} T_{ij} - \frac{1}{2} g^{kl} T_{kl} g_{ij} & \\ &= \mathrm{Rc}_{ij} \\ &= \mathrm{Rm}_{ikj}{}^{k} \\ &= \partial_k \Gamma_{ij}{}^{k} - \partial_i \Gamma_{kj}{}^{k} + \Gamma_{ij}{}^{q} \Gamma_{kq}{}^{k} - \Gamma_{kj}{}^{q} \Gamma_{iq}{}^{k} \\ &= \frac{1}{2} g^{km} \left(\partial_k \partial_m g_{ij} + \partial_i \partial_j g_{km} - \partial_k \partial_j g_{im} - \partial_i \partial_m g_{jk} \right) \\ &\quad - \frac{1}{2} g^{km} g^{ln} \partial_k g_{ln} \left(\partial_i g_{jm} + \partial_j g_{im} - \partial_m g_{ij} \right) \\ &\quad + \frac{1}{2} g^{km} g^{ln} \partial_k g_{ln} \left(\partial_k g_{jm} + \partial_j g_{km} - \partial_m g_{jk} \right) \\ &\quad + \frac{1}{4} g^{mn} g^{ko} \left(\partial_i g_{jn} + \partial_j g_{in} - \partial_n g_{ij} \right) \left(\partial_k g_{mo} + \partial_m g_{ko} - \partial_o g_{km} \right) \\ &\quad - \frac{1}{4} g^{mn} g^{ko} \left(\partial_k g_{jn} + \partial_j g_{kn} - \partial_n g_{kj} \right) \left(\partial_i g_{mo} + \partial_m g_{io} - \partial_o g_{im} \right) . \end{aligned} \tag{15.1}$$

Needless to say, this is not a very nice equation. Even ignoring the complicated mix of lower order terms, it is a degenerate nonlinear system. However, we have not made full use of the geometry at hand.

15.1. The 3+1 split. Locally, any observer $\gamma : I \to M$ gives rise to a splitting of spacetime into space+time: at any given time, which we without loss of generality take to be 0, the "instantaneous rest space" $\gamma'(0)^{\perp} \subset T_{\gamma(0)} M$ may be locally "integrated" to obtain a spacelike hypersurface $\Sigma_0 \subset M$ by shooting out (spacelike) geodesics from $o \doteqdot \gamma(0)$ in directions $v \in \gamma'(0)^{\perp}$. This hypersurface consists of events which γ may interpret as occurring simultaneously with $\gamma(0)$. By parallel translating $U_o \doteqdot \gamma'(0)$ along these geodesics, we may then shoot out (timelike) geodesics from each $p \in \Sigma_0$ in the timelike direction of U_p (the parallel translate of U_o at p). We may interpret these geodesics as a family of freefallers which are instantaneously comoving with γ at time 0.

Denote by U the vector field on the resulting open subset $N \subset M$ which gives at each $q \in N$ the tangent vector to the comoving freefaller at q, by $t : N \to \mathbb{R}$ the function which assigns to each $q \in N$ the proper time along the comoving freefaller which joins q to Σ_0, and, abusing notation, by $\Sigma_t \doteqdot \{p \in M : t(p) = t\}$ the t-level set of the function t. By the Gauss lemma, the metric restricted to N takes the form

$$g|_N = -dt \otimes dt + h \,,$$

where $h(U, \cdot) = 0$ and $\iota_t^* h$ is a Riemannian metric on Σ_t for each t (we will drop the inclusion map ι_t hereafter).

Our goal is to view Einstein's equation as an evolution equation for h with initial data $(\Sigma_t, h, \partial_t h)$ at $t = 0$, say (since we expect a hyperbolic equation for h, we expect that we should need to prescribe both h and $\partial_t h$ at the initial time).

Denote by $\nabla^\top$ the connection induced on Σ_t. Observe that

$$\begin{aligned}(\mathcal{L}_U h)(X,Y) &= U(h(X,Y)) - h(\mathcal{L}_U X, Y) - h(X, \mathcal{L}_U Y)\\ &= h(\nabla_U X - [U,X], Y) + h(X, \nabla_U Y - [U,Y])\\ &= h(\nabla_X U, Y) + h(X, \nabla_Y U)\\ &= 2A(X,Y)\,, \end{aligned} \tag{15.2}$$

where A is the second fundamental form of Σ_t corresponding to U. We interpret this as the (first order) evolution equation for the spatial metric. Indeed, if we compliment t with "spatial" coordinates for Σ_t, then

$$\partial_t h_{ij} = (\mathcal{L}_{\partial_t} h)_{ij} = 2A_{ij} \tag{15.3}$$

where here i and j range over the "spatial" indices, $1, 2, 3$.

Remark 15.1. *More generally, if t is any function whose level sets are spacelike hypersurfaces, then, setting*

$$U \doteqdot \frac{\operatorname{grad} t}{|\operatorname{grad} t|}\,,$$

we can decompose $\partial_t = dt^\sharp$ as

$$\partial_t = \alpha U + \beta$$

*for some function α (the **lapse function**) and some vector β tangent to the level sets Σ_t (the **shift vector**). The metric then takes the form*

$$g = -\alpha^2 dt \otimes dt + \frac{1}{2}(dt \otimes \beta^\flat + \beta^\flat \otimes dt) + h\,,$$

and the time derivative of h becomes

$$\mathcal{L}_{\partial_t} h = 2\alpha A + \mathcal{L}_\beta h\,.$$

To obtain an evolution equation for A, consider

$$\begin{aligned}
\mathrm{Rm}(X,U,U,Y) &= g(\nabla_U(\nabla_X U) - \nabla_X(\nabla_U U) - \nabla_{[U,X]}U, Y) \\
&= g(\nabla_U(\nabla_X U), Y) - g(\nabla_X(\nabla_U U), Y) - g(\nabla_{[U,X]}U, Y) \\
&= Ug(\nabla_X U, Y) - g(\nabla_X U, \nabla_U Y) - g(\nabla_X(\nabla_U U), Y) \\
&\quad - g(\nabla_{[U,X]^\top}U + \nabla_{[U,X]^\perp}U, Y) \\
&= Ug(\nabla_X U, Y) - g(\nabla_X U, \nabla_Y U + [U,Y]) - g(\nabla_X(\nabla_U U), Y) \\
&\quad - g(\nabla_{[U,X]^\top}U, Y) - g([U,X],U)g(\nabla_U U, Y) \\
&= UA(X,Y) - A([U,X]^\top, Y) - A(X,[U,Y]^\top) \\
&\quad - A^2(X,Y) - g(\nabla_X(\nabla_U U), Y) - g([U,X],U)g(\nabla_U U, Y)\,,
\end{aligned}$$

If we define the Lie derivative $\mathcal{L}_U A$ of A in the time direction U by the equation

$$U(A(X,Y)) = (\mathcal{L}_U A)(X,Y) + A\big((\mathcal{L}_U X)^\top, Y\big) + A\big(X, (\mathcal{L}_U Y)^\top\big)\,,$$

then, since we have assumed that the integral curves of U are geodesic, we obtain

$$\mathrm{Rm}(X,U,U,Y) = (\mathcal{L}_U A)(X,Y) + A^2(X,Y)\,, \tag{15.4}$$

which, by (15.3), can be interpreted as a second order time evolution of h.

By the Gauss equation,

$$\begin{aligned}
\mathrm{Rc}(X,Y) =& -\mathrm{Rm}(X,U,Y,U) + \mathrm{tr}_{T\Sigma_t}\mathrm{Rm}(X,\cdot,Y,\cdot) \\
=& \ \mathrm{Rm}(X,U,U,Y) + \mathrm{Rc}^h(X,Y) - A^2(X,Y) + HA(X,Y)
\end{aligned} \tag{15.5}$$

for directions X, Y tangent to Σ_t, where $H \doteqdot \mathrm{tr}_h(A)$ is the mean curvature of Σ_t.

Applying (15.5) to (15.4) yields

$$(\mathcal{L}_U A)(X,Y) = \mathrm{Rc}(X,Y) - \mathrm{Rc}^h(X,Y) + 2A^2(X,Y) - HA(X,Y) \tag{15.6}$$

for X, Y tangent to the level sets Σ_t. By Einstein's equation, the right hand side of (15.6) only involves source terms and the geometric data (Σ_t, h, A).

Taking the trace of (15.5) yields

$$\begin{aligned}
\mathrm{R} =& -\mathrm{Rc}(U,U) + \mathrm{tr}_{T\Sigma_t}(\mathrm{Rc}) \\
=& -\mathrm{Rc}(U,U) - \mathrm{tr}_{T\Sigma_t}\mathrm{Rm}(\cdot,U,\cdot,U) + \mathrm{R}^h - |A|^2 + H^2 \\
=& -2\,\mathrm{Rc}(U,U) + \mathrm{R}^h - |A|^2 + H^2\,,
\end{aligned}$$

where

$$\mathrm{R}^h \doteqdot \mathrm{tr}_{T\Sigma_t}(\mathrm{Rc}^h)$$

is the scalar curvature of (Σ_t, h).

Thus,

$$\begin{aligned} G(U,U) &= \mathrm{Rc}(U,U) - \frac{1}{2}\,\mathrm{R}\,g(U,U) \\ &= \mathrm{Rc}(U,U) + \frac{1}{2}\,\mathrm{R} \\ &= \frac{1}{2}\left(\mathrm{R}^h - |A|^2 + H^2\right). \end{aligned} \tag{15.7}$$

The mixed space-time components of G are given by taking the trace of the Codazzi equation,

$$\mathrm{Rm}(X,Y,Z,U) = \nabla^\top_X A(Y,Z) - \nabla^\top_Y A(X,Z)\,,$$

which yields

$$\mathrm{Rc}(Y,U) = \mathrm{div}_h\, A(Y) - YH$$

for X tangent to Σ_t. Thus,

$$\begin{aligned} G(X,U) &= \mathrm{Rc}(X,U) - \frac{1}{2}\,\mathrm{R}\,g(X,U) \\ &= \mathrm{div}_h\, A(X) - XH\,. \end{aligned} \tag{15.8}$$

Writing

$$T = \rho dt \otimes dt + \frac{1}{2}\left(dt \otimes J + J \otimes dt\right) + P\,,$$

where $P(U,\cdot) = J(U) = 0$, equations (15.6), (15.7), and (15.8) split Einstein's equation into the equations

$$\frac{1}{2}\left(\mathrm{R}^h + H^2 - |A|^2\right) = \rho \tag{15.9a}$$

$$\mathrm{div}_h\, A - dH = J \tag{15.9b}$$

$$\mathcal{L}_t A + \mathrm{Rc}^h - 2A^2 + HA = P + \frac{1}{2}(\rho - \mathrm{tr}_{T\Sigma_t}(P))h\,. \tag{15.9c}$$

The first two equations, (15.9a)-(15.9b), only involve the geometric data (Σ_t, h, A) plus the source terms (ρ, J). As a consequence of the second Bianchi identity, they are automatically satisfied at time t if they are satisfied initially and if the spatial equation (15.9c) is satisfied. So equations (15.9a)-(15.9b) only serve as **constraint equations** for the initial data. The remaining equation is formally a second order evolution equation for h.

15.2. Solving the Einstein equation via harmonic coordinates. We saw that the constraint equations (15.9a)-(15.9b) are necessary conditions on (Σ, h, A) for the data to arise from a spacelike hypersurface in a spacetime satisfying the Einstein equation. It turns out that they are also sufficient conditions for the existence of a spacetime satisfying the Einstein equations in which (Σ, h) embeds isometrically with second fundamental form A. We will describe the procedure for constructing such a spacetime in this section.

The idea is to solve the system of equations (15.1) obtained by writing the Einstein equation in coordinates, and use the resulting solution, g_{ij}, to construct our spacetime solution "by hand". Unfortunately, the system (15.1) is highly degenerate, which prevents the immediate application of "standard" PDE results. This is not a bug, however, but a feature — it results from the general covariance of the Einstein equation under diffeomorphisms, as required by the equivalence principle; we will mod-out this covariance by choosing a special class of coordinates, resulting in a non-degenerate system of hyperbolic equations for the metric coefficients.

Harmonic coordinates (a.k.a. **wave coordinates** in the Lorentzian context) on a pseudo-Riemannian manifold are local coordinates $x : U \to \mathbb{R}^n$ which satisfy

$$-\Delta x^i = 0\,, \tag{15.10}$$

where Δ is the Laplace–Beltrami operator,

$$\Delta u \doteqdot \mathrm{tr}_g(\nabla^2 u)\,.$$

If we assume that the coordinates in which we are working are harmonic, then, by (15.12), we obtain the four additional equations

$$\begin{aligned} 0 = -\Delta x^n &= -g^{kl}\left(\frac{\partial x^n}{\partial x^k \partial x^l} - \Gamma_{kl}{}^m \frac{\partial x^n}{\partial x^m}\right) \\ &= g^{kl}\Gamma_{kl}{}^n \\ &= g^{mn} g^{kl}\left(\partial_k g_{lm} + \partial_l g_{km} - \partial_m g_{kl}\right). \end{aligned} \tag{15.11}$$

We will use these equations to eliminate the degenerate terms in (15.1).

Remark 15.2. *Note that harmonic coordinates always exist about any point on any Lorentzian (or Riemannian) manifold: given any $p \in M$, choose any coordinates $\phi : U \to \mathbb{R}^n$ about p, and denote by g^{kl} and $\Gamma_{kl}{}^m$ the components of the metric and connection in the ϕ-coordinates. Classical* PDE *methods allow us to solve the boundary value problem*

$$\begin{aligned} 0 &= -(g^{kl}\circ\phi^{-1})\left(\frac{\partial x^i}{\partial y^k \partial y^l} - (\Gamma_{kl}{}^m \circ \phi^{-1})\frac{\partial x^i}{\partial y^m}\right), \\ x^i(\phi(p)) &= \phi^i(p), \quad \frac{\partial x^i}{\partial y^j}(\phi(p)) = \frac{\partial(\phi^i\circ\phi^{-1})}{\partial y^j}(p)\,. \end{aligned} \tag{15.12}$$

for functions $\{x^0, x^1, x^2, x^3\}$ on some ball about $\phi(p) \in \mathbb{R}^n$. It follows that the functions $x^i \circ \phi$ are harmonic. Since $\frac{\partial x^i}{\partial y^j}(\phi(p))d\phi^j$ are linearly independent, $(\frac{\partial x^i}{\partial y^j}\circ\phi)d\phi^j$ are linearly independent in a neighbourhood of p, so $\{x^i\circ\phi\}_{i=0}^3$ define local coordinates on this neighbourhood.

Differentiating (15.11) yields

$$\begin{aligned} 0 = & -g_{in}\partial_j(\Delta x^n) - g_{jn}\partial_i(\Delta x^n) \\ & = g_{in}\partial_j(g^{km}g^{ln}\Gamma_{kml}) + g_{jn}\partial_i(g^{km}g^{ln}\Gamma_{kml}) \\ & = g_{in}\partial_j(g^{km}g^{ln})\Gamma_{kml} + g_{in}g^{km}g^{ln}\partial_j\Gamma_{kml} \\ & \quad + g_{jn}\partial_i(g^{km}g^{ln})\Gamma_{kml} + g_{jn}g^{km}g^{ln}\partial_i\Gamma_{kml}\,. \end{aligned}$$

The two terms involving derivatives of the connection coefficients are

$$\begin{aligned} g^{km}\left(\partial_i\Gamma_{kmj} + \partial_j\Gamma_{kmi}\right) &= \frac{1}{2}g^{km}\big(\partial_i\partial_k g_{mj} + \partial_i\partial_m g_{kj} - \partial_i\partial_j g_{km} \\ &\quad + \partial_j\partial_k g_{mi} + \partial_j\partial_m g_{ki} - \partial_j\partial_i g_{km}\big) \\ &= g^{km}\big(\partial_i\partial_m g_{kj} + \partial_j\partial_k g_{mi} - \partial_j\partial_i g_{km}\big), \end{aligned}$$

where we exploited the fact that the metric coefficients are symmetric. Thus,

$$\begin{aligned} g^{km}\big(\partial_i\partial_m g_{kj} + \partial_j\partial_k g_{mi} - \partial_j\partial_i g_{km}\big) &= -g_{in}\partial_j(g^{km}g^{ln})\Gamma_{kml} - g_{jn}\partial_i(g^{km}g^{ln})\Gamma_{kml} \\ &= -\partial_j g^{km}\Gamma_{kmi} - \partial_j g^{km}\Gamma_{kmj} \\ &= g^{kp}g^{mq}\left(\partial_j g_{pq}\Gamma_{kmi} + \partial_j g_{pq}\Gamma_{kmi}\right). \end{aligned}$$

Using this identityt, the coordinate formula (15.1) for the Einstein equation becomes the inhomogeneous quasi-linear wave equation

$$T_{ij} - \frac{1}{2}g^{kl}T_{kl}g_{ij} = \frac{1}{2}g^{km}\partial_k\partial_m g_{ij} + Q(g,\partial g)\,, \tag{15.13}$$

where $Q(g,\partial g)$ is a term which is quadratic in the components g^{kl} and $\partial_i g_{jk}$.

Given a suitable stress-energy tensor T, initial data (Σ, h, A) satisfying the constraint equations, and a local harmonic coordinate chart (x, U) on Σ, the reduced Einstein equation (15.13) with initial conditions

$$\big[g_{ij}\big]\big|_{x(U)} = \begin{bmatrix} -1 & 0 \\ 0 & h_{ij} \end{bmatrix} \quad \text{and} \quad \big[\partial_{x_0} g_{ij}\big]\big|_{x(U)} = \begin{bmatrix} 0 & 0 \\ 0 & A_{ij} \end{bmatrix}$$

can be attacked using PDE methods. We do not want to stray into the vast wilderness of PDE theory here, so let us simply assume that we do indeed obtain a solution g_{ij} on $U \times (-\delta, \delta)$ for some $\delta > 0$ (uniquely and continuously depending on the initial data). Then we still need to show 1. that the functions g_{ij} do indeed correspond to a solution to the Einstein equation (this is not obvious, since a solution to the reduced Einstein equation only corresponds to a solution to the Einstein equation if the coordinates are *harmonic* with respect to g — a priori, only a subset of the harmonic coordinate conditions hold, and only on the initial hypersurface, $\{0\} \times x(U)$), and 2. that the local solutions can be patched together to obtain a global solution (M, g) in which (Σ, h) embeds isometrically with second fundamental form equal to A.

It turns out that the harmonic conditions hold automatically, as a consequence of the second Bianchi identity; so the coefficients g_{ij} do indeed correspond to a (local) solution to the Einstein equation. It is also not too difficult to "patch together" local solutions to obtain a solution spacetime in which (Σ, h) embeds globally. Local uniqueness of solutions to the reduced equation guarantees that there is a unique "largest" spacetime arising from (Σ, h) by solving the reduced Einstein equation[54]. Putting everything together, we arrive at the **Choquet-Bruhat Theorem**:

Theorem 15.3 (Choquet-Bruhat, 1952, 1962, Choquet-Bruhat–Geroch, 1969)**.** *Let (Σ, h) be a smooth Riemannian three-manifold equipped with a symmetric covariant two-tensor field A. If h and A satisfy the constraint equations* (15.9a)-(15.9b)*, then there exists a smooth Lorentzian four manifold (M, g) satisfying the vacuum Einstein equation and the following five properties:*

(1) (Σ, h) embeds isometrically into (M, g) with second fundamental form A.

(2) Every inextendable causal (timelike or null) curve in M intersects Σ in exactly one point[55]*.*

(3) Every globally hyperbolic solution to the vacuum Einstein equation with Cauchy hypersurface Σ maps isometrically into (M, g).

(4) If (M', g') is a spacetime satisfying the vacuum Einstein equation and (i)-(iii) with data (Σ, h, A) replaced by (Σ', h', A'), and there is a diffeomorphism $\phi : S \to S'$ from $S \subset \Sigma$ to $S' \subset \Sigma'$ such that $\phi^* h' = h$ and $\phi^* A' = A$, then the set $D(S) \subset \Sigma$ of points in M which are connected to S by causal curves[56] *is isometric to $D(S') \subset \Sigma$, the set of points in M' which are connected to S' by causal curves.*

(5) (M, g) depends continuously[57] *on (Σ, h, A).*

A full proof of Theorem 15.3 can be found in Hawking and Ellis [**4**] (see also Wald [**8**]).

The Choquet-Bruhat Theorem also holds for the non-vacuum Einstein equation, so long as the stress-energy tensor comes from matter fields satisfying suitable dynamical equations.

[54] There could be larger spacetimes containing Σ, but these will have points which cannot be reached from Σ by causal curves.

[55] The hypersurface Σ is called a **Cauchy hypersurface** for (M, g). A spacetime which admits a Cauchy hypersurface is called **globally hyperbolic**.

[56] The set $D(S)$ is called the **Cauchy development** of S.

[57] In an appropriate topology. See [**4**].

16. The Penrose singularity theorem

> *"The theorems predict singularities in two situations. One is in the future in the gravitational collapse of stars and other massive bodies. Such singularities would be an end of time, at least for particles moving on the incomplete geodesics. The other situation in which singularities are predicted is in the past at the beginning of the present expansion of the universe. This led to the abandonment of attempts (mainly by the Russians) to argue that there was a previous contracting phase and a non singular bounce into expansion. Instead almost everyone now believes that the universe, and time itself, had a beginning at the Big Bang. This is a discovery far more important than a few miscellaneous unstable particles but not one that has been so well recognized by Nobel prizes."* — Stephen Hawking and Roger Penrose, *The nature of space and time*.[58]

We have seen that both the Schwarzschild and Friedmann metrics are "singular": in both cases, the norm of the curvature tensor becomes infinite[59] in certain coordinate limits ($r \to 0$ and $t \to 0$, respectively). Such a limit point cannot be a part of our spacetime solution. On the other hand, it does not seem too objectionable for spacetime to become infinitely curved somewhere, so long as the point of infinite curvature is "inaccessible" (infinitely far away in proper time for any observer, say). A terrifying feature of the Schwarzschild and Friedmann solutions is that their "points" of infinite curvature *are* accessible, in the sense that there are observers which reach them in finite proper time (to the past, in the case of the Friedmann solution). At this final time, something rather unpleasant must happen to the observer, as the curve cannot be extended for proper times greater than or equal to the time at which the singularity is reached. It is this behaviour, *geodesic incompleteness*, that we investigate here.

Recall that the geodesic equation

$$\begin{cases} 0 = \gamma'' \doteq \nabla_s \gamma' = \left(\frac{d^2\gamma^k}{ds^2} + \frac{d\gamma^i}{ds} \frac{d\gamma^j}{ds} \Gamma_{ij}{}^k \circ \gamma \right) \partial_k \\ (\gamma(0), \gamma'(0)) = (p, v) \end{cases} \tag{16.1}$$

always has a solution $\gamma : (-s_0, s_0) \to M$, at least for sufficiently small s_0.

We will refer to a geodesic $\gamma : (-s_0, s_0) \to M$ whose tangent vector γ' is everywhere timelike, spacelike or null, respectively, as **timelike**, **spacelike**

[58] Stephen Hawking (1942–2018); Roger Penrose (1931–).

[59] Recall though that the nature of the blow-up differs between the two cases in that the Schwarzschild metric, being Ricci flat, has a Weyl curvature singularity, whereas the Friedmann metric, being conformal to a constant curvature metric, has a Ricci curvature singularity.

or **null**, respectively. Note that every geodesic is either timelike, spacelike or null, since

$$0 = \frac{d}{ds} g(\gamma', \gamma') = 2g(\gamma'', \gamma') = 0 \quad \Longrightarrow \quad g(\gamma', \gamma') \equiv g(\gamma'(0), \gamma'(0)) \,.$$

Definition 16.1. *A Lorentzian manifold (M, g) is called **timelike**, **null**, or **spacelike geodesically complete**, respectively, if solutions to* (16.1) *with timelike, null, or spacelike, respectively initial condition are defined for all $s \in \mathbb{R}$.*

If (M, g) is Riemannian, then, by the Hopf–Rinow theorem, (spacelike) geodesic completeness is equivalent to good-old-fashioned metric space completeness, where the **Riemannian distance function** is defined by

$$d(p, q) \doteqdot \inf_{\gamma} \int_0^1 |\gamma'(s)| ds \,,$$

with the infimum taken over piecewise regular curves $\gamma : [0, 1] \to M$ satisfying $\gamma(0) = p$ and $\gamma(1) = q$. Moreover, if the infimum is achieved, it is achieved by, and only by, a geodesic, since Riemannian geodesics are the local minimizers[60] of the **length functional**

$$L(\gamma) \doteqdot \int_a^b \sqrt{|g(\gamma'(s), \gamma'(s))|} ds \,.$$

In the Lorentzian case, however, this does not work so nicely, since L is not differentiable near a null geodesic. On the other hand, if we restrict attention to spacelike curves, we find that spacelike geodesics are local length minimizers of L. If we restrict attention to timelike curves, we find that timelike geodesics are also stationary points of L, although they turn out to be local length *maximizers*.

Alternatively, geodesics, both in the Riemannian and pseudo-Riemannian settings, are local minimizers of the **Dirichlet energy**

$$E(\gamma) \doteqdot \frac{1}{2} \int_a^b |\gamma'(s)|^2 ds \,.$$

16.1. Singularity theorems. Broadly speaking, a singularity theorem is a statement giving sufficient conditions for a spacetime to have a singularity. They are usually of the following form.

Generic singularity theorem. *If the spacetime (M, g) satisfies the following:*

(1) An energy (curvature) condition;

(2) A causality (topological) condition;

[60] Amongst piecewise regular curves with fixed endpoints.

(3) Somewhere gravity is strong enough to trap light,

then it is causally geodesic incomplete.

Now, although the Schwarzschild and Friedmann solutions are singular, until Penrose's singularity theorem (see below) appeared in 1965, it was not at all clear that singularities were really to be found in nature — it appeared possible that the Schwarzschild and Friedmann singularities were a consequence of the high degree of symmetry/homogeneity assumed in their derivation, a degree of which cannot in reality occur in nature; stellar collapse, for example, can never happen in a perfectly symmetric way, and the universe is clearly *not* homogeneous. It was conjectured by some that any lack of symmetry would cause all the matter to "miss" the singular point and reexpand. Penrose's theorem shows that singularities, at least the Schwarzschild and Friedmann singularities, cannot be perturbed away, at least within a very large and physically reasonable class of spacetimes.

Theorem 16.2 (Penrose singularity theorem)**.** *Let (M, g) be a smooth* ***spacetime*** *(a four dimensional, time oriented Lorentzian manifold (M, g) that satisfies Einstein's equation) satisfying the following conditions:*

(1) $\mathrm{Rc}(N, N) \geq 0$ *for all null vectors $N \in TM$;*

(2) M contains a non-compact Cauchy hypersurface $\Sigma \subset M$;

(3) M contains a closed, trapped surface $\tau \subset M$.

Then (M, g) is null geodesically incomplete.

Before proving Penrose's theorem, we had best discuss its conditions:

(1) The first condition, known as the **null energy condition**, is the eminently reasonable assumption that gravity always has a non-diverging effect on light rays. By Einstein's equation, it is equivalent to $T(N, N) \geq 0$ for all null vectors N. This is implied, in particular, by the weak energy condition — $T(u, u) \geq 0$ for all timelike u — which (according to Hawking and Ellis [**4**]) has been experimentally confirmed for all known forms of energy[61].

(2) A **Cauchy hypersurface** is a spacelike hypersurface (i.e. a three dimensional submanifold whose tangent vectors are spacelike) of M having the property that it intersects every smooth, inextendible causal (i.e. timelike or null) curve exactly once. For such a spacetime, every event is either in the causal future or the causal past of Σ, and is determined uniquely by Σ. Cauchy hyersurfaces can be used as initial data for Einstein's equation (see §15). The existence of a *non-compact* Cauchy surface assumes that "space" is

[61] Although there is some debate about the role of quantum vacuum energy.

non-compact, which certainly may not be true; however, the conclusion still for spacetimes with compact space slices, so long as there is *some* observer that doesn't fall into the "horizon" (beyond which lies the region described by condition 3). On the other hand, if every observer does fall behind the horizon, a theorem of Hawking and Penrose states, under slightly different conditions, that we should expect a singularity anyway (this theorem applies to "cosmological" singularities such as those of the Friedmann–Lemaître–Robertson–Walker universes).

(3) A closed (i.e. compact and boundaryless), **trapped** surface $\tau \subset M$ can be thought of as an *event horizon*[62]: it is a smooth, closed, two-dimensional, spacelike submanifold of M whose mean curvature vector H is past pointing timelike. This condition means that both outward and inward directed causal curves are converging at τ. Indeed, if $V \in \Gamma(T\tau)$ is a future pointing causal vector field on τ, then the first variation formula yields

$$\frac{d}{d\varepsilon}\Big|_{\varepsilon=0} A(\tau + \varepsilon V) = -\int g(H, V) d\mu \leq 0$$

with strict inequality unless $V \equiv 0$, where A is the area functional and

$$\tau + \varepsilon V \doteqdot \{\gamma_{(p,V(p))}(\varepsilon) : p \in \tau\}$$

is obtained by moving τ a parameter distance ε along the geodesics with initial data $(p, V(p))$, $p \in \tau$. So the area is locally decreasing about any $p \in \tau$ in the direction of any nonzero future pointing causal vector field — *all causal curves emanating from the surface are pulled inwards, (even light emitted in the outwards normal direction!).*

Proof of the Penrose singularity theorem. The proof is a *reductio ad absurdum*; it proceeds as follows: Assuming that (M, g) is null geodesically complete, we can show that $\partial J_+(\tau)$, the boundary of the causal future of τ (including τ itself), is a compact, boundaryless manifold (loosely speaking, the light rays emitted outwards from τ must converge and 'close up'). On the other hand $\partial J_+(\tau)$ maps in a continuous one to one manner into the non-compact hypersurface Σ. This implies either that $\partial J_+(\tau)$ is non-compact, or that it has a boundary, neither of which are true; unless, of course, our assumption that (M, g) were null geodesically complete is false.

First, we introduce some notation alluded to above: If $p \in M$, then we denote by $J_+(p)$ the set of points in M lying on future pointing causal

[62] In fact, the closed trapped surface lies strictly inside what we usually think of as the horizon, since we require the outward fired lightrays to be *strictly converging.*

(timelike or null) curves beginning at p. We call $J_+(p)$ the **causal future** of p. Considering only the future pointing timelike curves emanating from p, we obtain the **chronological future**[63] of p, denoted by $I_+(p)$. There are analogous definitions for $J_-(p)$ and $I_-(p)$, the **causal** and **chronological past** of p. The **causal** (**chronological**) **future** (**past**) of a set $P \subset M$ is defined as the union of the causal (chronological) future (past) of all of its points and denoted, for example, by $J_+(P)$. It is well to familiarize yourself with these notations, as we will use them heavily in what follows.

We first prove a rather general causality result for spacelike surfaces.

Claim 16.3. *Each point*[64] $p \in E_+(\tau) \doteqdot J_+(\tau) \setminus I_+(\tau)$ *lies on a future pointing null geodesic emanating from* τ *orthogonally.*

Thinking of τ as the surface of a sphere, Claim 16.3 simply says that the fastest we can transmit information from τ to some point p in space is via a light ray emitted radially outwards from τ in the direction of p.

We prove Claim 16.3 in two parts, first showing that non-orthogonal null geodesics from τ to p can be deformed into timelike curves from τ to p, and then showing that non-geodesic null curves from p to q can either be deformed into timelike curves from p to q or reparametrised to make them geodesic.

Proposition 16.4. *If* $n : [0, b] \to M$ *is a future pointing null geodesic from* $n(0) \in \tau$ *to* $q = n(b)$ *which is not orthogonal to* τ*, then there is a timelike curve from* τ *to* q *arbitrarily close to* n*.*

Proof. The idea is to move $n(0)$ along τ in the direction that makes $n'(0)$ more orthogonal to τ, while keeping the endpoint q fixed. This deformation decreases $g(n', n')$, therefore making the curve timelike.

Since $n'(0)$ is not normal to τ, there is a vector $u \in T_{n(0)}\tau$ such that $g(u, n'(0)) > 0$. We will move the endpoint $n(0)$ along τ in the direction of u. First parallel translate u along n to obtain a vector field U along n. Now, moving n with velocity U won't keep the endpoint q fixed, so we linearly rescale U to form the vector field $V(s) \doteqdot (1 - s/b)U(s)$. We now have $V(b) = 0$ and $V(0) \in T_{n(0)}\tau$. So consider a variation $\omega : [0, b] \times (-\epsilon_0, \epsilon_0) \to M$ of n, through curves joining τ to q, in the direction of V. That is, with longitudinal velocity $\partial_\epsilon|_{\epsilon=0}\omega = V$. Note that $V' \doteqdot \nabla_s V = -U/b$, so that $g(V', n') = -g(U, n')/b = -g(u, n'(0))/b < 0$. The second equality holds because n' is also parallel along n (n is geodesic), and parallel translation is an isometry.

[63] We consider the trivial curve to be a null curve, so that $p \in J_+(p)$ but $p \notin I_+(p)$.

[64] The set $E_+(q) \doteqdot J_+(q) \setminus I_+(q)$ is called the **future horismos** of q.

We will now show that, for ϵ_1 sufficiently small, the curves $\omega_\epsilon(s) \doteqdot \omega(s,\epsilon)$ for $0 < \epsilon < \epsilon_1$ are timelike.

Observe that $g(\omega'_\epsilon(s), \omega'_\epsilon(s))|_{\epsilon=0} = g(n'(s), n'(s)) = 0$ for all s. On the other hand,

$$\begin{aligned}\partial_\epsilon\big|_{\epsilon=0} g(\omega'_\epsilon(s), \omega'_\epsilon(s)) &= \partial_\epsilon\big|_{\epsilon=0} g(\partial_s\omega(\epsilon,s), \partial_s\omega(\epsilon,s)) \\ &= 2g(\nabla_\epsilon\partial_s\omega(\epsilon,s), \partial_s\omega(\epsilon,s))|_{\epsilon=0} \\ &= 2g(\nabla_s\partial_\epsilon\omega(\epsilon,s), \partial_s\omega(\epsilon,s))|_{\epsilon=0} \\ &= 2g(V'(s), n'(s)) < 0\,.\end{aligned}$$

Therefore, for all s, $g(\omega'_\epsilon(s), \omega'_\epsilon(s))$ is a strictly decreasing function of ϵ at $\epsilon = 0$ (where it vanishes). It follows that $g(\omega'_\epsilon(s), \omega'_\epsilon(s))$ is negative for sufficiently small positive ϵ. This completes the proof. □

Proposition 16.5. *If $n : [0,b] \to M$ is a future pointing non-geodesic null curve joining $n(0) \in \tau$ to a point $n(b) \in J_+(\tau)$, then either*

(a) there is a geodesic reparametrisation[65] of n; or

(b) there is a point on n that lies in $I_+(\tau)$.

In the second case, $n(b)$ cannot lie on $E_+(\tau)$.

Proof. To see why this holds, we consider the following local construction: Observe that two future pointing light cones, one of whose vertex lies on the other, can only intersect on a single null ray, unless the two vertices coincide. The 'lower cone' represents $E_+(p)$ and the 'upper cone' represents the possible directions for a null curve through its vertex. Notice that the upper cone, apart from the line of intersection, lies entirely inside the 'lower' cone. Therefore, all of the possible tangent directions to n either point in the direction of a null geodesic or into the interior of the lower cone. We now formalise this argument:

First observe that, if n is a null curve, then

$$0 = \partial_s g(n', n') = 2g(n'', n')\,,$$

so that n'' lies in the orthogonal compliment of n', which implies that it is either spacelike or null. If n'' is parallel to n', then n has a geodesic reparametrisation. So it suffices to assume that the function $g(n'', n'') \geq 0$ is not identically zero. We will use this fact to construct a variation field which vanishes at the endpoints of the interval and satisfies $g(V', n') < 0$. It then follows as in Proposition 16.4 that a variation of n in the direction of V (which this time fixes *both* endpoints) gives is a nearby timelike curve joining $n(0)$ and $n(b)$. It follows that $n(b) \in I_+(n(0)) \subset I_+(\tau)$. So let's construct V.

[65] In this case, we call n **pregeodesic**.

Let U be a parallel future pointing timelike vector field along n and define $V = \alpha U + \beta n''$, where α and β are determined as follows: First, we chose β to be a smooth non-negative function vanishing at the endpoints 0 and b, and normalised such that

$$\int_0^b \beta \frac{g(n'', n'')}{g(U, n')} = -b$$

This is possible because $g(n'', n'')/g(U, n') < 0$. Now define α by

$$\alpha(u) \doteqdot \int_0^u \left[\beta \frac{g(n'', n'')}{g(U, n')} + 1\right] ds\,.$$

Then $\alpha(0) = \alpha(b) = 0$ and

$$\alpha'(s) = \beta \frac{g(n'', n'')}{g(U, n')} + 1 > \beta \frac{g(n'', n'')}{g(U, n')}\,.$$

We now have

$$g(V', n') = \alpha' g(U, n') - \beta g(n'', n'') < 0$$

as required. □

Exercise 16.1. *If γ is a curve satisfying $\gamma'' = \alpha\gamma'$ for some function α along γ, then there is a reparametrisation of γ which is geodesic.*

This completes the proof of Claim 16.3. We will now show that $\partial J_+(\tau)$, the boundary of the future of τ, is precisely $E_+(\tau)$. This depends crucially on the existence of a Cauchy surface, and can otherwise fail[66].

Claim 16.6. *$\partial J_+(\tau) = E_+(\tau)$.*

We will need the following proposition, whose proof would distract from our current purposes, but may be found in [**4, 5, 7**].

Proposition 16.7. *A spacetime contains a Cauchy hypersurface if and only if it is **globally hyperbolic**; that is, if*

(i) For all $p \in M$, every neighbourhood of p contains a neighbourhood of p which no future pointing causal curve intersects more than once[67].

(ii) For every $p, q \in M$, the set $J_+(p) \cap J_-(q)$ is compact.

Corollary 16.8. *For any $p \in M$, the sets $J_+(p)$ and $J_-(p)$ are closed.*

[66] A simple example is given by $\mathbb{R}^{3,1}$ with a point on the lightcone removed. It is easy to see that $\partial J_+(0) \neq E_+(0)$. **Exercise**: Minkowski space with a point removed doesn't have a Cauchy surface, and isn't globally hyperbolic.

[67] This condition is known as the **strong causality condition**. In particular, it rules out closed causal curves.

Proof. Suppose that $J_+(p)$ is not closed, so that there is a point $q \in \overline{J_+(p)} \setminus J_+(p)$. Consider a point $r \in J_+(q)$. Then $q \in \overline{J_+(p) \cap J_-(r)}$. But, by the preceding proposition, $J_+(p) \cap J_-(r)$ is closed, so that $\overline{J_+(p) \cap J_-(r)} = J_+(p) \cap J_-(r)$, which implies $q \in J_+(p)$, a contradiction. □

It follows straightforwardly that $J_+(P)$ (and similarly, $J_-(P)$) is closed for any compact set P (such as τ).

We now show that the boundary of $J_+(\tau)$ doesn't contain points that lie on timelike curves emanating from τ.

Lemma 16.9. *For any $P \subset M$, $\operatorname{Int} J_+(P) = I_+(P)$.*

Proof. We first prove that $I_+(P)$ is an open set, so that $I_+(P) \subset \operatorname{Int} J_+(P)$. This is because, for any $p \in M$, the set $I_+(p)$ is an open set, since a small variation of a timelike curve (with one end fixed) gives a timelike curve. More explicitly, let $\gamma : [0, b] \to M$ be a future pointing timelike curve with $\gamma(0) = p$ and $\gamma(b) = q$. Let K be some compact, convex set in T_qM containing the origin. For each $v \in K$, we can consder the variation $\omega_v(s, \epsilon) \doteqdot \exp_{\gamma(s)}(\epsilon V(s))$ of γ, where V is the parallel translate of v along γ linearly rescaled such that $V(0) = 0$ and $V(b) = v$, and exp is the *exponential map*[68]. That is, we take each point of γ and move it a small distance ϵ along the geodesic in the direction of V. Since $g(\gamma', \gamma') < 0$, by continuity, for each $v \in K$ there is a (sufficiently small) positive ϵ_v such that for all $\epsilon \in (0, \epsilon_v)$ the curve $\gamma_{v,\epsilon}(s) = \omega_v(s, \epsilon)$ satisfies $g(\gamma'_{v,\epsilon}, \gamma'_{v,\epsilon}) < 0$. That is, for sufficiently small ϵ, the perturbed geodesic is still timelike. Since K is compact, $\epsilon_0 \doteqdot \inf_{v \in K} \epsilon_v$ is non-zero, and we find that, for every $v \in K$ and every $\epsilon \in (0, \epsilon_0)$, the curve $\gamma_{v,\epsilon}(s) = \omega_v(s, \epsilon)$ is timelike. Moreover, since $V(0) = 0$, the variation fixes the initial point p. Therefore, the set $N_q \doteqdot \{\omega_v(b, \epsilon) : v \in K,\ \epsilon < \epsilon_0\}$, which contains q, is contained in $I_+(p)$. Moreover, we can always choose K small enough that exp is a diffeomorphism. Then, as the diffeomorphic image of an open set, the set N_q is open. We have demonstrated that every $q \in I_+(p)$ has an open neighbourhood contained in $I_+(p)$, which implies that $I_+(p)$ is open. Since $p \in P$ was arbitrary, $I_+(P) = \cup_{p \in P} I_+(p)$ is a union of open sets, and is therefore itself open. Since $I_+(P)$ is an open subset of $J_+(P)$, we have $I_+(P) \subset \operatorname{Int} J_+(P)$.

To prove the opposite inclusion, consider any point $q \in \operatorname{Int} J_+(P)$. Since $\operatorname{Int} J_+(P)$ is open, there is a convex[69] neighbourhood O_q of q such that $O_q \subset \operatorname{Int} J_+(P)$. Now consider the open set $A \doteqdot I_-(O_q) \cap \operatorname{Int} J_+(P) \subset$

[68] The exponential map maps a tangent vector v to the point $\gamma_v(1)$, where γ_v is the geodesic with initial data $\gamma'(0) = v$. One should imagine wrapping the tangent space at p down onto M, with straight lines emanating from the origin mapped to geodesics emanating from p. There is a neighbourhood of 0 in T_pM, for which this map is a diffeomorphism. We call this neighbourhood a *normal neighbourhood* of p.

[69] Convex means that any geodesic joining two points of the set lies entirely in the set.

$J_+(P)$. This set is non-empty since it contains q. Moreover, we observe that $q \in I_+(A) \subset I_+(J_+(P)) \subset I_+(P)$. Therefore, $\operatorname{Int} J_+(P) \subset I_+(P)$, which completes the proof. □

Returning to the task at hand, we now have $\operatorname{Int} J_+(\tau) = I_+(\tau)$, and, since $J_+(\tau)$ is closed, $\partial J_+(\tau) = J_+(\tau) \setminus I_+(\tau)$. This proves Claim 16.6.

We now show that, if (M, g) is geodesically complete, then $\partial J_+(\tau) = E_+(\tau)$ is compact. Roughly speaking, this means that the light rays emanating outwards from τ converge on each other, so that, if there are no holes in the manifold for them to fall out of, they eventually meet, forming a 'closed light cone' which bounds the future of τ. This is a consequence for the causality of M which follows from the two main geometric assumptions: that (M, g) is null non-diverging, and that τ is trapped.

Claim 16.10. *If (M, g) is null geodesically complete, then $\partial J_+(\tau)$ is compact.*

We need to define the mean curvature, H, of τ: It is the trace (with respect to the metric induced on τ) of the *second fundamental form* of τ, which, in turn, is defined as the normal part of the connection restricted to τ. More precisely, given vector fields U, V tangent to τ, we can consider the covariant derivative $\nabla_U V$. Now, although U and V are tangent to τ, the derivative $\nabla_U V$ may not be! The second fundamental form, h, measures the lack of tangentness of $\nabla_U V$:

$$h(U, V) \doteqdot (\nabla_U V)^{\perp} ,$$

where $\perp$ denotes the orthogonal projection, giving the part of $\nabla_U V$ normal to τ. We often say that h measures the **extrinsic curvature** of τ.

Exercise 16.2.

(1) h is symmetric: $h(U, V) = h(V, U)$.

(2) At each point p of τ, h defines a tensor ($h_p \in T_p^*\tau \otimes T_p^*\tau \otimes N_p\tau$, where $N\tau$ is the orthogonal compliment of $T_p\tau$ in T_pM). That is, $h(fU, V) = h(U, fV) = fh(U, V)$ for any smooth function f.

The mean curvature (at $p \in \tau$) is then the normal vector

$$H = \operatorname{tr}_\tau h = h(e_1, e_1) + h(e_2, e_2) ,$$

where $\{e_1, e_2\}$ is an orthonormal basis for $T_p\tau$.

Definition 16.11. *Let P be a spacelike submanifold of (M, g) and $n : [0, b) \to M$ a null geodesic normal to P (with $n(0) \in P$). We say that n has a **focal point** at $a \in (0, b)$ if the **index form**,*

$$I(X, Y) \doteqdot \int_0^c \left[\langle \nabla_s X, \nabla_s Y\rangle - \operatorname{Rm}(X, n', Y, n')\right] ds - \langle n'(0), h(X(0), Y(0)) ,$$

is positive definite (with respect to vector fields $X(s), Y(s)$ along n which are not tangent to n) for all $c < a$, but not for $c = a$.

Remark 16.12. *For more about focal points (including a more natural definition), see* [**5**]. *Although this is not apparent from the definition, a focal point is, loosely speaking, a point at which infinitesimally nearby normal geodesics intersect. For example, the centre is a focal point of the sphere.*

The definition given here comes from analysing critical points of the energy functional[70] *amongst curves emanating normally from P. In fact, $I(V,V)$ is precisely the second variation of E, at n, in the direction of the vector field V; so our definition says that focal points are those at which n is ceases to be locally energy minimising (amongst nearby curves normal to P).*

Lemma 16.13. *Assuming null geodesic completeness, every (maximally extended) future pointing null geodesic $n : [0, \infty) \to M$ emanating from τ orthogonally has a focal point. Moreover, the focal point is reached before the geodesic parameter reaches a value of $1/k_n$, where*[71] *$k_n \doteqdot g(H, n'(0))$.*

Proof. Given a future pointing null geodesic $n : [0, \infty) \to M$, we will show that the index form is not positive definite on $[0, 1/k]$, where $k \doteqdot g(H, n'(0))$.

Consider a basis $\{e_1, e_2, \nu_1, \nu_2\}$ of $T_{n(0)}M$ satisfying

$$\begin{aligned} g(e_i, e_j) &= \delta_{ij} \\ g(\nu_i, e_j) = g(\nu_i, \nu_i) &= 0 \\ g(\nu_1, \nu_2) &= 1\,. \end{aligned}$$

We can further choose $\nu_1 = n'(0)$. Now parallel transport this basis along n to obtain a frame $\{E_1, E_2, N_1, N_2\}$ along n satisfying $\nabla_s E_i = \nabla_s N_i = 0$ for each $i = 1, 2$. Since n is geodesic, we have $N_1 = n'$. Now consider the fields $\{fE_i\}_{i=1}^2$, where $f(s) \doteqdot 1 - ks$. We have

$$\begin{aligned} I(fE_i, fE_i) &= \int_0^{\frac{1}{k}} \Big[f'^2 - f^2 \operatorname{Rm}(E_i, n', E_i, n') \Big] ds - \langle n'(0), h(e_i, e_i) \rangle \\ &= k - \int_0^{\frac{1}{k}} f^2 \operatorname{Rm}(E_i, n', E_i, n') ds - \langle n'(0), h(e_i, e_i) \rangle\,. \end{aligned}$$

Adding the two equations yields

$$I(fE_1, fE_1) + I(fE_2, fE_2) = -\int_0^{\frac{1}{k}} f^2 \sum_{i=1}^{2} \operatorname{Rm}(E_i, n', E_i, n') ds\,.$$

[70] See the remark following Definition 16.1.

[71] Note that $k_n > 0$ since H is past pointing timelike.

But $\sum_{i=1}^{2} R(E_i, n', n', E_i) = \mathrm{Rc}(n', n')$, since the null components vanish. (This is because the null components give the cross terms $R(N_1, n', n', N_2)$ and $R(N_2, n', n', N_1)$, which vanish because $N_1 = n'$). Since $\mathrm{Rc}(n', n') \geq 0$, we obtain

$$I(fE_1, fE_1) + I(fE_2, fE_2) \leq 0\,.$$

The proof is completed by noting that, since $\{E_i, N_i\}_{i=1}^{2}$ forms a basis along n, and $N_1 = n'$, neither E_1 nor E_2 can be tangent to n. □

The next proposition says that points on n beyond a focal point cannot lie on $\partial J_+(\tau)$.

Proposition 16.14. *Let $n : [0, b] \to M$ be a null geodesic through $q = n(b)$, emanating from τ orthogonally. If n has a focal point at $a < b$, then there is a timelike geodesic from τ to q arbitrarily close to n. If instead q lies on n at or before its first focal point, then there is no timelike curve joining τ to q.*

Proof. This is a variation argument similar in spirit to Propositions 16.4 and 16.5. See [**5**]. □

It follows that, for any null geodesic n generating $\partial J_+(\tau)$, any point on n beyond its focal point lies in $I_+(\tau) = \mathrm{Int}\, J_+(\tau)$, while every point before or at a focal point lies on the boundary. We conclude that $\partial J_+(\tau)$ is a closed set. At each point p of τ we may choose a local basis $\{\mu, \nu\}$ for the normal bundle of τ consisting of a pair of future pointing null vectors. Let T denote the set of the null normals composing these local bases. Then T is compact, since it is a double cover of τ. In particular, $k \doteqdot \inf_{\nu \in T} g(H, \nu)$ is attained by some $\nu \in T$, and, since H is past pointing timelike, k is positive. By Lemma 16.13, each geodesic emanating from τ with initial data $\nu \in T$ has a focal point at or before its parameter reaches the value k. Therefore $\partial J_+(\tau)$ is a subset of the set

$$\begin{aligned} E &\doteqdot \{n(s) \in M : n \text{ is a geodesic with } n'(0) \in T,\ s \in [0, k]\} \\ &= \exp\{s\nu : s \in [0, 1], \nu \in T\}\,, \end{aligned}$$

where exp is the exponential map. Therefore, E, as the continuous image of a compact set, is compact. It follows that $\partial J_+(\tau)$, a closed subset of a compact set, is compact. This proves Claim 16.10.

We will now show that $\partial J_+(\tau)$ fits injectively inside Σ.

Claim 16.15. *There is a continuous injective map $\rho : \partial J_+(\tau) \to \Sigma$.*

Proof. Since (M, g) is time oriented, it has a timelike vector field, j. Each of the integral curves of j is timelike, and therefore intersects the Cauchy

surface Σ exactly once. Therefore the map $\rho : \partial J_+(\tau) \to \Sigma$ defined by tracing a point $p \in \partial J_+(\tau)$ along the integral curve of j through p until it meets Σ is injective. Since the flow of j is continuous, so is ρ. □

This proves Claim 16.15. We will use it to derive a contradiction to the geodesic completeness of (M, g). We will need the following lemma.

Lemma 16.16. *$\partial J_+(\tau)$ is a 3–dimensional topological manifold (without boundary).*

Sketch proof. Consider any point $p \in \partial J_+(\tau)$. Fix an orthonormal basis $\{e_i\}_{i=0}^3$ for T_pM (such that e_0 is timelike). We can define *normal coordinates* on a neighbourhood U_p of p by setting $y^i(q) = v^i$, where $q = \exp_p(v^i e_i)$.

Now, for sufficiently small $u = (u_1, u_2, u_3) \in \textit{proposition}^3$, we can consider the unique timelike curve in U_p defined by $\gamma_u(s) = \exp_p(se_0 + u_1e_1 + u_2e_2 + u_3e_3)$. For small enough s, $\gamma_u(s)$ lies outside of $J_+(\tau)$, whereas for large enough s, $\gamma_u(s)$ lies inside $J_+(\tau)$. Therefore $\gamma_u(s)$ intersects $\partial J_+(\tau)$. Moreover, γ_u can only intersect $\partial J_+(\tau)$ once, otherwise there would be a timelike curve joining two of its points, which leads easily to a contradiction. This gives us a unique triple of 'coordinates', $u^i(p) = u_i$, for each point of $\partial J_+(\tau) \cap U_p$. It now suffices to show that the y^0 coordinate is a Lipschitz function of the u^i coordinates, which follows from the lack of timelike separation of points of $\partial J_+(\tau)$ (see [**5**, Corollary 14.27], [**4**, Proposition 6.3.1] and [**7**, Proposition 2.16]). □

By Claim 16.15, we obtain the following dichotomy: the map ρ is either onto, or its image has a boundary. Since $\partial J_+(\tau)$ is compact, and Σ is non-compact[72], the first option cannot hold. On the other hand, by Lemma 16.16, neither can the second.

We have arrived at a contradiction, demonstrating the falseness of our assumption that (M, g) is null geodesically complete, confirming the statement of the theorem. □

[72]In fact, we observe that only a single integral curve of j not intersecting $\partial J_+(\tau)$ suffices, justifying the claim in the seond remark following the statement of the theorem.

Bibliography

[1] Benn, I. M., and Tucker, R. W. *An introduction to spinors and geometry with applications in physics*. Adam Hilger, Ltd., Bristol, 1989. Reprint of the 1987 original.

[2] Bondi, H. *Relativity and ccommon sense*, vol. 31 of *Educational Books*. Heinemann, New York, 1964.

[3] Carroll, S. *Spacetime and geometry*. Addison Wesley, San Francisco, CA, 2004. An introduction to general relativity.

[4] Hawking, S. W., and Ellis, G. F. R. *The large scale structure of space-time*. Cambridge University Press, London-New York, 1973. Cambridge Monographs on Mathematical Physics, No. 1.

[5] O'Neill, B. *Semi-Riemannian geometry*, vol. 103 of *Pure and Applied Mathematics*. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York, 1983. With applications to relativity.

[6] Rowe, E. G. P. *Geometrical physics in Minkowski spacetime*. Springer Monographs in Mathematics. Springer-Verlag London, Ltd., London, 2001. With a foreword by Wojtek J. Zakrzewski.

[7] Senovilla, J. M. M. Singularity theorems and their consequences. *Gen. Relativity Gravitation 30*, 5 (1998), 701–848.

[8] Wald, R. M. *General relativity*. University of Chicago Press, Chicago, IL, 1984.