GENERAL RELATIVITY

Benjamin Crowell
Fullerton College

LibreTexts™

Fullerton College

General Relativity

Benjamin Crowell

TABLE OF CONTENTS

Licensing

A detailed breakdown of this resource's licensing can be found in **Back Matter/Detailed Licensing**.

CHAPTER OVERVIEW

1: Geometric Theory of Spacetime

"I always get a slight brain-shiver, now [that] space and time appear conglomerated together in a gray, miserable chaos." – Sommerfeld

---

1.1: Time and Causality

Updating Plato's allegory of the cave, imagine two super-intelligent twins, Alice and Betty. They're raised entirely by a robotic tutor on a sealed space station, with no access to the outside world. The robot, in accord with the latest fad in education, is programmed to encourage them to build up a picture of all the laws of physics based on their own experiments, without a textbook to tell them the right answers. Putting yourself in the twins' shoes, imagine giving up all your preconceived ideas about space and time, which may turn out according to relativity to be completely wrong, or perhaps only approximations that are valid under certain circumstances.

Causality is one thing the twins will notice. Certain events result in other events, forming a network of cause and effect. One general rule they infer from their observations is that there is an unambiguously defined notion of *betweenness*: if Alice observes that event 1 causes event 2, and then 2 causes 3, Betty always agrees that 2 lies between 1 and 3 in the chain of causality. They find that this agreement holds regardless of whether one twin is standing on her head (i.e., it's invariant under rotation), and regardless of whether one twin is sitting on the couch while the other is zooming around the living room in circles on her nuclear fusion scooter (i.e., it's also invariant with respect to different states of motion).

You may have heard that relativity is a theory that can be interpreted using non-Euclidean geometry. The invariance of betweenness is a basic geometrical property that is shared by both Euclidean and non-Euclidean geometry. We say that they are both *ordered* geometries. With this geometrical interpretation in mind, it will be useful to think of events not as actual notable occurrences but merely as an ambient sprinkling of *points* at which things *could* happen. For example, if Alice and Betty are eating dinner, Alice could choose to throw her mashed potatoes at Betty. Even if she refrains, there was the potential for a causal linkage between her dinner and Betty's forehead.

Betweenness is very weak. Alice and Betty may also make a number of conjectures that would say much more about causality. For example:

  i. that the universe's entire network of causality is connected, rather than being broken up into separate parts;
 ii. that the events are globally ordered, so that for *any* two events 1 and 2, either 1 could cause 2 or 2 could cause 1, but not both;
iii. not only are the events ordered, but the ordering can be modeled by sorting the events out along a line, the time axis, and assigning a number tt, time, to each event.

To see what these conjectures would entail, let's discuss a few examples that may draw on knowledge from outside Alice and Betty's experiences.

- According to the Big Bang theory, it seems likely that the network is connected, since all events would presumably connect back to the Big Bang. On the other hand, if (i) were false we might have no way of finding out, because the lack of causal connections would make it impossible for us to detect the existence of the other universes represented by the other parts disconnected from our own universe.
- If we had a time machine[1], we could violate (ii), but this brings up paradoxes, like the possibility of killing one's own grandmother when she was a baby, and in any case nobody knows how to build a time machine.
- There are nevertheless strong reasons for believing that (ii) is false. For example, if we drop Alice into one black hole, and Betty into another, they will never be able to communicate again, and therefore there is no way to have any cause and effect relationship between Alice's events and Betty's.[2]

> **closed timelike curve (CTC)**
>
> [1] The possibility of having time come back again to the same point is often referred to by physicists as a *closed timelike curve (CTC)*. Kip Thorne, in his popularization *Black Holes and Time Warps*, recalls experiencing some anxiety after publishing a paper with "Time Machines" in the title, and later being embarrassed when a later paper on the topic was picked up by the National Enquirer with the headline PHYSICISTS PROVE TIME MACHINES EXIST. "CTC" is safer because nobody but physicists know what it means.

Since (iii) implies (ii), we suspect that (iii) is false as well. But Alice and Betty build clocks, and these clocks are remarkably successful at describing cause-and-effect relationships within the confines of the quarters in which they've lived their lives: events with higher clock readings never cause events with lower clock readings. They announce to their robot tutor that they've discovered a universal thing called time, which explains all causal relationships, and which their experiments show flows at the same rate everywhere within their quarters.

"Ah," the tutor sighs, his metallic voice trailing off.

"I know that `ah', Tutorbot," Betty says. "Come on, can't you just tell us what we did wrong?"

"You know that my pedagogical programming doesn't allow that."

"Oh, sometimes I just want to strangle whoever came up with those stupid educational theories," Alice says.

The twins go on strike, protesting that the time theory works perfectly in every experiment they've been able to imagine. Tutorbot gets on the commlink with his masters and has a long, inaudible argument, which, judging from the hand gestures, the twins imagine to be quite heated. He announces that he's gotten approval for a field trip for one of the twins, on the condition that she remain in a sealed environment the whole time so as to maintain the conditions of the educational experiment.

"Who gets to go?" Alice asks.

"Betty," Tutorbot replies, "because of the mashed potatoes."

"But I refrained!" Alice says, stamping her foot.

"Only one time out of the last six that I served them."

The next day, Betty, smiling smugly, climbs aboard the sealed spaceship carrying a duffel bag filled with a large collection of clocks for the trip. Each clock has a duplicate left behind with Alice. The clock design that they're proudest of consists of a tube with two mirrors at the ends. A flash of light bounces back and forth between the ends, with each round trip counting as one "tick," one unit of time. The twins are convinced that this one will run at a constant rate no matter what, since it has no moving parts that could be affected by the vibrations and accelerations of the journey.

Betty's field trip is dull. She doesn't get to see any of the outside world. In fact, the only way she can tell she's not still at home is that she sometimes feels strong sensations of acceleration. (She's grown up in zero gravity, so the pressing sensation is novel to her.) She's out of communication with Alice, and all she has to do during the long voyage is to tend to her clocks. As a crude check, she verifies that the light clock seems to be running at its normal rate, judged against her own pulse. The pendulum clock gets out of synch with the light clock during the accelerations, but that doesn't surprise her, because it's a mechanical clock with moving parts. All of the nonmechanical clocks seem to agree quite well. She gets hungry for breakfast, lunch, and dinner at the usual times.

When Betty gets home, Alice asks, "Well?"

"Great trip, too bad you couldn't come. I met some cute boys, went out dancing, ..."

"You did not. What about the clocks?"

"They all checked out fine. See, Tutorbot? The time theory still holds up."

"That was an anticlimax," Alice says. "I'm going back to bed now."

"Bed?" Betty exclaims. "It's three in the afternoon."

The twins now discover that although all of Alice's clocks agree among themselves, and similarly for all of Betty's (except for the ones that were obviously disrupted by mechanical stresses), Alice's and Betty's clocks disagree with one another. A week has passed for Alice, but only a couple of days for Betty.

1.2: Experimental Tests of the Nature of Time

The Hafele-Keating Experiment

In 1971, J.C. Hafele and R.E. Keating of the U.S. Naval Observatory brought atomic clocks aboard commercial airliners and went around the world, once from east to west and once from west to east.[1] (The clocks had their own tickets, and occupied their own seats.) As in the parable of Alice and Betty, Hafele and Keating observed that there was a discrepancy between the times measured by the traveling clocks and the times measured by similar clocks that stayed at the lab in Washington. The result was that the east-going clock lost an amount of time $\Delta t_E = -59 \pm 10 \; ns$, while the west going one gained $\Delta t_W = +273 \pm 7 \; ns$. This establishes that time is not universal and absolute.

> **Note**
>
> Hafele and Keating, Science, 177 (1972), 168
>
>
>
> Figure 1.2.1 - The clock took up two seats, and two tickets were bought for it under the name of "Mr. Clock."

Nevertheless, causality was preserved. The nanosecond-scale effects observed were small compared to the three-day lengths of the plane trips. There was no opportunity for paradoxical situations such as, for example, a scenario in which the east-going experimenter arrived back in Washington before he left and then proceeded to convince himself not to take the trip.

Hafele and Keating were testing specific quantitative predictions of relativity, and they verified them to within their experiment's error bars. At this point in the book, we aren't in possession of enough relativity to be able to make such calculations, but, like Alice and Betty, we can inspect the empirical results for clues as to how time works.

The opposite signs of the two results suggests that the rate at which time flows depends on the motion of the observer. The east-going clock was moving in the same direction as the earth's rotation, so its velocity relative to the earth's center was greater than that of the ones that remained in Washington, while the west-going clock's velocity was correspondingly reduced.[2] The signs of the $\Delta t's$ show that the moving clocks were slower.

> **Note**
>
> These differences in velocity are not simply something that can be eliminated by choosing a different frame of reference, because the clocks' motion isn't in a straight line. The clocks back in Washington, for example, have a certain acceleration toward the earth's axis, which is different from the accelerations experienced by the traveling clocks.

On the other hand, the asymmetry of the results, with $|\Delta t_E| \neq |\Delta t_w|$, implies that there was a second effect involved, simply due to the planes' being up in the air. Relativity predicts that time's rate of flow also changes with height in a gravitational field. The deeper reasons for such an effect are given in section 1.5.

Although Hafele and Keating's measurements were on the ragged edge of the state of the art in 1971, technology has now progressed to the point where such effects have everyday consequences. The satellites of the Global Positioning System (GPS) orbit at a speed of $1.9 \times 10^3 m/s$, an order of magnitude faster than a commercial jet. Their altitude of 20,000 km is also much greater than that of an aircraft. For both these reasons, the relativistic effect on time is stronger than in the Hafele-Keating experiment. The atomic clocks aboard the satellites are tuned to a frequency of 10.22999999543 MHz, which is perceived on the ground as 10.23 MHz. (This frequency shift will be calculated in example 11).

Muons

Although the Hafele-Keating experiment is impressively direct, it was not the first verification of relativistic effects on time, it did not completely separate the kinematic and gravitational effects, and the effect was small. An early experiment demonstrating a large and purely kinematic effect was performed in 1941 by Rossi and Hall, who detected cosmic-ray muons at the summit and base of Mount Washington in New Hampshire. The muon has a mean lifetime of 2.2 μs, and the time of flight between the top and bottom of the mountain (about 2 km for muons arriving along a vertical path) at nearly the speed of light was about 7 μs, so in the absence of relativistic effects, the flux at the bottom of the mountain should have been smaller than the flux at the top by about an order of magnitude. The observed ratio was much smaller, indicating that the "clock" constituted by nuclear decay processes was dramatically slowed down by the motion of the muons.

Gravitational red-shifts

The first experiment that isolated the gravitational effect on time was a 1925 measurement by W.S. Adams of the spectrum of light emitted from the surface of the white dwarf star Sirius B. The gravitational field at the surface of Sirius B is $4 \times 10^5 g$, and the gravitational potential is about 3,000 times greater than at the Earth's surface. The emission lines of hydrogen were red-shifted, i.e., reduced in frequency, and this effect was interpreted as a slowing of time at the surface of Sirius relative to the surface of the Earth. Historically, the mass and radius of Sirius were not known with better than order of magnitude precision in 1925, so this observation did not constitute a good quantitative test.

The first such experiment to be carried out under controlled conditions, by Pound and Rebka in 1959, is analyzed quantitatively in example 7.

The first high-precision experiment of this kind was Gravity Probe A, a 1976 experiment[3] in which a space probe was launched vertically from Wallops Island, Virginia, at less than escape velocity, to an altitude of 10,000 km, after which it fell back to earth and crashed down in the Atlantic Ocean. The probe carried a hydrogen maser clock which was used to control the frequency of a radio signal. The radio signal was received on the ground, the nonrelativistic Doppler shift was subtracted out, and the residual blueshift was interpreted as the gravitational effect effect on time, matching the relativistic prediction to an accuracy of 0.01%.

> **Note**
>
> Vessot at al., Physical Review Letters 45 (1980) 2081
>
>
>
> Figure 1.2.2 - Gravity Probe A.
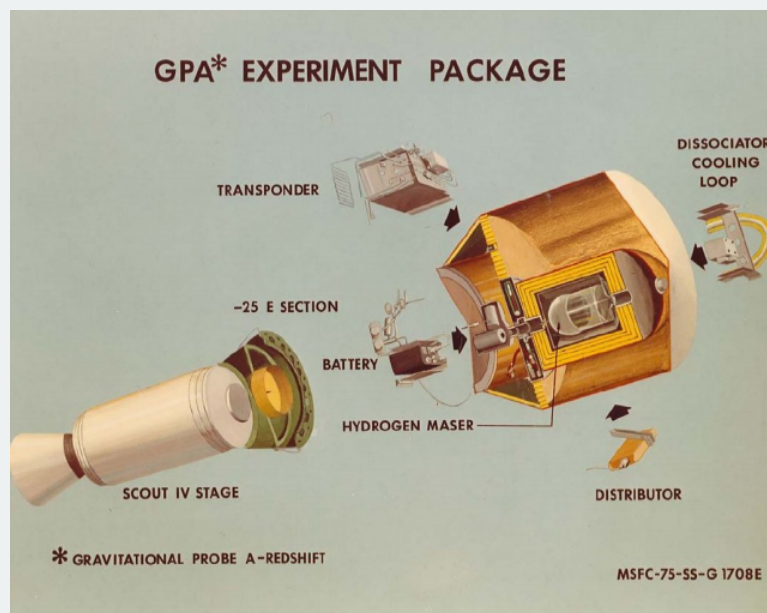
1.3: Non-simultaneity and Maximum Speed of Cause and Effect

We've seen that time flows at different rates for different observers. Suppose that Alice and Betty repeat their Hafele-Keating-style experiment, but this time they are allowed to communicate during the trip. Once Betty's ship completes its initial acceleration away from Betty, she cruises at constant speed, and each girl has her own equally valid inertial frame of reference. Each twin considers herself to be at rest, and says that the other is the one who is moving. Each one says that the other's clock is the one that is slow. If they could pull out their phones and communicate instantaneously, with no time lag for the propagation of the signals, they could resolve the controversy. Alice could ask Betty, "What time does your clock read right *now*?" and get an immediate answer back.

By the symmetry of their frames of reference, however, it seems that Alice and Betty should *not* be able to resolve the controversy *during* Betty's trip. If they could, then they could release two radar beacons that would permanently establish two inertial frames of reference, A and B, such that time flowed, say, more slowly in B than in A. This would violate the principle that motion is relative, and that all inertial frames of reference are equally valid. The best that they can do is to compare clocks once Betty returns, and verify that the net result of the trip was to make Betty's clock run more slowly *on the average*.

Alice and Betty can never satisfy their curiosity about exactly when during Betty's voyage the discrepancies accumulated or at what rate. This is information that they can never obtain, but they could obtain it if they had a system for communicating instantaneously. We conclude that instantaneous communication is impossible. There must be some maximum speed at which signals can propagate — or, more generally, a maximum speed at which cause and effect can propagate — and this speed must for example be greater than or equal to the speed at which radio waves propagate. It is also evident from these considerations that simultaneity itself cannot be a meaningful concept in relativity.

1.4: Ordered Geometry

Let's try to put what we've learned into a general geometrical context.

Euclid's familiar geometry of two-dimensional space has the following axioms,[6] which are expressed in terms of operations that can be carried out with a compass and unmarked straightedge:

- **E1** Two points determine a line.
- **E2** Line segments can be extended.
- **E3** A unique circle can be constructed given any point as its center and any line segment as its radius.
- **E4** All right angles are equal to one another.
- **E5** *Parallel postulate:* Given a line and a point not on the line, no more than one line can be drawn through the point and parallel to the given line.[7]

> **Note**
>
> This is a form known as Playfair's axiom, rather than the version of the postulate originally given by Euclid.

The modern style in mathematics is to consider this type of axiomatic system as a self-contained sandbox, with the axioms, and any theorems proved from them, being true or false only in relation to one another. Euclid and his contemporaries, however, believed them to be empirical facts about physical reality. For example, they considered the fifth postulate to be less obvious than the first four, because in order to verify physically that two lines were parallel, one would theoretically have to extend them to an infinite distance and make sure that they never crossed. In the first 28 theorems of the *Elements*, Euclid restricts himself entirely to propositions that can be proved based on the more secure first four postulates. The more general geometry defined by omitting the parallel postulate is known as *absolute geometry*.

What kind of geometry is likely to be applicable to general relativity? We can see immediately that Euclidean geometry, or even absolute geometry, would be far too specialized. We have in mind the description of events that are points in both space and time. Confining ourselves for ease of visualization to one dimension worth of space, we can certainly construct a plane described by coordinates (t, x), but imposing Euclid's postulates on this plane results in physical nonsense. Space and time are physically distinguishable from one another. But postulates 3 and 4 describe a geometry in which distances measured along non-parallel axes are comparable, and figures may be freely rotated without affecting the truth or falsehood of statements about them; this is only appropriate for a physical description of different spacelike directions, as in an (x,y) plane whose two axes are indistinguishable.

We need to throw most of the specialized apparatus of Euclidean geometry overboard. Once we've stripped our geometry to a bare minimum, then we can go back and build up a different set of equipment that will be better suited to relativity.

The stripped-down geometry we want is called *ordered geometry*, and was developed by Moritz Pasch around 1882. As suggested by the parable of Alice and Betty, ordered geometry does not have any global, all-encompassing system of measurement. When Betty goes on her trip, she traces out a particular path through the space of events, and Alice, staying at home, traces another. Although events play out in cause-and-effect order along each of these paths, we do not expect to be able to measure times along paths A and B and have them come out the same. This is how ordered geometry works: points can be put in a definite order along any particular line, but not along different lines. Of the four primitive concepts used in Euclid's E1-E5 — point, line, circle, and angle — only the non-metrical notions of point (i.e., event) and line are relevant in ordered geometry. In a geometry without measurement, there is no concept of measuring distance (hence no compasses or circles), or of measuring angles. The notation [ABC] indicates that event B lies on a line segment joining A and C, and is strictly between them.

The axioms of ordered geometry are as follows:[8]

- **O1** Two events determine a line.
- **O2** Line segments can be extended: given A and B, there is at least one event such that [ABC] is true.
- **O3** Lines don't wrap around: if [ABC] is true, then [BCA] is false.
- **O4** Betweenness: For any three distinct events A, B, and C lying on the same line, we can determine whether or not B is between A and C (and by statement 3, this ordering is unique except for a possible over-all reversal to form [CBA]).
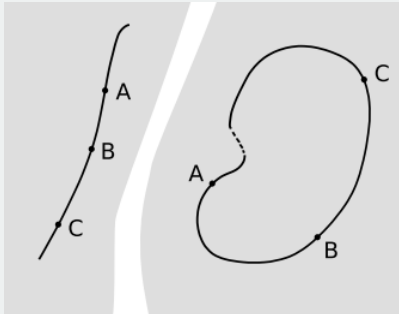
Figure 1.4.1 - Axioms O2 (left) and O3 (right).

The axioms are summarized for convenient reference. This is meant to be an informal, readable summary of the system, pitched to the same level of looseness as Euclid's E1-E5. Modern mathematicians have found that systems like these actually need quite a bit more technical machinery to be perfectly rigorous, so if you look up an axiomatization of ordered geometry, or a modern axiomatization of Euclidean geometry, you'll typically find a much more lengthy list of axioms than the ones presented here. The axioms I'm omitting take care of details like making sure that there are more than two points in the universe, and that curves can't cut through one another without intersecting. The classic, beautifully written book on these topics is H.S.M. Coxeter's *Introduction to Geometry*, which is "introductory" in the sense that it's the kind of book a college math major might use in a first upper-division course in geometry.

O1-O2 express the same ideas as Euclid's E1-E2. Not all lines in the system will correspond physically to chains of causality; we could have a line segment that describes a snapshot of a steel chain, and O3-O4 then say that the order of the links is well defined. But O3 and O4 also have clear physical significance for lines describing causality. O3 forbids time travel paradoxes, like going back in time and killing our own grandmother as a child; The figure above illustrates why a violation of O3 is referred to as a closed timelike curve. O4 says that events are guaranteed to have a well-defined cause-and-effect order only if they lie on the same line. This is completely different from the attitude expressed in Newton's famous statement: "Absolute, true and mathematical time, of itself, and from its own nature flows equably without regard to anything external ..."

If you're dismayed by the austerity of a system of geometry without any notion of measurement, you may be more appalled to learn that even a system as weak as ordered geometry makes some statements that are too strong to be completely correct as a foundation for relativity. For example, if an observer falls into a black hole, at some point he will reach a central point of infinite density, called a singularity. At this point, his chain of cause and effect terminates, violating \O2). It is also an open question whether O3's prohibition on time-loops actually holds in general relativity; this is Stephen Hawking's playfully named chronology protection conjecture. We'll also see that in general relativity O1 is almost always true, but there are exceptions.
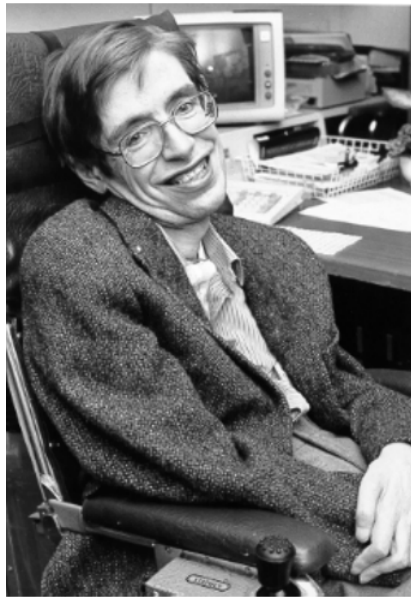
Figure 1.4.2 - Stephen Hawking (1942-).

This page titled 1.4: Ordered Geometry is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Benjamin Crowell via source content that was edited to the style and standards of the LibreTexts platform.

1.5: The Equivalence Principle (Part 1)

Proportionality of Inertial and Gravitational Mass

What physical interpretation should we give to the "lines" described in ordered geometry? Galileo described an experiment (which he may or may not have actually performed) in which he simultaneously dropped a cannonball and a musket ball from a tall tower. The two objects hit the ground simultaneously, disproving Aristotle's assertion that objects fell at a speed proportional to their weights. On a graph of spacetime with x and t axes, the curves traced by the two objects, called their *world-lines*, are identical parabolas. (The paths of the balls through $x - y - z$ space are straight, not curved.) One way of explaining this observation is that what we call "mass" is really two separate things, which happen to be equal. *Inertial mass*, which appears in Newton's $a = \frac{F}{m}$, describes how difficult it is to accelerate an object. *Gravitational mass* describes the strength with which gravity acts. The cannonball has a hundred times more gravitational mass than the musket ball, so the force of gravity acting on it is a hundred times greater. But its inertial mass is also precisely a hundred times greater, so the two effects cancel out, and it falls with the same acceleration. This is a special property of the gravitational force. Electrical forces, for example, do not behave this way. The force that an object experiences in an electric field is proportional to its charge, which is unrelated to its inertial mass, so different charges placed in the same electric field will in general have *different* motions.
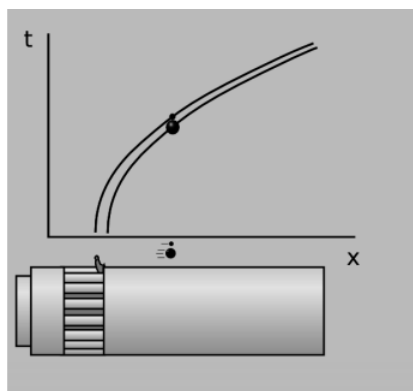
Figure 1.5.1 - The cannonball and the musketball have identical parabolic world-lines. On this type of space-time plot, space is conventionally shown on the horizontal axis, so the tower has to be depicted on its side.

Geometrical Treatment of Gravity

Einstein realized that this special property of the gravitational force made it possible to describe gravity in purely geometrical terms. We define the world-lines of small objects[9] acted on by gravity to be the lines described by the axioms of the geometry. Since we normally think of the "lines" described by Euclidean geometry and its kin as *straight* lines, this amounts to a redefinition of what it means for a line to be straight. By analogy, imagine stretching a piece of string taut across a globe, as we might do in order to plan an airplane flight or aim a directional radio antenna. The string may not appear straight as viewed from the three-dimensional Euclidean space in which the globe is embedded, but it is as straight as possible in the sense that it is the path followed by a radio wave,[10] or by an airplane pilot who keeps her wings level and her rudder straight. The world-"line" of an object acted on by nongravitational forces is not considered to be a straight "line" in the sense of O1-O4. When necessary, one eliminates this ambiguity in the overloaded term "line" by referring to the lines of O1-O4 as geodesics. The world-line of a low-mass object acted on only by gravity is one type of geodesic.[11]

Figure 1.5.2 - A piece of string held taut on a globe forms a geodesic from Mexico City to London. Although it appears curved, it is the analog of a straight line in the non-Euclidean geometry confined to the surface of the Earth. Similarly, the world-lines of figure a appear curved, but they are the analogs of straight lines in the non-Euclidean geometry used to describe gravitational fields in general relativity.

Notes

[9] The reason for the restriction to small objects is essentially gravitational radiation. The object should also be electrically neutral, and neither the object nor the surrounding spacetime should contain any exotic forms of negative energy. This is discussed in more detail later. See also problem 1.

[10] Radio waves in the HF band tend to be trapped between the ground and the ionosphere, causing them to curve over the horizon, allowing long-distance communication.

[11] For more justification of this statement, see section 9, problem 1.

We can now see the deep physical importance of statement O1, that two events determine a line. To predict the trajectory of a golf ball, we need to have some initial data. For example, we could measure event A when the ball breaks contact with the club, and event B an infinitesimal time after A.[12] This pair of observations can be thought of as fixing the ball's initial position and velocity, which should be enough to predict a unique world-line for the ball, since relativity is a deterministic theory. With this interpretation, we can also see why it is not necessarily a disaster for the theory if O1 fails sometimes. For example, event A could mark the launching of two satellites into circular orbits from the same place on the Earth, heading in opposite directions, and B could be their subsequent collision on the opposite side of the planet. Although this violates O1, it doesn't violate determinism. Determinism only requires the validity of O1 for events infinitesimally close together. Even for randomly chosen events far apart, the probability that they will violate O1 is zero.
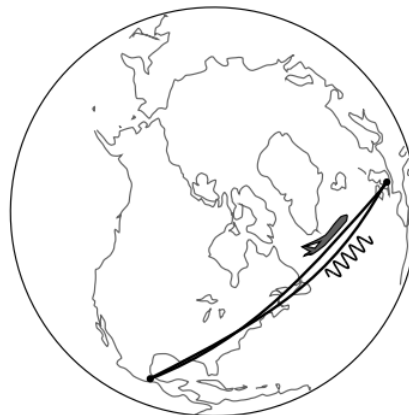
Figure 1.5.3 - If the geodesics defined by an airplane and a radio wave differ from one another, then it is not possible to treat both problems exactly using the same geometrical theory. In general relativity, this would be analogous to a violation of the equivalence principle. General relativity's validity as a purely geometrical theory of gravity requires that the equivalence principle be exactly satisfied in all cases.

> **Note**
>
> Regarding infinitesimals, see Example 3.

Eötvös Experiments

Einstein's entire system breaks down if there is any violation, no matter how small, of the proportionality between inertial and gravitational mass, and it therefore becomes very interesting to search experimentally for such a violation. For example, we might wonder whether neutrons and protons had slightly different ratios of gravitational and inertial mass, which in a Galileo-style experiment would cause a small difference between the acceleration of a lead weight, with a large neutron-to-proton ratio, and a wooden one, which consists of light elements with nearly equal numbers of neutrons and protons. The first high-precision experiments of this type were performed by Eötvös around the turn of the twentieth century, and they verified the equivalence of inertial and gravitational mass to within about one part in 10^8. These are generically referred to as Eötvös experiments.
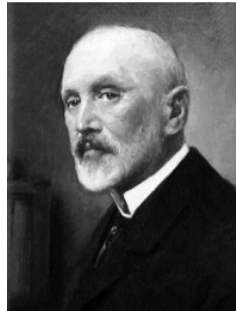
Figure 1.5.4: Lorand Eötvös (1848-1919).

Figure 1.5.5 below shows a strategy for doing Eötvös experiments that allowed a test to about one part in 10^{12}. The top panel is a simplified version. The platform is balanced, so the gravitational masses of the two objects are observed to be equal. The objects are made of different substances. If the equivalence of inertial and gravitational mass fails to hold for these two substances, then the force of gravity on each mass will not be exact proportion to its inertia, and the platform will experience a slight torque as the earth spins.
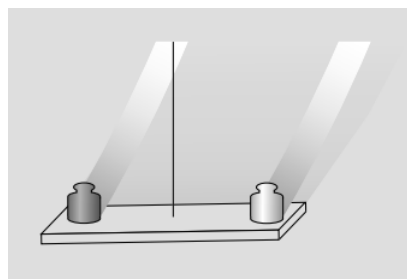
Figure_1.5.5b.png" />

***Figure* 1.5.5**: *An Eötvös experiment. Top: simplified version. Bottom: realistic version by Braginskii and Panov. (Drawing after Braginskii and Panov.)*

The bottom panel shows a more realistic drawing of an experiment by Braginskii and Panov.[13] The whole thing was encased in a tall vacuum tube, which was placed in a sealed basement whose temperature was controlled to within 0.02 °C. The total mass of the platinum and aluminum test masses, plus the tungsten wire and the balance arms, was only 4.4 g. To detect tiny motions, a laser beam was bounced off of a mirror attached to the wire. There was so little friction that the balance would have taken on the order of several years to calm down completely after being put in place; to stop these vibrations, static electrical forces were applied through the two circular plates to provide very gentle twists on the ellipsoidal mass between them.

In the 45 years since Braginskii and Panov's work, improvements have been made in more direct experimental tests of the equivalence principle, in which the test masses simply free-fall. The best earthbound experiment of this type[14] has given a bound of 10^{-9}, while a new experiment in orbit[15] has tightened this to 10^{-14}.

Equivalence of Gravitational Fields and Accelerations

One consequence of the Eötvös experiments' null results is that it is not possible to tell the difference between an acceleration and a gravitational field. At certain times during Betty's field trip, she feels herself pressed against her seat, and she interprets this as evidence that she's in a space vessel that is undergoing violent accelerations and decelerations. But it's equally possible that Tutorbot has simply arranged for her capsule to be hung from a rope and dangled into the gravitational field of a planet. Suppose that the first explanation is correct. The capsule is initially at rest in outer space, where there is no gravity. Betty can release a pencil and a lead ball in the air inside the cabin, and they will stay in place. The capsule then accelerates, and to Betty, who has adopted a frame of reference tied to its deck, ceiling and walls, it appears that the pencil and the ball fall to the deck. They are guaranteed to stay side by side until they hit the deckplates, because in fact they aren't accelerating; they simply appear to accelerate, when in reality it's the deckplates that are coming up and hitting them. But now consider the second explanation, that the capsule has been dipped into a gravitational field. The ball and the pencil will still fall side by side to the floor, because they have the same ratio of gravitational to inertial mass.

The Equivalence Principle

This leads to one way of stating a central principle of relativity known as the *equivalence principle*: Accelerations and gravitational fields are equivalent. There is no experiment that can distinguish one from the other.[16]

> **Note**
>
> This statement of the equivalence principle is summarized, along with some other forms of it to be encountered later.

To see what a radical departure this is, we need to compare with the completely different picture presented by Newtonian physics and special relativity. Newton's law of inertia states that "Every object perseveres in its state of rest, or of uniform motion in a straight line, unless it is compelled to change that state by forces impressed thereon."[17] Newton's intention here was to clearly state a contradiction of Aristotelian physics, in which objects were supposed to naturally stop moving and come to rest in the absence of a force. For Aristotle, "at rest" meant at rest relative to the Earth, which represented a special frame of reference. But if motion doesn't naturally stop of its own accord, then there is no longer any way to single out one frame of reference, such as the one tied to the Earth, as being special. An equally good frame of reference is a car driving in a straight line down the interstate at constant speed. The earth and the car both represent valid *inertial* frames of reference, in which Newton's law of inertia is valid. On the other hand, there are other, noninertial frames of reference, in which the law of inertia is violated. For example, if the car decelerates suddenly, then it appears to the people in the car as if their bodies are being jerked forward, even though there is no physical object that could be exerting any type of forward force on them. This distinction between inertial and noninertial frames of reference was carried over by Einstein into his theory of special relativity, published in 1905.

> **Note**
>
> paraphrased from a translation by Motte, 1729

But by the time he published the general theory in 1915, Einstein had realized that this distinction between inertial and noninertial frames of reference was fundamentally suspect. How do we know that a particular frame of reference is inertial? One way is to verify that its motion relative to some other inertial frame, such as the Earth's, is in a straight line and at constant speed. But how does the whole thing get started? We need to bootstrap the process with at least one frame of reference to act as our standard. We can look for a frame in which the law of inertia is valid, but now we run into another difficulty. To verify that the law of inertia holds, we have to check that an observer tied to that frame doesn't see objects accelerating for no reason. The trouble here is that by the equivalence principle, there is no way to determine whether the object is accelerating "for no reason" or because of a gravitational force. Betty, for example, cannot tell by any local measurement (i.e., any measurement carried out within the capsule) whether she is in an inertial or a noninertial frame.
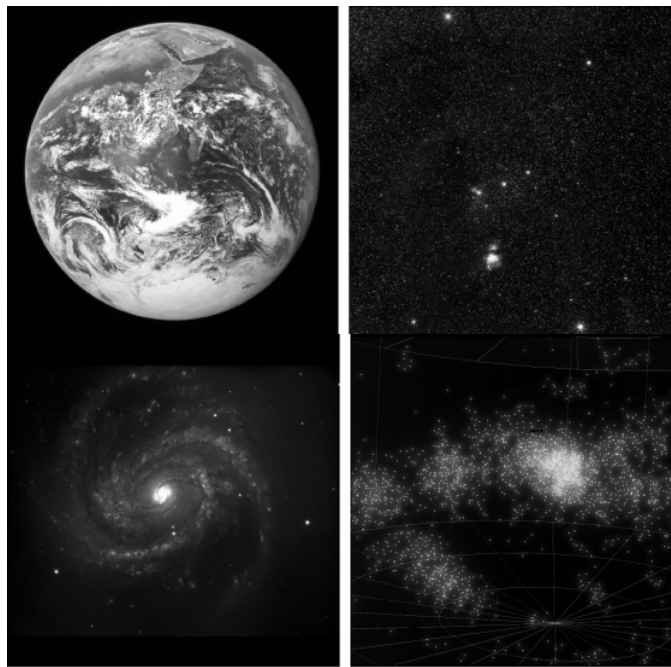
Figure 1.5.6: Wouldn't it be nice if we could define the meaning of a Newtonian inertial frame of reference? Newton makes it sound easy: to define an inertial frame, just find some object that is not accelerating because it is not being acted on by any external forces. But what object would we use? The earth? The "fixed stars?" Our galaxy? Our supercluster of galaxies? All of these are accelerating — relative to something.

We could hope to resolve the ambiguity by making non-local measurements instead. For example, if Betty had been allowed to look out a porthole, she could have tried to tell whether her capsule was accelerating relative to the stars. Even this possibility ends up not being satisfactory. The stars in our galaxy are moving in circular orbits around the galaxy. On an even larger scale, the universe is expanding in the aftermath of the Big Bang. It spent about the first half of its history decelerating due to gravitational attraction, but the expansion is now observed to be accelerating, apparently due to a poorly understood phenomenon referred to by the catch-all term "dark energy." In general, there is no distant background of physical objects in the universe that is not accelerating.

Lorentz Frames

The conclusion is that we need to abandon the entire distinction between Newton-style inertial and noninertial frames of reference. The best that we can do is to single out certain frames of reference defined by the motion of objects that are not subject to any nongravitational forces. A falling rock defines such a frame of reference. In this frame, the rock is at rest, and the ground is accelerating. The rock's world-line is a straight line of constant x = 0 and varying t. Such a free-falling frame of reference is called a Lorentz frame. The frame of reference defined by a rock sitting on a table is an inertial frame of reference according to the Newtonian view, but it is not a Lorentz frame.

In Newtonian physics, inertial frames are preferable because they make motion simple: objects with no forces acting on them move along straight world-lines. Similarly, Lorentz frames occupy a privileged position in general relativity because they make motion simple: objects move along "straight" world-lines if they have no nongravitational forces acting on them.

Example 1: The artificial horizon

The pilot of an airplane cannot always easily tell which way is up. The horizon may not be level simply because the ground has an actual slope, and in any case the horizon may not be visible if the weather is foggy. One might imagine that the problem could be solved simply by hanging a pendulum and observing which way it pointed, but by the equivalence principle the pendulum cannot tell the difference between a gravitational field and an acceleration of the aircraft relative to the ground — nor can any other accelerometer, such as the pilot's inner ear.

Figure 1.5.7

For example, when the plane is turning to the right, accelerometers will be tricked into believing that "down" is down and to the left. To get around this problem, airplanes use a device called an artificial horizon, which is essentially a gyroscope. The gyroscope has to be initialized when the plane is known to be oriented in a horizontal plane. No gyroscope is perfect, so over time it will drift. For this reason the instrument also contains an accelerometer, and the gyroscope is automatically restored to agreement with the accelerometer, with a time-constant of several minutes. If the plane is flown in circles for several minutes, the artificial horizon will be fooled into indicating that the wrong direction is vertical.

Example 2: no antigravity

This whole chain of reasoning was predicated on the null results of Eötvös experiments. In the Rocky and Bullwinkle cartoons, there is a non-Eötvösian substance called upsidasium, which falls up instead of down. Its ratio of gravitational to inertial mass is apparently negative. If such a substance could be found, it would falsify the equivalence principle. Example 10.
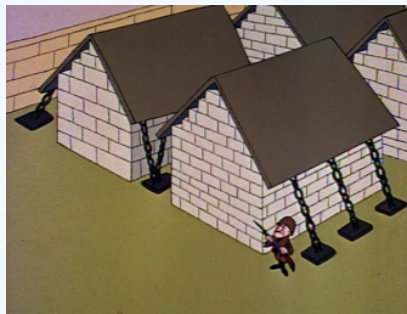
Figure 1.5.8: Bars of upsidasium are kept in special warehouses, bolted to the ground. © Jay Ward Productions, used under U.S. fair use exception to copyright law.

Operational Definition of a Lorentz Frame

We can define a Lorentz frame in operational terms using an idealized variation (Figure 1.5.9) on a device actually built by Harold Waage at Princeton as a lecture demonstration to be used by his partner in crime John Wheeler. Build a sealed chamber whose contents are isolated from all nongravitational forces. Of the four known forces of nature, the three we need to exclude are the strong nuclear force, the weak nuclear force, and the electromagnetic force. The strong nuclear force has a range of only about 1 fm (10^{-15} m), so to exclude it we merely need to make the chamber thicker than that, and also surround it with enough paraffin wax to keep out any neutrons that happen to be flying by. The weak nuclear force also has a short range, and although shielding against neutrinos is a practical impossibility, their influence on the apparatus inside will be negligible. To shield against electromagnetic forces, we surround the chamber with a Faraday cage and a solid sheet of mu-metal. Finally, we make sure that the chamber is not being touched by any surrounding matter, so that short-range residual electrical forces (sticky forces, chemical bonds, etc.) are excluded. That is, the chamber cannot be supported; it is free-falling.
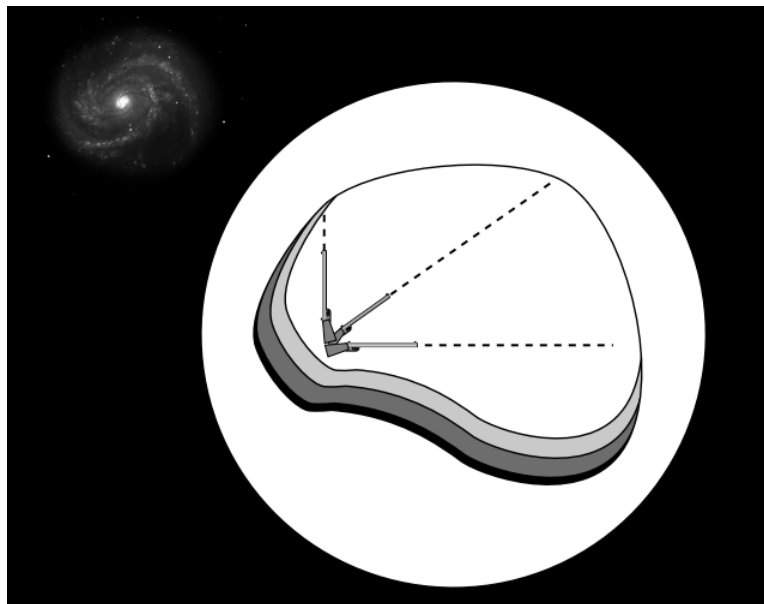
Figure 1.5.9: The spherical chamber, shown in a cutaway view, has layers of shielding to exclude all known nongravitational forces. Once the chamber has been calibrated by marking the three dashed-line trajectories under free-fall conditions, an observer inside the chamber can always tell whether she is in a Lorentz frame.

Crucially, the shielding does *not* exclude gravitational forces. There is in fact no known way of shielding against gravitational effects such as the attraction of other masses or the propagation of gravitational waves. Because the shielding is spherical, it exerts no gravitational force of its own on the apparatus inside.

Inside, an observer carries out an initial calibration by firing bullets along three Cartesian axes and tracing their paths, which she *defines* to be linear.

We've gone to elaborate lengths to show that we can really determine, without reference to any external reference frame, that the chamber is not being acted on by any nongravitational forces, so that we know it is free-falling. In addition, we also want the observer to be able to tell whether the chamber is rotating. She could look out through a porthole at the stars, but that would be missing the whole point, which is to show that without reference to any other object, we can determine whether a particular frame is a Lorentz frame. One way to do this would be to watch for precession of a gyroscope. Or, without having to resort to additional apparatus, the observer can check whether the paths traced by the bullets change when she changes the muzzle velocity. If they do, then she infers that there are velocity-dependent Coriolis forces, so she must be rotating. She can then use flywheels to get rid of the rotation, and redo the calibration.

After the initial calibration, she can always tell whether or not she is in a Lorentz frame. She simply has to fire the bullets, and see whether or not they follow the precalibrated paths. For example, she can detect that the frame has become non-Lorentzian if the chamber is rotated, allowed to rest on the ground, or accelerated by a rocket engine.

It may seem that the detailed construction of this elaborate thought-experiment does nothing more than confirm something obvious. It is worth pointing out, then, that we don't really know whether it works or not. It works in general relativity, but there are other theories of gravity, such as Brans-Dicke gravity, that are also consistent with all known observations, but in which the apparatus in Figure 1.5.8 doesn't work. Two of the assumptions made above fail in this theory: gravitational shielding effects exist, and Coriolis effects become undetectable if there is not enough other matter nearby.

Locality of Lorentz Frames

It would be convenient if we could define a single Lorentz frame that would cover the entire universe, but we can't. In Figure 1.5.9, two girls simultaneously drop down from tree branches — one in Los Angeles and one in Mumbai. The girl free-falling in Los Angeles defines a Lorentz frame, and in that frame, other objects falling nearby will also have straight world-lines. But in the LA girl's frame of reference, the girl falling in Mumbai does not have a straight world-line: she is accelerating up toward the LA girl with an acceleration of about 2g.
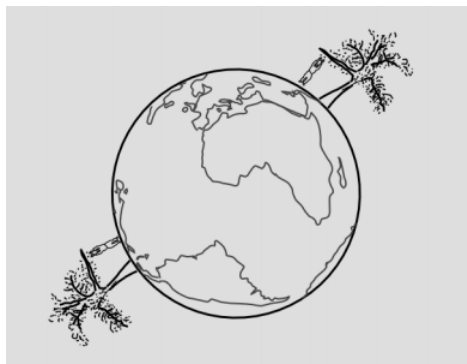
Figure 1.5.10: Two local Lorentz frames.

A second way of stating the equivalence principle is that it is always possible to define a local Lorentz frame in a particular neighborhood of spacetime.[18] It is not possible to do so on a universal basis.

> **Note**
>
> This statement of the equivalence principle is summarized, along with some other forms of it, in Appendix C.

The locality of Lorentz frames can be understood in the analogy of the string stretched across the globe. We don't notice the curvature of the Earth's surface in everyday life because the radius of curvature is thousands of kilometers. On a map of LA, we don't notice any curvature, nor do we detect it on a map of Mumbai, but it is not possible to make a flat map that includes both LA and Mumbai without seeing severe distortions.

Terminology

The meanings of words evolve over time, and since relativity is now a century old, there has been some confusing semantic drift in its nomenclature. This applies both to "inertial frame" and to "special relativity."

Early formulations of general relativity never refer to "inertial frames," "Lorentz frames," or anything else of that flavor. The very first topic in Einstein's original systematic presentation of the theory[19] is an example (Figure 1.5.10) involving two planets, the purpose of which is to convince the reader that *all* frames of reference are created equal, and that any attempt to make some of them into second-class citizens is invidious. Other treatments of general relativity from the same era follow Einstein's lead.[20] The trouble is that this example is more a statement of Einstein's aspirations for his theory than an accurate depiction of the physics that it actually implies. General relativity really does allow an unambiguous distinction to be made between Lorentz frames and non-Lorentz frames, as described earlier. Einstein's statement should have been weaker: the laws of physics (such as the Einstein field equation) are the same in all frames (Lorentz or non-Lorentz). This is different from the situation in Newtonian mechanics and special relativity, where the laws of physics take on their simplest form only in Newton-inertial frames.
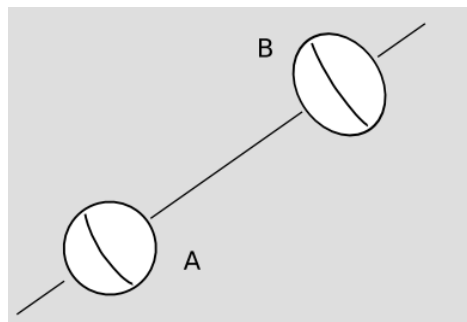
Figure 1.5.11 - One planet rotates about its axis and the other does not. As discussed in more detail later, Einstein believed that general relativity was even more radically egalitarian about frames of reference than it really is. He thought that if the planets were alone in an otherwise empty universe, there would be no way to tell which planet was really rotating and which was not, so that B's equatorial bulge would have to disappear. There would be no way to tell which planet's surface was a Lorentz frame.

> **Note**
>
> Two that I believe were relatively influential are Born's 1920 *Einstein's Theory of Relativity* and Eddington's 1924 *The Mathematical Theory of Relativity*. Born follows Einstein's "Foundation" paper slavishly. Eddington seems only to mention inertial frames in a few places where the context is Newtonian.

Because Einstein didn't want to make distinctions between frames, we ended up being saddled with inconvenient terminology for them. The least verbally awkward choice is to hijack the term "inertial," redefining it from its Newtonian meaning. We then say that the Earth's surface is not an inertial frame, in the context of general relativity, whereas in the Newtonian context it is an inertial frame to a very good approximation. This usage is fairly standard,[21] but would have made Newton confused and Einstein unhappy. If we follow this usage, then we may sometimes have to say "Newtonianinertial" or "Einstein-inertial." A more awkward, but also more precise, term is "Lorentz frame," as used in this book; this seems to be widely understood.[22]

The distinction between special and general relativity has undergone a similar shift over the decades. Einstein originally defined the distinction in terms of the admissibility of accelerated frames of reference. This, however, puts us in the absurd position of saying that special relativity, which is supposed to be a generalization of Newtonian mechanics, cannot handle accelerated frames of reference in the same way that Newtonian mechanics can. In fact both Newtonian mechanics and special relativity treat Newtonian-noninertial frames of reference in the same way: by modifying the laws of physics so that they do not take on their most simple form (e.g., violating Newton's third law), while retaining the ability to change coordinates back to a preferred frame in which the simpler laws apply. It was realized fairly early on[23] that the important distinction was between special relativity as a theory of flat spacetime, and general relativity as a theory that described gravity in terms of curved spacetime. All relativists writing since about 1950 seem to be in agreement on this more modern redefinition of the terms.[24]

In an accelerating frame, the equivalence principle tells us that measurements will come out the same as if there were a gravitational field. But if the spacetime is flat, describing it in an accelerating frame doesn't make it curved. (Curvature is a physical property of spacetime, and cannot be changed from zero to nonzero simply by a choice of coordinates.) Thus relativity allows us to have gravitational fields in flat space — but only for certain special configurations like this one. Special relativity is capable of operating just fine in this context. For example, Chung et al.[25] did a high-precision test of special relativity in 2009 using a matter interferometer in a vertical plane, specifically in order to test whether there was any violation of special relativity in a uniform gravitational field. Their experiment is interpreted purely as a test of special relativity, not general relativity.

Chiao's Paradox

The remainder of this subsection deals with the subtle question of whether and how the equivalence principle can be applied to charged particles. You may wish to skip it on a first reading. The short answer is that using the equivalence principle to make conclusions about charged particles is like the attempts by slaveholders and abolitionists in the 19th century U.S. to support their positions based on the Bible: you can probably prove whichever conclusion was the one you set out to prove.
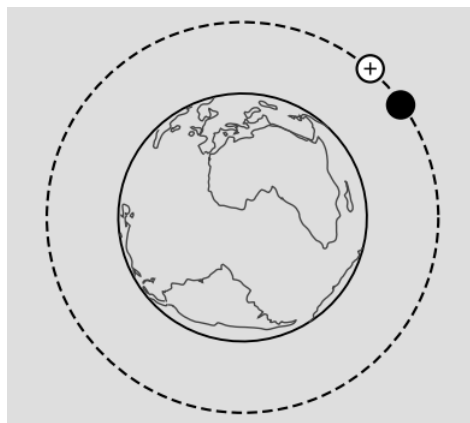
Figure 1.5.12 - Chiao's paradox: a charged particle and a neutral particle are in orbit around the earth. Will the charged particle radiate, violating the equivalence principle?

The equivalence principle is not a single, simple, mathematically well defined statement.[26] As an example of an ambiguity that is still somewhat controversial, 90 years after Einstein first proposed the principle, consider the question of whether or not it applies to charged particles. Raymond Chiao[27] proposes the following thought experiment, which I'll refer to as Chiao's paradox. Let a

neutral particle and a charged particle be set, side by side, in orbit around the earth. Assume (unrealistically) that the space around the earth has no electric or magnetic field. If the equivalence principle applies regardless of charge, then these two particles must go on orbiting amicably, side by side. But then we have a violation of conservation of energy, since the charged particle, which is accelerating, will radiate electromagnetic waves (with very low frequency and amplitude). It seems as though the particle's orbit must decay.

> **Note**
>
> A good recent discussion of this is "Theory of gravitation theories: a noprogress report," Sotiriou, Faraoni, and Liberati, http://arxiv.org/abs/0707.2748

The resolution of the paradox, as demonstrated by hairy calculations[28] is interesting because it exemplifies the local nature of the equivalence principle. When a charged particle moves through a gravitational field, in general it is possible for the particle to experience a reaction from its own electromagnetic fields. This might seem impossible, since an observer in a frame momentarily at rest with respect to the particle sees the radiation fly off in all directions at the speed of light. But there are in fact several different mechanisms by which a charged particle can be reunited with its long-lost electromagnetic offspring. An example (not directly related to Chiao's scenario) is the following.

> **Note**
>
> The first detailed calculation appears to have been by Cécile and Bryce DeWitt, "Falling Charges," Physics 1 (1964) 3. This paper is unfortunately very difficult to obtain now. A more recent treatment by Grøn and Næss is accessible at arxiv.org/abs/0806.0464v1. A full exposition of the techniques is given by Poisson, "The Motion of Point Particles in Curved Spacetime," www.livingreviews.org/lrr-2004-6.

Bring a laser very close to a black hole, but not so close that it has strayed inside the event horizon, which is the spherical point of no return from within which nothing can escape. Example 14 gives a plausibility argument based on Newtonian physics that the radius[29] of the event horizon should be something like $r_H = \frac{GM}{c^2}$, and section 6.3 derives the relativistically correct factor of 2 in front, so that $r_H = \frac{2GM}{c^2}$. It turns out that at r $= \left(\frac{3}{2}\right) r_H$, a ray of light can have a circular orbit around the black hole. Since this is greater than r_H, we can, at least in theory, hold the laser stationary at this value of r using a powerful rocket engine. If we point the laser in the azimuthal direction, its own beam will come back and hit it.

> **Note**
>
> Because relativity describes gravitational fields in terms of curvature of spacetime, the Euclidean relationship between the radius and circumference of a circle fails here. The r coordinate should be understood here not as the radius measured from the center but as the circumference divided by 2π.

Since matter can experience a back-reaction from its own electromagnetic radiation, it becomes plausible how the paradox can be resolved. The equivalence principle holds *locally*, i.e., within a small patch of space and time. If Chiao's charged and neutral particle are released side by side, then they will obey the equivalence principle for at least a certain amount of time — and "for at least a certain amount of time" is all we should expect, since the principle is local. But after a while, the charged particle will start to experience a back-reaction from its own electromagnetic fields, and this causes its orbit to decay, satisfying conservation of energy. Since Chiao's particles are orbiting the earth, and the earth is not a black hole, the mechanism clearly can't be as simple as the one described above, but Grøn and Næss show that there are similar mechanisms that can apply here, e.g., scattering of light waves by the nonuniform gravitational field.

It is worth keeping in mind the DeWitts' caution that "The questions answered by this investigation are of conceptual interest only, since the forces involved are far too small to be detected experimentally" (see problem 8).

References

[13] V.B. Braginskii and V.I. Panov, Soviet Physics JETP 34, 463 (1972).

[14] Carusotto et al., "Limits on the violation of g-universality with a Galileotype experiment," Phys Lett A183 (1993) 355. Freely available online at researchgate.net.

[15] Touboul et al., "The MICROSCOPE mission: first results of a space test of the Equivalence Principle," arxiv.org/abs/1712.01176

[19] Einstein, "The Foundation of the General Theory of Relativity," 1916. An excerpt is given later.

[21] Misner, Thorne, and Wheeler, *Gravitation*, 1973, p. 18

[22] *ibid*, p. 19

[23] Eddington, *op. cit.*

[24] Misner, Thorne, and Wheeler, *op. cit.*, pp.163-164. Penrose, *The Road to Reality*, 2004, p. 422. Taylor and Wheeler, *Spacetime Physics*, 1992, p. 132. Schutz, *A First Course in General Relativity*, 2009, pp. 3, 141. Hobson, *General Relativity: An Introduction for Physicists*, 2005, sec. 1.14.

[25] arxiv.org/abs/0905.1929

[27] arxiv.org/abs/quant-ph/0601193v7

1.6: The Equivalence Principle (Part 2)

Gravitational Red-shifts

Earlier, we saw experimental evidence that the rate of flow of time changes with height in a gravitational field. We can now see that this is required by the equivalence principle.

By the equivalence principle, there is no way to tell the difference between experimental results obtained in an accelerating laboratory and those found in a laboratory immersed in a gravitational field.[30] In a laboratory accelerating upward, a photon emitted from the floor and would be Doppler-shifted toward lower frequencies when observed at the ceiling, because of the change in the receiver's velocity during the photon's time of flight. The effect is given by

$$\frac{\Delta E}{E} = \frac{\Delta f}{f} = \frac{ay}{c^2} \tag{1.6.1}$$

where a is the lab's acceleration, y is the height from floor to ceiling, and c is the speed of light.

> **Note**
>
> Problem 4 verifies, in one specific example, that this way of stating the equivalence principle is implied by the one stated earlier.

> **Exercise 1.6.1**
>
> Self-check: Verify this statement.
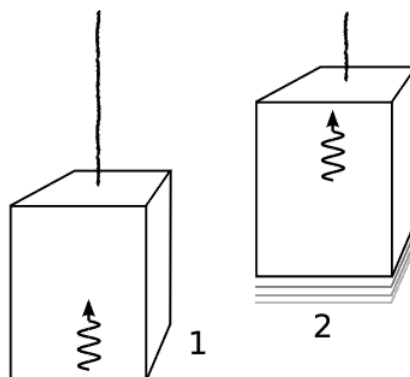
Figure 1.6.13 - 1. A photon is emitted upward from the floor of the elevator. The elevator accelerates upward. 2. By the time the photon is detected at the ceiling, the elevator has changed its velocity, so the photon is detected with a Doppler shift.

By the equivalence principle, we find that when such an experiment is done in a gravitational field g, there should be a gravitational effect on the energy of a photon equal to $\frac{\Delta E}{E} = \frac{gy}{c^2}$. Since the quantity gy is the gravitational potential (gravitational energy per unit mass), the photon's fractional loss of energy is the same as the (Newtonian) loss of energy experienced by a material object of mass m and initial kinetic energy mc^2.

The interpretation is as follows. Classical electromagnetism requires that electromagnetic waves have inertia. For example, if a plane wave strikes an ohmic surface, as in Figure 1.5.14, the wave's electric field excites oscillating currents in the surface. These currents then experience a magnetic force from the wave's magnetic field, and application of the right-hand rule shows that the resulting force is in the direction of propagation of the wave. Thus the light wave acts as if it has momentum. The equivalence principle says that whatever has inertia must also participate in gravitational interactions. Therefore light waves must have weight, and must lose energy when they rise through a gravitational field.
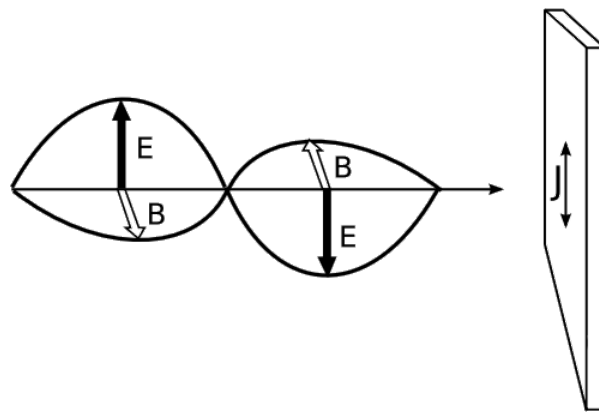
Figure 1.5.15. See Nichols and Hull, "The pressure due to radiation," Phys. Rev. (Series I) 17 (1903) 26.

Exercise 1.6.2

Self-check: Verify the application of the right-hand rule described above.
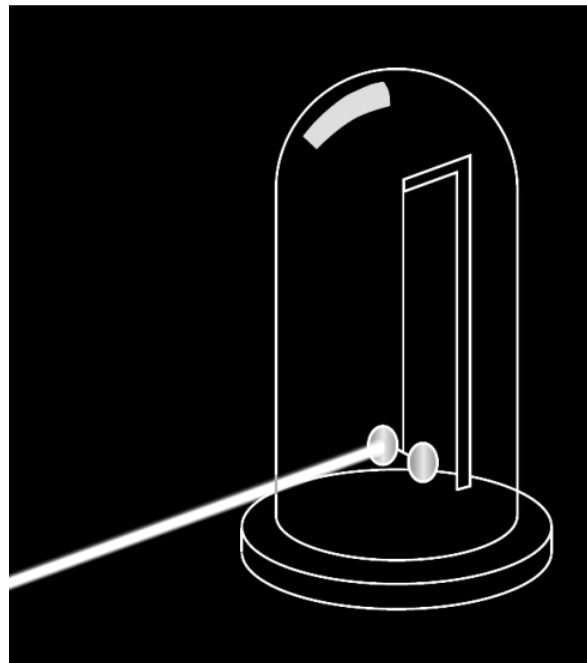
Figure 1.6.15 - A simplified drawing of the 1903 experiment by Nichols and Hull that verified the predicted momentum of light waves. Two circular mirrors were hung from a fine quartz fiber, inside an evacuated bell jar. A 150 mW beam of light was shone on one of the mirrors for 6 s, producing a tiny rotation, which was measurable by an optical lever (not shown). The force was within 0.6% of the theoretically predicted value of 0.001 μN. For comparison, a short clipping of a single human hair weighs $\sim 1\ \mu$N.

Further interpretation:

- The quantity mc^2 is famous, even among people who don't know what m and c stand for. This is the first hint of where it comes from. The full story is given in section 4.2.
- The relation p = $\frac{E}{c}$ between the energy and momentum of a light wave follows directly from Maxwell's equations, by the argument above; however, we will see in section 4.2 that according to relativity this relation must hold for any massless particle
- What we have found agrees with Niels Bohr's correspondence principle, which states that when a new physical theory, such as relativity, replaces an older one, such as Newtonian physics, the new theory must agree with the old one under the experimental conditions in which the old theory had been verified by experiments. The gravitational mass of a beam of light with energy E is $\frac{E}{c^2}$, and since c is a big number, it is not surprising that the weight of light rays had never been detected before Einstein trying to detect it.

- This book describes one particular theory of gravity, Einstein's theory of general relativity. There are other theories of gravity, and some of these, such as the *Brans-Dicke theory*, do just as well as general relativity in agreeing with the presently available experimental data. Our prediction of gravitational Doppler shifts of light only depended on the equivalence principle, which is one ingredient of general relativity. Experimental tests of this prediction only test the equivalence principle; they do not allow us to distinguish between one theory of gravity and another if both theories incorporate the equivalence principle.
- If an object such as a radio transmitter or an atom in an excited state emits an electromagnetic wave with a frequency f, then the object can be considered to be a type of clock. We can therefore interpret the gravitational red-shift as a gravitational time dilation: a difference in the rate at which time itself flows, depending on the gravitational potential. This is consistent with the empirical results presented in Section 1.2.

Example 3: Chiao's paradox revisited

The equivalence principle says that electromagnetic waves have gravitational mass as well as inertial mass, so it seems clear that the same must hold for static fields. In Chiao's paradox, the orbiting charged particle has an electric field that extends out to infinity. When we measure the mass of a charged particle such as an electron, there is no way to separate the mass of this field from a more localized contribution. The electric field "falls" through the gravitational field, and the equivalence principle, which is local, cannot guarantee that all parts of the field rotate uniformly about the earth, even in distant parts of the universe. The electric field pattern becomes distorted, and this distortion causes a radiation reaction which back-reacts on the particle, causing its orbit to decay.

The Pound-Rebka Experiment

The 1959 Pound-Rebka experiment at Harvard[31] was one of the first high-precision, relativistic tests of the equivalence principle to be carried out under controlled conditions, and in this section we will discuss it in detail.
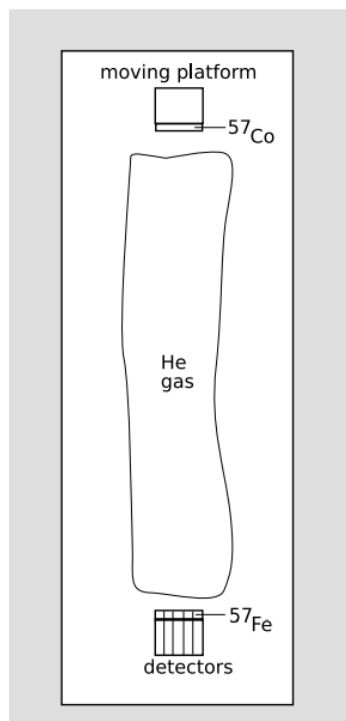
Figure 1.6.16 - The Pound-Rebka experiment.

When y is on the order of magnitude of the height of a building, the value of $\frac{\Delta E}{E} = \frac{gy}{c^2}$ is $\sim 10^{-14}$, so an extremely high-precision experiment is necessary in order to detect a gravitational red-shift. A number of other effects are big enough to obscure it entirely, and must somehow be eliminated or compensated for. These are listed below, along with their orders of magnitude in the experimental design finally settled on by Pound and Rebka.

(1) *Classical Doppler broadening due to temperature.* Thermal motion causes Doppler shifts of emitted photons, corresponding to the random component of the emitting atom's velocity vector along the direction of emission.	$\sim 10^{-6}$
(2) *The recoil Doppler shift.* When an atom emits a photon with energy E and momentum $p = \frac{E}{c}$, conservation of momentum requires that the atom recoil with momentum $p = -\frac{E}{c}$ and energy $\frac{p^2}{2m}$. This causes a downward Doppler shift of the energy of the emitted photon. A similar effect occurs on absorption, doubling the problem.	$\sim 10^{-12}$
(3) *Natural line width.* The Heisenberg uncertainty principle says that a state with a half-life τ must have an uncertainty in its energy of at least $\sim \frac{h}{\tau}$, where h is Planck's constant.	$\sim 10^{-12}$
(4) *Special-relativistic Doppler shift due to temperature.* Section 1.2 presented experimental evidence that time flows at a different rate depending on the motion of the observer. Therefore the thermal motion of an atom emitting a photon has an effect on the frequency of the photon, even if the atom's motion is not along the line of emission. The equations needed in order to calculate this effect will not be derived until section 2.2; a quantitative estimate is given in example 13. For now, we only need to know that this leads to a temperature-dependence in the average frequency of emission, in addition to the broadening of the bell curve described by effect (1) above.	$\sim 10^{-14}$ per degree C

The most straightforward way to mitigate effect (1) is to use photons emitted from a solid. At first glance this would seem like a bad idea, since electrons in a solid emit a continuous spectrum of light, not a discrete spectrum like the ones emitted by gases; this is because we have N electrons, where N is on the order of Avogadro's number, all interacting strongly with one another, so by the correspondence principle the discrete quantum-mechanical behavior must be averaged out. But the protons and neutrons within one nucleus do not interact much at all with those in other nuclei, so the photons emitted by a nucleus do have a discrete spectrum. The energy scale of nuclear excitations is in the keV or MeV range, so these photons are x-rays or gamma-rays. Furthermore, the timescale of the random vibrations of a nucleus in a solid are extremely short. For a velocity on the order of 100 m/s, and vibrations with an amplitude of $\sim 10^{-10}$ m, the time is about 10^{-12} s. In many cases, this is much shorter than the half-life of the excited nuclear state emitting the gamma-ray, and therefore the Doppler shift averages out to nearly zero.
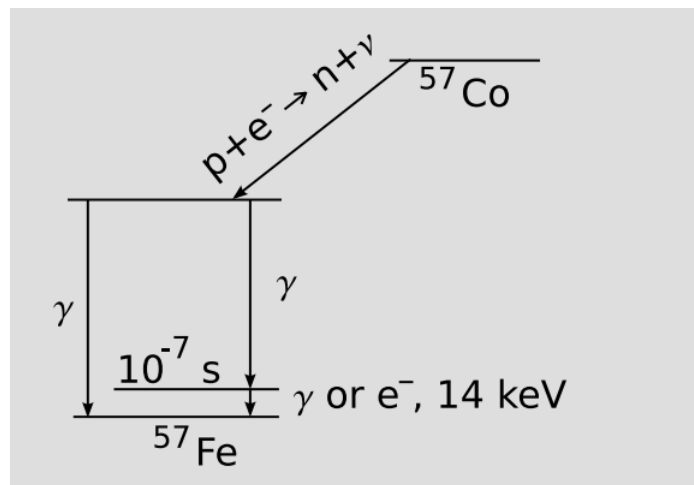
Figure 1.6.17 - Emission of 14 keV gamma-rays by ^{57}Fe. The parent nucleus ^{57}Co absorbs an electron and undergoes a weak-force decay process that converts it into ^{57}Fe, in an excited state. With 85% probability, this state decays to a state just above the ground state, with an excitation energy of 14 keV and a half-life of 10^{-7} s. This state finally decays, either by gamma emission or emission of an internal conversion electron, to the ground state.

Effect (2) is still much bigger than the 10^{-14} size of the effect to be measured. It can be avoided by exploiting the Mössbauer effect, in which a nucleus in a solid substance at low temperature emits or absorbs a gamma-ray photon, but with significant probability the recoil is taken up not by the individual nucleus but by a vibration of the atomic lattice as a whole. Since the recoil energy varies as $\frac{p^2}{2m}$, the large mass of the lattice leads to a very small dissipation of energy into the recoiling lattice. Thus if a photon is emitted and absorbed by identical nuclei in a solid, and for both emission and absorption the recoil momentum is taken up by the lattice as a whole, then there is a negligible energy shift. One must pick an isotope that emits photons with energies of about 10-100 keV. X-rays with energies lower than about 10 keV tend to be absorbed strongly by matter and are difficult to detect, whereas for gamma-ray energies $\gtrsim$ 100 keV the Mössbauer effect is not sufficient to eliminate the recoil effect completely enough.

If the Mössbauer effect is carried out in a horizontal plane, resonant absorption occurs. When the source and absorber are aligned vertically, p, gravitational frequency shifts should cause a mismatch, destroying the resonance. One can move the source at a small velocity (typically a few mm/s) in order to add a Doppler shift onto the frequency; by determining the velocity that compensates for the gravitational effect, one can determine how big the gravitational effect is.
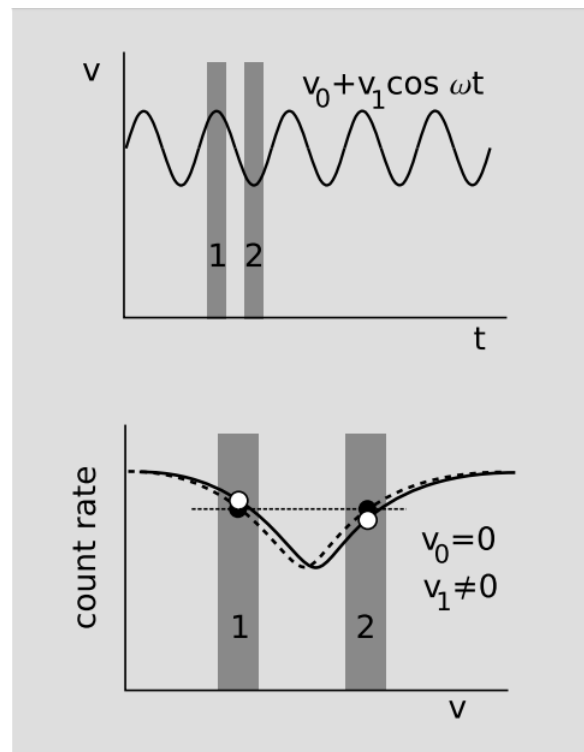
Figure 1.6.18 - Top: A graph of velocity versus time for the source. The velocity has both a constant component and an oscillating one with a frequency of 10-50 Hz. The constant component v_0 was used as a way of determining the calibration of frequency shift as a function of count rates. Data were acquired during the quarter-cycle periods of maximum oscillatory velocity, 1 and 2. Bottom:

Count rates as a function of velocity, for $v_0 = 0$ and $v_1 \neq 0$. The dashed curve and black circles represent the count rates that would have been observed if there were no gravitational effect. The gravitational effect shifts the resonance curve to one side (solid curve), resulting in an asymmetry of the count rates (open circles). The shift, and the resulting asymmetry, are greatly exaggerated for readability; in reality, the gravitational effect was 500 times smaller than the width of the resonance curve.

The typical half-life for deexcitation of a nucleus by emission of a gamma-ray with energy E is in the nanosecond range. To measure an gravitational effect at the 10^{-14} level, one would like to have a natural line width, (3), with $\frac{\Delta E}{E} \lesssim 10^{-14}$, which would require a half-life of $\gtrsim 10~\mu$s. In practice, Pound and Rebka found that other effects, such as (4) and electron-nucleus interactions that depended on the preparation of the sample, tended to put nuclei in one sample "out of tune" with those in another sample at the 10^{-13}-10^{-12} level, so that resonance could not be achieved unless the natural line width gave $\frac{\Delta E}{E} \gtrsim 10^{-12}$. As a result, they settled on an experiment in which 14 keV gammas were emitted by ^{57}Fe nuclei (Figure 1.5.17) at the top of a 22-meter tower, and absorbed by ^{57}Fe nuclei at the bottom. The 100-ns half-life of the excited state leads to $\frac{\Delta E}{E} \sim 10^{-12}$. This is 500 times greater than the gravitational effect to be measured, so, as described in more detail below, the experiment depended on high-precision measurements of small up-and-down shifts of the bell-shaped resonance curve.

The absorbers were seven iron films isotopically enhanced in ^{57}Fe, applied directly to the faces of seven sodium-iodide scintillation detectors (bottom of Figure 1.5.16). When a gamma-ray impinges on the absorbers, a number of different things can happen, of which we can get away with considering only the following: (a) the gamma-ray is resonantly absorbed in one of the ^{57}Fe absorbers, after which the excited nucleus decays by re-emission of another such photon (or a conversion electron), in a random direction; (b) the gamma-ray passes through the absorber and then produces ionization directly in the sodium iodide crystal. In case b, the gamma-ray is detected. In case a, there is a 50% probability that the re-emitted photon will come out in the upward direction, so that it cannot be detected. Thus when the conditions are right for resonance, a reduction in count rate is expected. The Mössbauer effect never occurs with 100% probability; in this experiment, about a third of the gammas incident on the absorbers were resonantly absorbed.

The choice of y = 22 m was dictated mainly by systematic errors. The experiment was limited by the strength of the gamma-ray source. For a source of a fixed strength, the count rate in the detector at a distance y would be proportional to y^{-2}, leading to statistical errors proportional to $1/\sqrt{count~rate} \propto y$. Since the effect to be measured is also proportional to y, the signal-to-noise ratio was independent of y. However, systematic effects such as (4) were easier to monitor and account for when y was fairly large.

A lab building at Harvard happened to have a 22-meter tower, which was used for the experiment. To reduce the absorption of the gammas in the 22 meters of air, a long, cylindrical mylar bag full of helium gas was placed in the shaft, p.

The resonance was a bell-shaped curve with a minimum at the natural frequency of emission. Since the curve was at a minimum, where its derivative was zero, the sensitivity of the count rate to the gravitational shift would have been nearly zero if the source had been stationary. Therefore it was necessary to vibrate the source up and down, so that the emitted photons would be Doppler shifted onto the shoulders of the resonance curve, where the slope of the curve was large. The resulting asymmetry in count rates is shown in Figure 1.5.18. A further effort to cancel out possible systematic effects was made by frequently swapping the source and absorber between the top and bottom of the tower.

For y = 22.6 m, the equivalence principle predicts a fractional frequency shift due to gravity of 2.46×10^{-15}. Pound and Rebka measured the shift to be $(2.56 \pm 0.25) \times 10^{-15}$. The results were in statistical agreement with theory, and verified the predicted size of the effect to a precision of 10%.

Figure_1.5.19b.png" />

Figure $\mathbf{1.6.19}$ - *Pound and Rebka at the top and bottom of the tower.*

References

[31] Phys. Rev. Lett. 4 (1960) 337

1.E: Geometric Theory of Spacetime (Exercises)

Problems

1. In classical mechanics, one hears the term "the acceleration of gravity," which doesn't literally make sense, since it is *objects* that accelerate. Explain why this term's usefulness is dependent on the equivalence principle.

2. The New Horizons space probe communicates with the earth using microwaves with a frequency of about 10 GHz. Estimate the sizes of the following frequency shifts in this signal, when the probe flies by Pluto in 2015, at a velocity of $\sim$ 10 A.U./year:

    a. the Doppler shift due to the probe's velocity;
    b. the Doppler shift due to the Earth's orbital velocity;
    c. the gravitational Doppler shift.

3. Euclid's axioms E1-E5 do not suffice to prove that there are an infinite number of points in the plane, and therefore they need to be supplemented by an extra axiom that states this (unless one finds the nonstandard realizations with finitely many points to be interesting enough to study for their own sake). Prove that the axioms of ordered geometry O1-O4 do not have this problem.

4. In the science fiction novel *Have Spacesuit — Will Travel*, by Robert Heinlein, Kip, a high school student, answers a radio distress call, encounters a flying saucer, and is knocked out and kidnapped by aliens. When he wakes up, he finds himself in a locked cell with a young girl named Peewee. Peewee claims they're aboard an accelerating spaceship. "If this was a spaceship," Kip thinks. "The floor felt as solid as concrete and motionless."
   The equivalence principle can be stated in a variety of ways. Earlier, it was stated as (1) gravitational and inertial mass are always proportional to one another. An alternative formulation is (2) that Kip has no way, by experiments or observations inside his sealed prison cell, to determine whether he's in an accelerating spaceship or on the surface of a planet, experiencing its gravitational field.

    a. Show that any violation of statement 1 also leads to a violation of statement 2.
    b. If we'd intended to construct a geometrical theory of gravity roughly along the lines of axioms O1-O4, which axiom is violated in this scenario?

5. Clock A sits on a desk. Clock B is tossed up in the air from the same height as the desk and then comes back down. Compare the elapsed times.

6. (a) Find the difference in rate between a clock at the center of the earth and a clock at the south pole. (b) When an antenna on earth receives a radio signal from a space probe that is in a hyperbolic orbit in the outer solar system, the signal will show both a kinematic red-shift and a gravitational blueshift. Compare the orders of magnitude of these two effects.

7. Consider the following physical situations: (1) a charged object lies on a desk on the planet earth; (2) a charged object orbits the earth; (3) a charged object is released above the earth's surface and dropped straight down; (4) a charged object is subjected to a constant acceleration by a rocket engine in outer space. In each case, we want to know whether the charge radiates. Analyze the physics in each case (a) based on conservation of energy; (b) by determining whether the object's motion is inertial in the sense intended by Isaac Newton; (c) using the most straightforward interpretation of the equivalence principle (i.e., not worrying about the issues discussed on p. that surround the ambiguous definition of locality).

8. Consider the physical situation depicted in Figure 1.5.12. Let ag be the gravitational acceleration and a$_r$ the acceleration of the charged particle due to radiation. Then $\frac{a_r}{a_g}$ measures the violation of the equivalence principle. The goal of this problem is to make an order-of-magnitude estimate of this ratio in the case of a neutron and a proton in low earth orbit.

    a. Let m the mass of each particle, and q the charge of the charged particle. Without doing a full calculation like the ones by the DeWitts and Grøn and Næss, use general ideas about the frequency-scaling of radiation (see section 9.2) to find the proportionality that gives the dependence of $\frac{a_r}{a_g}$ on q, m, and any convenient parameters of the orbit.
    b. Based on considerations of units, insert the necessary universal constants into your answer from part a.
    c. The result from part b will still be off by some unitless factor, but we expect this to be of order unity. Under this assumption, make an order-of-magnitude estimate of the violation of the equivalence principle in the case of a neutron and a proton in low earth orbit.

CHAPTER OVERVIEW

2: Geometry of Flat Spacetime

Topic hierarchy

Thumbnail: Light cone in 2D space plus a time dimension. (CC BY-SA 3.0; K. Aainsqatsi).

2.1: Introduction to Geometry of Flat Spacetime

The geometrical treatment of space, time, and gravity only requires as its basis the equivalence of inertial and gravitational mass. Given this assumption, we can describe the trajectory of any free-falling test particle as a geodesic. Equivalence of inertial and gravitational mass holds for Newtonian gravity, so it is indeed possible to redo Newtonian gravity as a theory of curved spacetime. This project was carried out by the French mathematician Cartan. The geometry of the local reference frames is very simple. The three space dimensions have an approximately Euclidean geometry, and the time dimension is entirely separate from them. This is referred to as a Euclidean spacetime with 3+1 dimensions. Although the outlook is radically different from Newton's, all of the predictions of experimental results are the same.

The experiments in Section 1.2 show, however, that there are real, experimentally verifiable violations of Newton's laws. In Newtonian physics, time is supposed to flow at the same rate everywhere, which we have found to be false.The flow of time is actually dependent on the observer's state of motion through space, which shows that the space and time dimensions are intertwined somehow. The geometry of the local frames in relativity therefore must not be as simple as Euclidean 3+1. Their actual geometry was implicit in Einstein's 1905 paper on special relativity, and had already been developed mathematically, without the full physical interpretation, by Hendrik Lorentz. Lorentz's and Einstein's work were explicitly connected by Minkowski in 1907, so a Lorentz frame is often referred to as a Minkowski frame.

Figure 2.0.1 - Hendrik Antoon Lorentz (1853-1928)

To describe this Lorentz geometry, we need to add more structure on top of the axioms O1-O4 of ordered geometry, but it will not be the additional Euclidean structure of E3-E4, it will be something different. To see how to proceed, let's start by thinking about what bare minimum of geometrical machinery is needed in order to set up frames of reference.

2.2: Affine Properties of Lorentz Geometry (Part 1)

Parallelism and measurement

We think of a frame of reference as a body of measurements or possible measurements to be made by some observer. Ordered geometry lacks measure. The following argument shows that merely by adding a notion of parallelism to our geometry, we automatically gain a system of measurement.

We only expect Lorentz frames to be local, but we do need them to be big enough to cover at least some amount of spacetime. If Betty does an Eötvös experiment by releasing a pencil and a lead ball side by side, she is essentially trying to release them at the same event A, so that she can observe them later and determine whether their world-lines stay right on top of one another at point B. That was all that was required for the Eötvös experiment, but in order to set up a Lorentz frame we need to start dealing with objects that are not right on top of one another. Suppose we release two lead balls in two different locations, at rest relative to one another. This could be the first step toward adding measurement to our geometry, since the balls mark two points in space that are separated by a certain distance, like two marks on a ruler, or the goals at the ends of a soccer field. Although the balls are separated by some finite distance, they are still close enough together so that if there is a gravitational field in the area, it is very nearly the same in both locations, and we expect the distance defined by the gap between them to stay the same. Since they are both subject only to gravitational forces, their world-lines are by definition straight lines (geodesics). The goal here is to end up with some kind of coordinate grid defining a (t,x) plane, and on such a grid, the two balls' world-lines are vertical lines. If we release them at events P and Q, then observe them again later at R and S, PQRS should form a rectangle on such a plot. In the figure, the irregularly spaced tick marks along the edges of the rectangle are meant to suggest that although ordered geometry provides us with a well-defined ordering along these lines, we have not yet constructed a complete system of measurement.
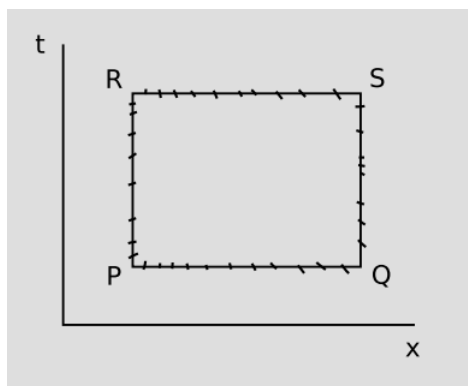
Figure 2.2.1 - Objects are released at rest at spacetime events P and Q. They remain at rest, and their world-lines define a notion of parallelism.

The depiction of PQSR as a rectangle, with right angles at its vertices, might lead us to believe that our geometry would have something like the concept of angular measure referred to in Euclid's E4, equality of right angles. But this is too naive even for the Euclidean 3+1 spacetime of Newton and Galileo. Suppose we switch to a frame that is moving relative to the first one, so that the balls are not at rest. In the Euclidean spacetime, time is absolute, so events P and Q would remain simultaneous, and so would R and S; the top and bottom edges PQ and RS would remain horizontal on the plot, but the balls' world-lines PR and QS would become slanted. The result would be a parallelogram. Since observers in different states of motion do not agree on what constitutes a right angle, the concept of angular measure is clearly not going to be useful here. Similarly, if Euclid had observed that a right angle drawn on a piece of paper no longer appeared to be a right angle when the paper was turned around, he would never have decided that angular measure was important enough to be enshrined in E4.
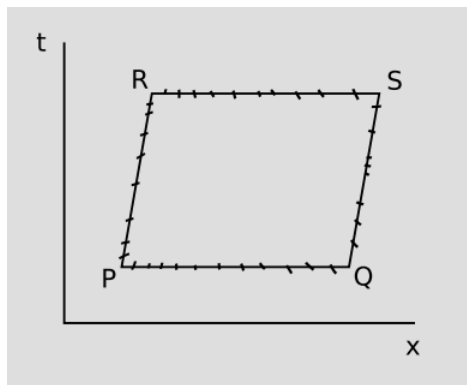
Figure 2.2.2 - There is no well-defined angular measure in this geometry. In a different frame of reference, the angles are not right angles.

In the context of relativity, where time is not absolute, there is not even any reason to believe that different observers must agree on the simultaneity of PQ and RS. Our observation that time flows differently depending on the observer's state of motion tells us specifically to expect this *not* to happen when we switch to a frame moving to the relative one. Thus in general we expect that PQRS will be distorted into a form like the one shown in Figure 2.1.3. We do expect, however, that it will remain a parallelogram; a Lorentz frame is one in which the gravitational field, if any, is constant, so the properties of spacetime are uniform, and by symmetry the new frame should still have PR=QS and PQ=RS.
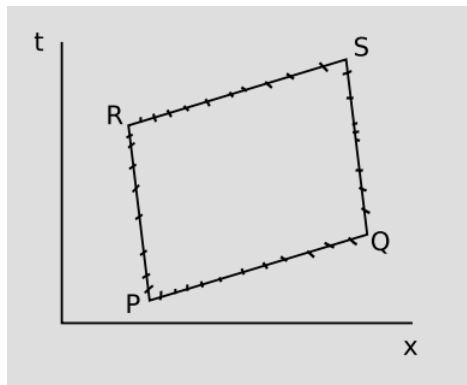
Figure 2.2.3 - Simultaneity is not well defined. The constant-time lines PQ and RS from figure a are not constant-time lines when observed in a different frame of reference.

With this motivation, we form the system of *affine geometry* by adding the following axioms to set O1-O4.[1] The notation [PQRS] means that events P, Q, S, and R form a parallelogram, and is defined as the statement that the lines determined by PQ and RS never meet at a point, and similarly for PR and QS.

- **A1** Constructibility of parallelograms: Given any P, Q, and R, there exists S such that [PQRS], and if P, Q, and R are distinct then S is unique.
- **A2** Symmetric treatment of the sides of a parallelogram: If [PQRS], then [QRSP], [QPSR], and [PRQS].
- **A3** Lines parallel to the same line are parallel to one another: If [ABCD] and [ABEF], then [CDEF].

> Note
>
> The axioms are summarized for convenient reference in Appendix C. This formulation is essentially the one given by Penrose, The Road to Reality, in section 14.1.

The following theorem is a stronger version of Playfair's axiom E5, the interpretation being that affine geometry describes a spacetime that is locally flat.

Theorem: Given any line ℓ and any point P not on the line, there exists a unique line through P that is parallel to ℓ.

This is stronger than E5, which only guarantees uniqueness, not existence. Informally, the idea here is that A1 guarantees the existence of the parallel, and A3 makes it unique.[2]

proof

Pick any two distinct points A and B on ℓ, and construct the uniquely determined parallelogram [ABPQ] (axiom A1). Points P and Q determine a line (axiom O1), and this line is parallel to ℓ (definition of the parallelogram). To prove that this line is unique, we argue by contradiction. Suppose some other parallel m to exist. If m crosses the infinite line BQ at some point Z, then both [ABPQ] and [ABPZ], so by A1, Q=Z, so the ℓ and m are the same. The only other possibility is that m is parallel to BQ, but then the following chain of parallelisms holds: PQ ∥ AB ∥ m ∥ BQ. By A3, lines parallel to another line are parallel to each other, so PQ ∥ BQ, but this is a contradiction, since they have Q in common.

Although these new axioms do nothing more than to introduce the concept of parallelism lacking in ordered geometry, it turns out that they also allow us to build up a concept of measurement. Let ℓ be a line, and suppose we want to define a number system on this line that measures how far apart events are. Depending on the type of line, this could be a measurement of time, of spatial distance, or a mixture of the two. First we arbitrarily single out two distinct points on ℓ and label them 0 and 1. Next, pick some auxiliary point q_0 not lying on ℓ. By A1, construct the parallelogram $01q_0q_1$. Next construct $q_0 1 q_1 2$. Continuing in this way, we have a scaffolding of parallelograms adjacent to the line, determining an infinite lattice of points 1, 2, 3, . . . on the line, which represent the positive integers. Fractions can be defined in a similar way. For example, $\frac{1}{2}$ is defined as the point such that when the initial lattice segment $0 \frac{1}{2}$ is extended by the same construction, the next point on the lattice is 1.
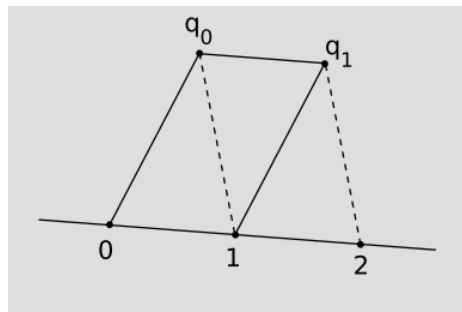
Figure 2.2.4 - Construction of an affine parameter.

The continuously varying variable constructed in this way is called an *affine parameter*. The time measured by a free-falling clock is an example of an affine parameter, as is the distance measured by the tick marks on a free-falling ruler. Since light rays travel along geodesics, the wave crests on a light wave can even be used analogously to the ruler's tick marks.

Example 1: centroids

The affine parameter can be used to define the centroid of a set of points. In the simplest example, finding the centroid of two points, we simply bisect the line segment as described above in the construction of the number $\frac{1}{2}$. Similarly, the centroid of a triangle can be defined as the intersection of its three medians, the lines joining each vertex to the midpoint of the opposite side.
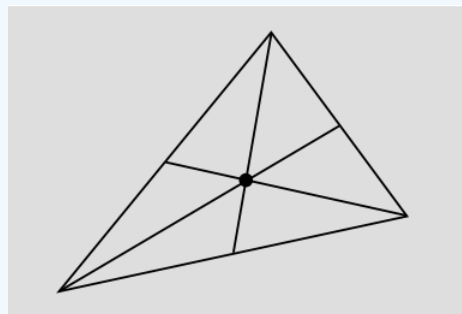
Figure 2.2.5 - Affine geometry gives a well-defined centroid for the triangle.

Example 2: Conservation of momentum

In nonrelativistic mechanics, the concept of the center of mass is closely related to the law of conservation of momentum. For example, a logically complete statement of the law is that if a system of particles is not subjected to any external force, and we pick a frame in which its center of mass is initially at rest, then its center of mass remains at rest in that frame. Since centroids are well defined in affine geometry, and Lorentz frames have affine properties, we have grounds to hope that it might be possible to generalize the definition of momentum relativistically so that the generalized version is conserved in a Lorentz frame. On the other hand, we don't expect to be able to define anything like a global Lorentz frame for the entire universe, so there is no such natural expectation of being able to define a global principle of conservation of momentum. This is an example of a general fact about relativity, which is that conservation laws are difficult or impossible to formulate globally.

Although the affine parameter gives us a system of measurement for free in a geometry whose axioms do not even explicitly mention measurement, there are some restrictions:

1. The affine parameter is defined only along straight lines, i.e., geodesics. Alice's clock defines an affine parameter, but Betty's does not, since it is subject to nongravitational forces.
2. We cannot compare distances along two arbitrarily chosen lines, only along a single line or two parallel lines.
3. The affine parameter is arbitrary not only in the choice of its origin 0 (which is to be expected in any case, since any frame of reference requires such an arbitrary choice) but also in the choice of scale. For example, there is no fundamental way of deciding how fast to make a clock tick.

We will eventually want to lift some of these restrictions by adding to our kit a tool called a metric, which allows us to define distances along arbitrary curves in space time, and to compare distances in different directions. The affine parameter, however, will not be entirely superseded. In particular, we'll find that the metric has a couple of properties that are not as nice as those of the affine parameter. The square of a metric distance can be negative, and the metric distance measured along a light ray is precisely zero, which is not very useful.

Exercise 2.2.1

Self-check: By the construction of the affine parameter above, affine distances on the same line are comparable. By another construction, verify the claim made above that this can be extended to distances measured along two different parallel lines.

Example 3: Area and volume

It is possible to define area and volume in affine geometry. This is a little surprising, since distances along different lines are not even comparable. However, we are already accustomed to multiplying and dividing numbers that have different units (a concept that would have given Euclid conniptions), and the situation in affine geometry is really no different. To define area, we extend the one-dimensional lattice to two dimensions. Any planar figure can be superimposed on such a lattice, and dissected into parallelograms, each of which has a standard area.

Figure 2.2.6 - The area of the viola can be determined by counting the parallelograms formed by the lattice. The area can be determined to any desired precision, by dividing the parallelograms into fractional parts that are as small as necessary.

Example 4: Area on a graph of v versus t

If an object moves at a constant velocity v for time t, the distance it travels can be represented by the area of a parallelogram in an affine plane with sides having lengths v and t. These two lengths are measured by affine parameters along two different directions, so they are not comparable. For example, it is meaningless to ask whether 1 m/s is greater than, less than, or equal to 1 s. If we were graphing velocity as a function of time on a conventional Cartesian graph, the v and t axes would be perpendicular, but affine geometry has no notion of angular measure, so this is irrelevant here.
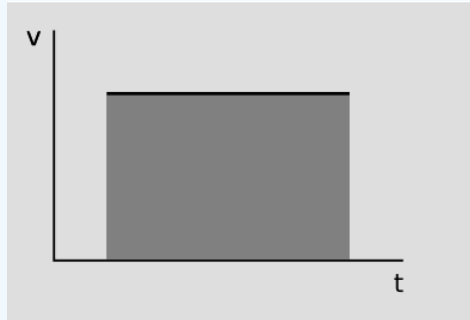
Figure 2.2.7

Exercise 2.2.2

Self-check: If multiplication is defined in terms of affine area, prove the commutative property ab = ba and the distributive rule a(b + c) = ab + bc from axioms A1-A3.

2.3: Affine Properties of Lorentz Geometry (Part 2)

Vectors

Vectors Distinguished from Scalars

We've been discussing subjects like the center of mass that in freshman mechanics would be described in terms of vectors and scalars, the distinction being that vectors have a direction in space and scalars don't. As we make the transition to relativity, we are forced to refine this distinction. For example, we used to consider time as a scalar, but the Hafele-Keating experiment shows that time is different in different frames of reference, which isn't something that's supposed to happen with scalars such as mass or temperature. In affine geometry, it doesn't make much sense to say that a vector has a magnitude and direction, since non-parallel magnitudes aren't comparable, and there is no system of angular measurement in which to describe a direction.

A better way of defining vectors and scalars is that scalars are absolute, vectors relative. If I have three apples in a bowl, then all observers in all frames of reference agree with me on the number three. But if my terrier pup pulls on the leash with a certain force vector, that vector has to be defined in relation to other things. It might be three times the strength of some force that we define as one newton, and in the same direction as the earth's magnetic field.

In general, measurement means comparing one thing to another. The number of apples in the bowl isn't a measurement, it's a count.

Affine Measurement of Vectors

Before even getting into the full system of affine geometry, let's consider the one-dimensional example of a line of time. We could use the hourly emergence of a mechanical bird from a pendulumdriven cuckoo clock to measure the rate at which the earth spins, but we could equally well take our planet's rotation as the standard and use it to measure the frequency with which the bird pops out of the door. Once we have two things to compare against one another, measurement is reduced to counting (Figure 2.1.4). Schematically, let's represent this measurement process with the following notation, which is part of a system called called birdtracks:[3]

$$c \to e = 24 \qquad\qquad (2.3.1)$$

> **Note**
>
> The system used in this book follows the one defined by Cvitanović, which was based closely on a graphical notation due to Penrose. For a more complete exposition, see the Wikipedia article "Penrose graphical notation" and Cvitanović's online book at birdtracks.eu.

Here c represents the cuckoo clock and e the rotation of the earth. Although the measurement relationship is nearly symmetric, the arrow has a direction, because, for example, the measurement of the earth's rotational period in terms of the clock's frequency is c $\to$ e = (24 hr)(1 hr^{-1}) = 24, but the clock's period in terms of the earth's frequency is e $\to$ c = $\frac{1}{24}$. We say that the relationship is not symmetric but "dual." By the way, it doesn't matter how we arrange these diagrams on the page. The notations c $\to$ e and e $\to$ c mean exactly the same thing, and expressions like this can even be drawn vertically.

Suppose that e is a displacement along some one-dimensional line of time, and we want to think of it as the thing being measured. Then we expect that the measurement process represented by c produces a real-valued result and is a linear function of e. Since the relationship between c and e is dual, we expect that c also belongs to some vector space. For example, vector spaces allow multiplication by a scalar: we could double the frequency of the cuckoo clock by making the bird come out on the half hour as well as on the hour, forming 2c. Measurement should be a linear function of both vectors; we say it is "bilinear."

Duality

The two vectors c and e have different units, hr^{-1} and hr, and inhabit two different one-dimensional vector spaces. The "flavor" of the vector is represented by whether the arrow goes into it or comes out. Just as we used notation like $\vec{v}$ in freshman physics to tell vectors apart from scalars, we can employ arrows in the birdtracks notation as part of the notation for the vector, so that instead of writing the two vectors as c and e, we can notate them as c $\to$ and $\to$ e. Performing a measurement is like plumbing. We join the two "pipes" in c $\to$ $\to$ e and simplify to c $\to$ e.

A confusing and nonstandardized jungle of notation and terminology has grown up around these concepts. For now, let's refer to a vector such as $\rightarrow$ e, with the arrow coming in, simply as a "vector," and the type like c $\rightarrow$ as a "dual vector." In the one-dimensional example of the earth and the cuckoo clock, the roles played by the two vectors were completely equivalent, and it didn't matter which one we expressed as a vector and which as a dual vector. Example 5 shows that it is sometimes more natural to take one quantity as a vector and another as a dual vector. Example 6 shows that we sometimes have no choice at all as to which is which.

Scaling

In birdtracks notation, a scalar is a quantity that has no external arrows at all. Since the expression c $\rightarrow$ e = 24 has no external arrows, only internal ones, it represents a scalar. This makes sense because it's a count, and a count is a scalar.

A convenient way of summarizing all of our categories of variables is by their behavior when we convert units, i.e., when we rescale our space. If we switch our time unit from hours to minutes, the number of apples in a bowl is unchanged, the earth's period of rotation gets 60 times bigger, and the frequency of the cuckoo clock changes by a factor of $\frac{1}{60}$. In other words, a quantity u under rescaling of coordinates by a factor α becomes $\alpha^p u$, where the exponents -1, 0, and $+1$ correspond to dual vectors, scalars, and vectors, respectively. We can therefore see that these distinctions are of interest even in one dimension, contrary to what one would have expected from the freshman-physics concept of a vector as something transforming in a certain way under rotations.

Geometrical Visualization

In two dimensions, there are natural ways of visualizing the different vector spaces inhabited by vectors and dual vectors. We've already been describing a vector like $\rightarrow$ e as a displacement. Its vector space is the space of such displacements.[4] A vector in the dual space such as c $\rightarrow$ can be visualized as a set of parallel, evenly spaced lines on a topographic map, $\frac{h}{2}$, with an arrowhead to show which way is "uphill." The act of measurement consists of counting how many of these lines are crossed by a certain vector, $\frac{h}{3}$.
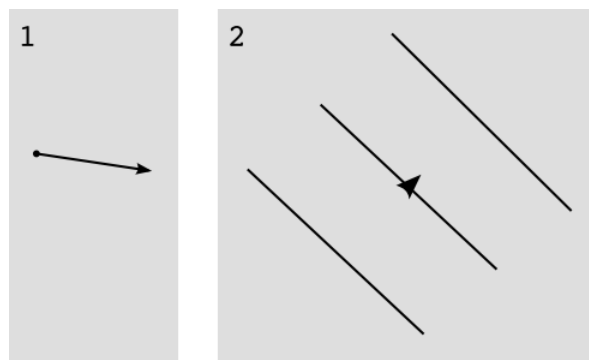
Figure_2.1.8b.png" />

***Figure* 2.3.8** - *1. A displacement vector. 2. A vector from the space dual to the space of displacements. 3. Measurement is reduced to counting. The cuckoo clock chimes 24 times in one rotation of the earth.*

> **Note**
>
> In terms of the primitive notions used in the axiomatization in section 2.1, a displacement could be described as an equivalence class of segments such that for any two segments in the class AB and CD, AB and CD form a parallelogram.

Given a scalar field f, its gradient ∇f at any given point is a dual vector. In birdtracks notation, we have to indicate this by writing it with an outward-pointing arrow, (grad f) $\rightarrow$. Because gradients occur so frequently, we have a special shorthand for them, which is simply a circle:

$$\widehat{f}\text{-}$$

In the context of spacetime with a metric and curvature, we'll see that the usual definition of the gradient in terms of partial derivatives should be modified with correction terms to form something called a covariant derivative. When we get to that point later, we'll commandeer the circle notation for that operation.
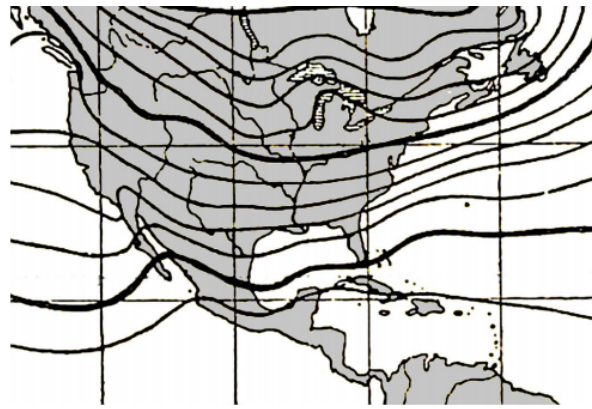
Figure 2.3.9 - Constant-temperature curves for January in North America, at intervals of 4 °C. The temperature gradient at a given point is a dual vector.

Example 5: force is a dual vector

The dot product dW = $\mathbf{F} \cdot d\mathbf{x}$ for computing mechanical work becomes, in birdtracks notation,

$$dW = F \rightarrow dx. \tag{2.3.2}$$

This shows that force is more naturally considered to be a dual vector rather than a vector. The symmetry between vectors and dual vectors is broken by considering displacements like $\rightarrow dx$ to be vectors, and this asymmetry then spreads to other quantities such as force.

The same result can be obtained from Newton's second law; see example 21.

Example 6: Systems without a metric

The freshman-mechanics way of thinking about vectors and scalar products contains the hidden assumption that we have, besides affine measurement, an additional piece of measurement apparatus called the metric (section 3.5). Without yet having to formally define what we mean by a metric, we can say roughly that it supplies the conveniences that we're used to having in the Euclidean plane, but that are not present in affine geometry. In particular, it allows us to define the notion that one vector is perpendicular to another vector, or that one dual vector is perpendicular to another dual vector.

Let's start with an example where the hidden assumption is valid, and we do have a metric. Let a billiard ball of unit mass be constrained by a diagonal wall to have C ≤ 0, where C = y − x. The Lagrangian formalism just leads to the expected Newtonian expressions for the momenta conjugate to x and y, $p_x = \dot{x}$, $p_y = \dot{y}$, and these form a dual vector p → . The force of constraint is $F \rightarrow = \frac{dp \rightarrow}{dt}$. Let w → = (grad C) → be the gradient of the constraint function. The vectors F → and w → both belong to the space of dual vectors, and they are parallel to each other. Since we do happen to have a metric in this example, it is also possible to say, as most people would, that the force is perpendicular to the wall.

Now consider the example shown in Figure 2.1.10. The arm's weight is negligible compared to the unit mass of the gripped weight, and both the upper and lower arm have unit length. Elbows don't bend backward, so we have a constraint C ≤ 0, where C = $\theta - \varphi$, and as before we can define define a dual vector w → = (grad C) → that is parallel to the line of constraint in the (θ, φ) plane. The conjugate momenta (which are actually angular momenta) turn out to be $p_\theta = \dot{\theta} + \cos(\varphi - \theta)\dot{\varphi}$ and a similar expression for p_φ. As in the example of the billiard ball, the force of constraint is parallel to w → . There is no metric that naturally applies to the (θ, φ) plane, so we have no notion of perpendicularity, and it doesn't make sense to say that F → is perpendicular to the line of constraint.
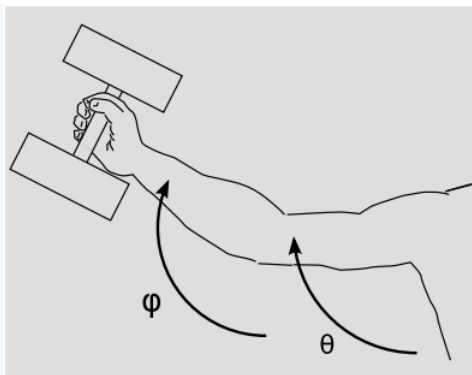
Figure 2.3.10 - There is no natural metric on the space (θ, φ).

Finally we remark that since four-dimensional Galilean spacetime lacks a metric (see section 3.5), the distinction between vectors and dual vectors in Galilean relativity is a real and physically important one. The only reason people were historically able to ignore this distinction was that Galilean spacetime splits into independent time and spatial parts, with the spatial part having a metric.

Example 7: No simultaneity without a metric

We'll see in section 2.2 that one way of defining the distinction between Galilean and Lorentz geometry is that in Lorentzian spacetime, simultaneity is observer-dependent. Without a metric, there can be no notion of simultaneity at all, not even a frame-dependent one. In Figure 2.1.11, the fact that the observer considers events P and Q to be simultaneous is represented by the fact that the observer's displacement vector $\to o$ is perpendicular to the displacement $\to s$ from P to Q. In affine geometry, we can't express perpendicularity.
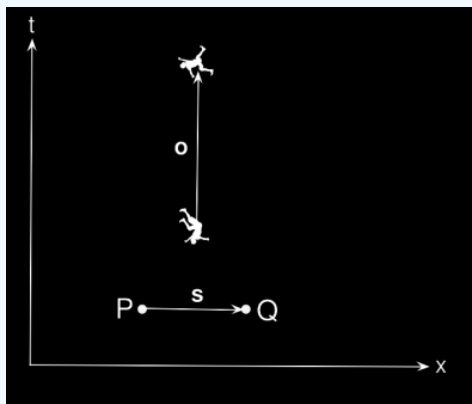
Figure 2.3.11 - The free-falling observer considers P and Q to be simultaneous.

Abstract Index Notation

Expressions in birdtracks notation such as

$$\boxed{C} \!\!\to\!\! s$$

can be awkward to type on a computer, which is why we've already been occasionally resorting to more linear notations such as (grad C) $\to$ s. As we encounter more complicated birdtracks, the diagrams will sometimes look like complicated electrical schematics, and the problem of generating them on a keyboard will get more acute. There is in fact a systematic way of representing any such expression using only ordinary subscripts and superscripts. This is called abstract index notation, and was introduced by Roger Penrose at around the same time he invented birdtracks. For practical reasons, it was the abstract index notation that caught on.

The idea is as follows. Suppose we wanted to describe a complicated birdtrack verbally, so that someone else could draw it. The diagram would be made up of various smaller parts, a typical one looking something like the scalar product u $\to$ v. The verbal

instructions might be: "We have an object u with an arrow coming out of it. For reference, let's label this arrow as a. Now remember that other object v I had you draw before? There was an arrow coming into that one, which we also labeled a. Now connect up the two arrows labeled a."

Shortening this lengthy description to its bare minimum, Penrose renders it like this: $u_a v^a$. Subscripts depict arrows coming out of a symbol (think of water flowing from a tank out through a pipe below). Superscripts indicate arrows going in. When the same letter is used as both a superscript and a subscript, the two arrows are to be piped together.

Abstract index notation evolved out of an earlier one called the Einstein summation convention, in which superscripts and subscripts referred to specific coordinates. For example, we might take 0 to be the time coordinate, 1 to be x, and so on. A symbol like u_γ would then indicate a component of the dual vector u, which could be its x component if γ took on the value 1. Repeated indices were summed over.

The advantage of the birdtrack and abstract index notations is that they are coordinate-independent, so that an equation written in them is valid regardless of the choice of coordinates. The Einstein and Penrose notations look very similar, so for example if we want to take a general result expressed in Penrose notation and apply it in a specific coordinate system, there is essentially no translation required. In fact, the two notations look so similar that we need an explicit way to tell which is which, so that we can tell whether or not a particular result is coordinate-independent. We therefore use the convention that Latin indices represent abstract indices, whereas Greek ones imply a specific coordinate system and can take on numerical values, e.g., $\gamma = 1$.

2.4: Relativistic Properties of Lorentz Geometry (Part 1)

We now want to pin down the properties of the Lorentz geometry that are left unspecified by the affine treatment. We need some further input from experiments in order to show us how to proceed. We take the following as empirical facts about flat spacetime:[5]

- **L1** *Spacetime is homogeneous and isotropic*. No time or place has special properties that make it distinguishable from other points, nor is one direction in space distinguishable from another.[6]
- **L2** *Inertial frames of reference* exist. These are frames in which particles move at constant velocity if not subject to any forces. We can construct such a frame by using a particular particle, which is not subject to any forces, as a reference point.
- **L3** *Equivalence of inertial frames*: If a frame is in constant-velocity translational motion relative to an inertial frame, then it is also an inertial frame. No experiment can distinguish one preferred inertial frame from all the others.
- **L4** *Causality*: There exist events 1 and 2 such that $t_1 < t_2$ in all frames.
- **L5** *Relativity of time*: There exist events 1 and 2 and frames of reference (t, x) and (t', x') such that $t_1 < t_2$, but $t'_1 > t'_2$.

> **Note**
>
> These facts are summarized for convenience in Appendix C.

> **Note**
>
> For the experimental evidence on isotropy, see http://www.edu-observatory.org/physi...tropy_of_space.

L4 makes it possible to have an event 1 that causes an event 2, with all observers agreeing on which caused which. L5 is supported by the experimental evidence in section 1.2; if L5 were false, then space and time could work as imagined by Galileo and Newton.

Define affine parameters t and x for time and position, and construct a (t, x) plane. Axiom L1 guarantees that spacetime is flat, allowing us to do this; if spacetime had, for example, a curvature like that of a sphere, then the axioms of affine geometry would fail, and it would be impossible to lay out such a global grid of parallels. Although affine geometry treats all directions symmetrically, we're going beyond the affine aspects of the space, and t does play a different role than x here, as shown, for example, by L4 and L5.

In the (t, x) plane, consider a rectangle with one corner at the origin O. We can imagine its right and left edges as representing the world-lines of two objects that are both initially at rest in this frame; they remain at rest (L2), so the right and left edges are parallel.

How do we know that this is a rectangle and not some other kind of parallelogram? In purely affine geometry, there is no notion of perpendicularity, so this distinction is meaningless. But implicit in the existence of inertial frames (L2) is the assumption that spacetime has some additional structure that allows a particular observer to decide what events he considers to be simultaneous (example 7). He then considers his own world-line, i.e., his t axis, to be perpendicular to a proposed x axis if points on the x axis are simultaneous to him.

We now define a second frame of reference such that the origins of the two frames coincide, but they are in motion relative to one another with velocity v. The transformation L from the first frame to the second is referred to as a Lorentz boost with velocity v. L depends on v. By equivalence of inertial frames (L3), an observer in the new frame considers his own t axis to be perpendicular to his own x, even though they don't look that way in figure a. Thus, although we assume some notion of perpendicularity, we do not assume that it looks the same as the Euclidean one.

By homogeneity of spacetime (L1), L must be linear, so the original rectangle will be transformed into a parallelogram in the new frame; this is also consistent with L3, which requires that the world-lines on the right and left edges remain parallel. The left edge has inverse slope v. By L5 (no simultaneity), the top and bottom edges are no longer horizontal.
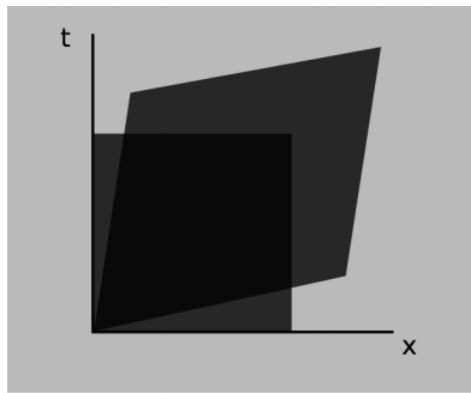
Figure 2.4.1 - Two objects at rest have world-lines that define a rectangle. In a second frame of reference in motion relative to the first one, the rectangle becomes a parallelogram.

For simplicity, let the original rectangle have unit area. Then the area of the new parallelogram is still 1, by the following argument. Let the new area be A, which is a function of v. By isotropy of spacetime (L1), A(v) = A(−v). Furthermore, the function A(v) must have some universal form for all geometrical figures, not just for a figure that is initially a particular rectangle; this follows because of our definition of affine area in terms of a dissection by a two-dimensional lattice, which we can choose to be a lattice of squares. Applying boosts +v and −v one after another results in a transformation back into our original frame of reference, and since A is universal for all shapes, it doesn't matter that the second transformation starts from a parallelogram rather than a square. Scaling the area once by A(v) and again by A(−v) must therefore give back the original square with its original unit area, A(v)A(−v) = 1, and since A(v) = A(−v), A(v) = ±1 for any value of v. Since A(0) = 1, we must have A(v) = 1 for all v. The argument is independent of the shape of the region, so we conclude that all areas are preserved by Lorentz boosts. (See section 4.6 for further interpretation of A.)

If we consider a boost by an infinitesimal velocity dv, then the vanishing change in area comes from the sum of the areas of the four infinitesimally thin slivers where the rectangle lies either outside the parallelogram (call this negative area) or inside it (positive). (We don't worry about what happens near the corners, because such effects are of order dv^2.) In other words, area flows around in the x − t plane, and the flows in and out of the rectangle must cancel. Let v be positive; the flow at the sides of the rectangle is then to the right. The flows through the top and bottom cannot be in opposite directions (one up, one down) while maintaining the parallelism of the opposite sides, so we have the following three possible cases:
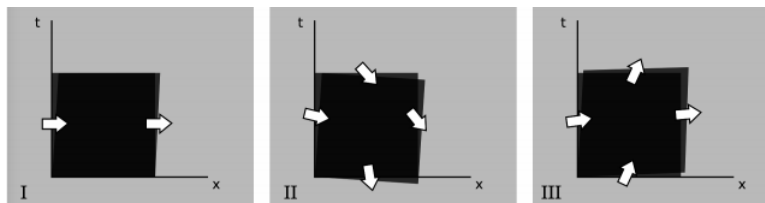
Figure 2.4.2 - Flows of area: (I) a shear that preserves simultaneity, (II) a rotation, (III) upward flow at all edges.

  I. There is no flow through the top and bottom. This case corresponds to Galilean relativity, in which the rectangle shears horizontally under a boost, and simultaneity is preserved, violating L5.
 II. Area flows downward at both the top and the bottom. The flow is clockwise at both the positive t axis and the positive x axis. This makes it plausible that the flow is clockwise everywhere in the (t, x) plane, and the proof is straightforward.[7] As v increases, a particular element of area flows continually clockwise. This violates L4, because two events with a cause and effect relationship could be time-reversed by a Lorentz boost.
III. Area flows upward at both the top and the bottom.

> **proof**
>
> By linearity of L, the flow is clockwise at the negative axes as well. Also by linearity, the handedness of the flow is the same at all points on a ray extending out from the origin in the direction θ. If the flow were counterclockwise somewhere, then it would have to switch handedness twice in that quadrant, at θ_1 and θ_2. But by writing out the vector cross product r × dr, where dr is the displacement caused by L(dv), we find that it depends on $\sin(2\theta + \delta)$, which does not oscillate rapidly enough to have two zeroes in the same quadrant.

Only case III is possible, and given case III, there must be at least one point P in the first quadrant where area flows neither clockwise nor counterclockwise.[8] The boost simply increases P's distance from the origin by some factor. By the linearity of the transformation, the entire line running through O and P is simply rescaled. This special line's inverse slope, which has units of velocity, apparently has some special significance, so we give it a name, c. We'll see later that c is the maximum speed of cause and effect whose existence we inferred in section 1.3. Any world-line with a velocity equal to c retains the same velocity as judged by moving observers, and by isotropy the same must be true for −c.

> Note
>
> This follows from the fact that, as shown in the preceding footnote, the handedness of the flow depends only on θ.

For convenience, let's adopt time and space units in which c = 1, and let the original rectangle be a unit square. The upper right tip of the parallelogram must slide along the line through the origin with slope +1, and similarly the parallelogram's other diagonal must have a slope of −1. Since these diagonals bisected one another on the original square, and since bisection is an affine property that is preserved when we change frames of reference, the parallelogram must be equilateral.

We can now determine the complete form of the Lorentz transformation. Let unit square PQRS, as described above, be transformed to parallelogram P'Q'R'S' in the new coordinate system (x', t'). Let the t' coordinate of R' be γ, interpreted as the ratio between the time elapsed on a clock moving from P' to R' and the corresponding time as measured by a clock that is at rest in the (x', t') frame. By the definition of v, R' has coordinates $(v\gamma, \gamma)$, and the other geometrical facts established above place Q' symmetrically on the other side of the diagonal, at $(\gamma, v\gamma)$. Computing the cross product of vectors P'R' and P'Q', we find the area of P'Q'R'S' to be $\gamma^2(1 - v^2)$, and setting this equal to 1 gives

$$\gamma = \frac{1}{\sqrt{1 - v^2}} \tag{2.4.1}$$

The result for the transformation L, a Lorentz boost along the x axis with velocity v, is:

$$t' = \gamma t + v\gamma x \tag{2.4.2}$$

$$x' = v\gamma t + \gamma x \tag{2.4.3}$$

The symmetry of P'Q'R'S' with respect to reflection across the diagonal indicates that the time and space dimensions are treated symmetrically, although they are not entirely interchangeable as they would have been in case II.
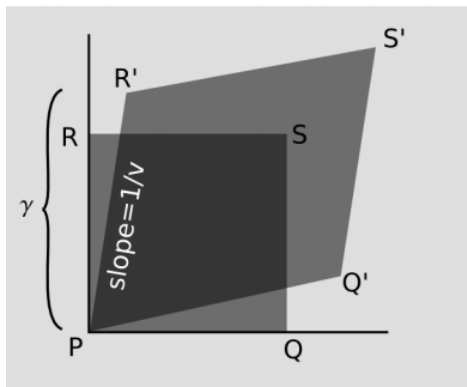
Figure 2.4.3 - Unit square PQRS is Lorentz-boosted to the parallelogram P'Q'R'S'.
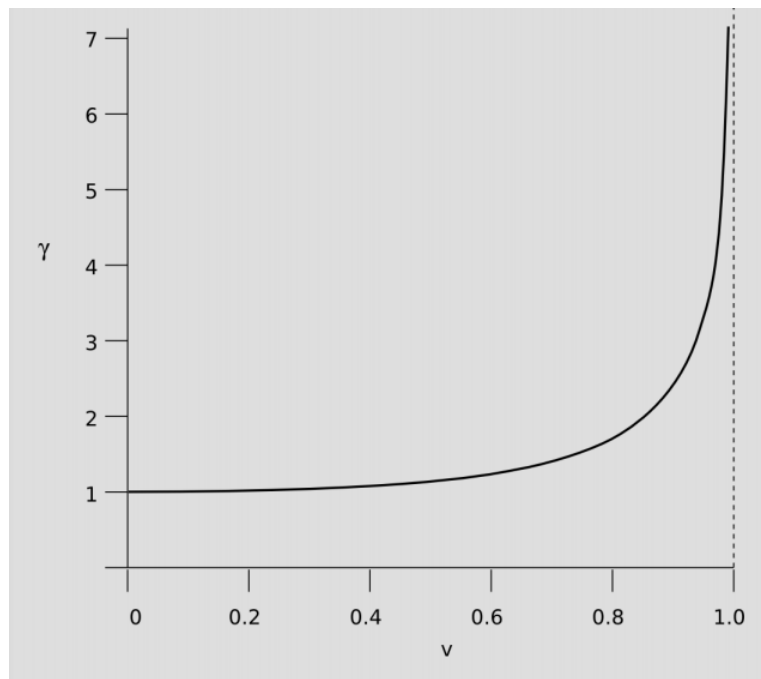
Figure 2.4.4 - The behavior of the γ factor.

A measuring rod, unlike a clock, sweeps out a two-dimensional strip on an $x - t$ graph. As in Galilean relativity, the two observers disagree on the positions of events at the two ends of their rods, but in addition they disagree on the simultaneity of such events. Calculation shows that a moving rod appears contracted by a factor γ.

In summary, a clock runs fastest according to an observer who is at rest relative to the clock, and a measuring rod likewise appears longest in its own rest frame.

The lack of a universal notion of simultaneity has a similarly symmetric interpretation. In prerelativistic physics, points in space have no fixed identity. A brass plaque commemorating a Civil War battle is not at the same location as the battle, according to an observer who perceives the Earth has having been hurtling through space for the intervening centuries. By symmetry, points in time have no fixed identity either.

In everyday life, we don't notice relativistic effects like time dilation, so apparently $\gamma \approx 1$, and v << 1, i.e., the speed c must be very large when expressed in meters per second. By setting c equal to 1, we have chosen a the distance unit that is extremely long in proportion to the time unit. This is an example of the correspondence principle, which states that when a new physical theory, such as relativity, replaces an old one, such as Galilean relativity, it must remain "backward-compatible" with all the experiments that verified the old theory; that is, it must agree with the old theory in the appropriate limit. Despite my coyness, you probably know that the speed of light is also equal to c. It is important to emphasize, however, that light plays no special role in relativity, nor was it necessary to assume the constancy of the speed of light in order to derive the Lorentz transformation; we will in fact prove that photons must travel at c, and that this must be true for any massless particle.

On the other hand, Einstein did originally develop relativity based on a different set of assumptions than our L1-L5. His treatment, given in his 1905 paper "On the electrodynamics of moving bodies," is reproduced later. It starts from the following two postulates:

- **P1** The principle of relativity: ". . . the phenomena of electrodynamics as well as of mechanics possess no properties corresponding to the idea of absolute rest."
- **P2** ". . . light is always propagated in empty space with a definite velocity c which is independent of the state of motion of the emitting body."

Einstein's P1 is essentially the same as our L3 (equivalence of inertial frames). He implicitly assumes something equivalent to our L1 (homogeneity and isotropy of spacetime). In his system, our L5 (relativity of time) is a theorem proved from the axioms P1-P2, whereas in our system, his P2 is a theorem proved from the axioms L1-L5.

2.5: Relativistic Properties of Lorentz Geometry (Part 2)

Example 8
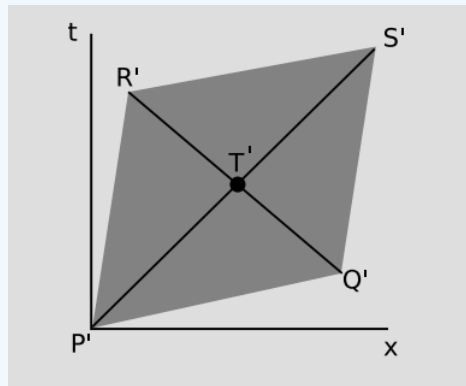
Figure 2.5.5 - Flashes of light travel along P'T' and Q'T'. The observer in this frame of reference judges them to have been emitted at different times, and to have traveled different distances.

Let the intersection of the parallelogram's two diagonals be T in the original (rest) frame, and T' in the Lorentz-boosted frame. An observer at T in the original frame simultaneously detects the passing by of the two flashes of light emitted at P and Q, and since she is positioned at the midpoint of the diagram in space, she infers that P and Q were simultaneous. Since the arrival of both flashes of light at the same point in spacetime is a concrete event, an observer in the Lorentz-boosted frame must agree on their simultaneous arrival. (Simultaneity is well defined as long as no spatial separation is involved.) But the distances traveled by the two flashes in the boosted frame are unequal, and since the speed of light is the same in all cases, the boosted observer infers that they were not emitted simultaneously.

Example 9

A different kind of symmetry is the symmetry between observers. If observer A says observer B's time is slow, shouldn't B say that A's time is fast? This is what would happen if B took a pill that slowed down all his thought processes: to him, the rest of the world would seem faster than normal. But this can't be correct for Lorentz boosts, because it would introduce an asymmetry between observers. There is no preferred, "correct" frame corresponding to the observer who didn't take a pill; either observer can correctly consider himself to be the one who is at rest. It may seem paradoxical that each observer could think that the other was the slow one, but the paradox evaporates when we consider the methods available to A and B for resolving the controversy. They can either (1) send signals back and forth, or (2) get together and compare clocks in person. Signaling doesn't establish one observer as correct and one as incorrect, because as we'll see in the following section, there is a limit to the speed of propagation of signals; either observer ends up being able to explain the other observer's observations by taking into account the finite and changing time required for signals to propagate. Meeting in person requires one or both observers to accelerate, as in the original story of Alice and Betty, and then we are no longer dealing with pure Lorentz frames, which are described by nonaccelerating observers.

Example 10: Einstein's goof

Einstein's original 1905 paper on special relativity, reproduced in Appendix A, contains a famous incorrect prediction, that "a spring-clock at the equator must go more slowly, by a very small amount, than a precisely similar clock situated at one of the poles under otherwise identical conditions". This was a reasonable prediction at the time, but we now know that it was incorrect because it neglected gravitational time dilation. In the description of the Hafele-Keating experiment using atomic clocks aboard airplanes, we saw that both gravity and motion had effects on the rate of flow of time. Earlier we found based on the equivalence principle that the gravitational redshift of an electromagnetic wave is $\frac{\Delta E}{E} = \Delta\phi$ (where c = 1 and ϕ is the gravitational potential gy), and that this could also be interpreted as a gravitational time dilation $\frac{\Delta t}{t} = \Delta\phi$.

The clock at the equator suffers a kinematic time dilation that would tend to cause it to run more slowly than the one at the pole. However, the earth is not a sphere, so the two clocks are at different distances from the earth's center, and the field they inhabit is also not the simple field of a sphere. This suggests that there may be an additional gravitational effect due to

$\Delta\phi \neq 0$. Expanding the Lorentz gamma factor in a Taylor series, we find that the kinematic effect amounts to $\frac{\Delta t}{t} = \gamma - 1 \approx \frac{v^2}{2}$. The mismatch in rates between the two clocks is

$$\frac{\Delta t}{t} \approx \frac{1}{2}v^2 - \Delta\phi, \qquad (2.5.1)$$

where $\Delta\phi = \phi_{equator} - \phi_{pole}$, and a factor of $frac1c^2$ on the right is suppressed because c = 1. But this expression for $\frac{\Delta t}{t}$ vanishes, because the surface of the earth's oceans is in equilibrium, and therefore a mass of water m can be brought from the north pole to the equator, with the change in potential energy $-m\Delta\phi$ being exactly sufficient to supply the necessary kinetic energy $(\frac{1}{2})mv^2$. We therefore find that a change of latitude should have no effect on the rate of a clock, provided that it remains at sea level.

This has been verified experimentally by Alley et al.[9] Alley's group flew atomic clocks from Washington, DC to Thule, Greenland, left them there for four days, and brought them back. The difference between the clocks that went to Greenland and other clocks that stayed in Washington was 38 ± 5 ns, which was consistent with the 35 ± 2 ns effect predicted purely based on kinematic and gravitational time dilation while the planes were in the air. If Einstein's 1905 prediction had been correct, then there would have been an additional difference of 224 ns due to the difference in latitude.

The perfect cancellation of kinematic and gravitational effects is not as fortuitous as it might seem; this is discussed in example 18.

---

[9] C.O. Alley, et al., in NASA Goddard Space Flight Center, Proc. of the 13th Ann. Precise Time and Time Interval (PTTI) Appl. and Planning Meeting, p. 687-724, 1981, available online at http://www.pttimeeting.org/archivemeetings/index9.html

Example 11: GPS

In the GPS system, as in example 10, both gravitational and kinematic time dilation must be considered. Let's determine the directions and relative strengths of the two effects in the case of a GPS satellite.

A radio photon emitted by a GPS satellite gains energy as it falls to the earth's surface, so its energy and frequency are increased by this effect. The observer on the ground, after accounting for all non-relativistic effects such as Doppler shifts and the Sagnac effect, would interpret the frequency shift by saying that time aboard the satellite was flowing more quickly than on the ground.

However, the satellite is also moving at orbital speeds, so there is a Lorentz time dilation effect. According to the observer on earth, this causes time aboard the satellite to flow more slowly than on the ground.

We can therefore see that the two effects are of opposite sign. Which is stronger?

For a satellite in low earth orbit, we would have $fracv^2r = g$, where r is only slightly greater than the radius of the earth. The relative effect on the flow of time is $\gamma - 1 \approx \frac{v^2}{2} = \frac{gr}{2}$. The gravitational effect, approximating g as a constant, is $-gy$, where y is the satellite's altitude above the earth. For such a satellite, the gravitational effect is down by a factor of $\frac{2y}{r}$, so the Lorentz time dilation dominates.

GPS satellites, however, are not in low earth orbit. They orbit at an altitude of about 20,200 km, which is quite a bit greater than the radius of the earth. We therefore expect the gravitational effect to dominate. To confirm this, we need to generalize the equation $\frac{\Delta t}{t} = \Delta\phi$ (with c = 1) from example 10 to the case where g is not a constant. Integrating the equation $\frac{dt}{t} = d\phi$, we find that the time dilation factor is equal to $e^{\Delta\phi}$. When $\Delta\phi$ is small, $e^{\Delta\phi} \approx 1 + \Delta\phi$, and we have a relative effect equal to $\Delta\phi$. The total effect for a GPS satellite is thus (inserting factors of c for calculation with SI units, and using positive signs for blueshifts)

$$\frac{1}{c^2}\left(+\Delta\phi - \frac{v^2}{2}\right) = 5.2 \times 10^{-10} - 0.9 \times 10^{-10}, \qquad (2.5.2)$$

where the first term is gravitational and the second kinematic. A more detailed analysis includes various time-varying effects, but this is the constant part. For this reason, the atomic clocks aboard the satellites are set to a frequency of 10.22999999543 MHz before launching them into orbit; on the average, this is perceived on the ground as 10.23 MHz. A more complete analysis of the general relativity involved in the GPS system can be found in the review article by Ashby.[10]

[10] N. Ashby, "Relativity in the Global Positioning System," http://www.livingreviews.org/lrr-2003-1

Exercise 2.5.1

Self-check: Suppose that positioning a clock at a certain distance from a certain planet produces a fractional change δ in the rate at which time flows. In other words, the time dilation factor is $1 + \delta$. Now suppose that a second, identical planet is brought into the picture, at an equal distance from the clock. The clock is positioned on the line joining the two planets' centers, so that the gravitational field it experiences is zero. Is the fractional time dilation now approximately 0, or approximately 2δ? Why is this only an approximation?

Example 12: Large time dilation

The time dilation effect in the Hafele-Keating experiment was very small. If we want to see a large time dilation effect, we can't do it with something the size of the atomic clocks they used; the kinetic energy would be greater than the total megatonnage of all the world's nuclear arsenals. We can, however, accelerate subatomic particles to speeds at which γ is large. An early, lowprecision experiment of this kind was performed by Rossi and Hall in 1941, using naturally occurring cosmic rays. Figure 2.2.6 shows a 1974 experiment[11] of a similar type which verified the time dilation predicted by relativity to a precision of about one part per thousand.
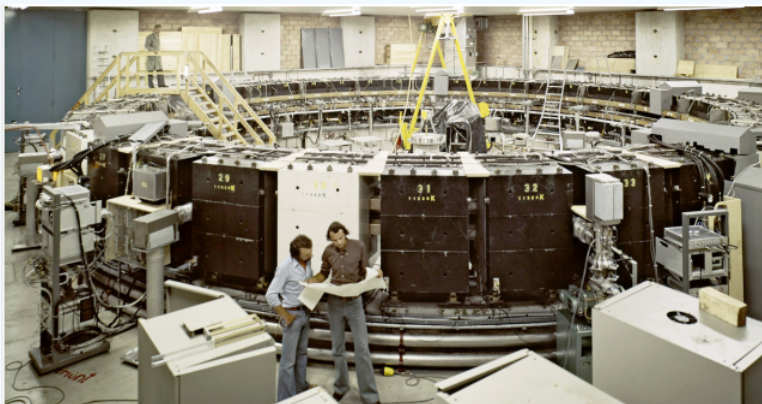
Figure 2.5.6 - Apparatus used for the test of relativistic time dilation described in example 12. The prominent black and white blocks are large magnets surrounding a circular pipe with a vacuum inside. (c) 1974 by CERN.

Muons were produced by an accelerator at CERN, near Geneva. A muon is essentially a heavier version of the electron. Muons undergo radioactive decay, lasting an average of only 2.197 μs before they evaporate into an electron and two neutrinos. The 1974 experiment was actually built in order to measure the magnetic properties of muons, but it produced a high-precision test of time dilation as a byproduct. Because muons have the same electric charge as electrons, they can be trapped using magnetic fields. Muons were injected into the ring shown in Figure 2.2.6, circling around it until they underwent radioactive decay. At the speed at which these muons were traveling, they had $\gamma = 29.33$, so on the average they lasted 29.33 times longer than the normal lifetime. In other words, they were like tiny alarm clocks that self-destructed at a randomly selected time. Figure 2.2.7 shows the number of radioactive decays counted, as a function of the time elapsed after a given stream of muons was injected into the storage ring. The two dashed lines show the rates of decay predicted with and without relativity. The relativistic line is the one that agrees with experiment.
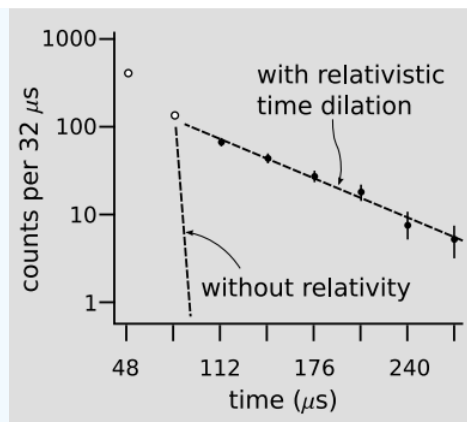
Figure 2.5.7 - Muons accelerated to nearly c undergo radioactive decay much more slowly than they would according to an observer at rest with respect to the muons. The first two data-points (unfilled circles) were subject to large systematic errors.

[11] Bailey at al., Nucl. Phys. B150(1979) 1

Example 13: Time dilation in the Pound-Rebka experiment

In the description of the Pound-Rebka experiment, the quantitative estimation of the frequency shift due to temperature was postponed. Classically, one expects only a broadening of the line, since the Doppler shift is proportional to $\frac{v_{\parallel}}{c}$, where $v_{\parallel}$, the component of the emitting atom's velocity along the line of sight, averages to zero. But relativity tells us to expect that if the emitting atom is moving, its time will flow more slowly, so the frequency of the light it emits will also be systematically shifted downward. This frequency shift should increase with temperature. In other words, the Pound-Rebka experiment was designed as a test of general relativity (the equivalence principle), but this special-relativistic effect is just as strong as the relativistic one, and needed to be accounted for carefully.

In Pound and Rebka's paper describing their experiment,[12] they refer to a preliminary measurement[13] in which they carefully measured this effect, showed that it was consistent with theory, and pointed out that a previous claim by Cranshaw et al. of having measured the gravitational frequency shift was vitiated by their failure to control for the temperature dependence.

It turns out that the full Debye treatment of the lattice vibrations is not really necessary near room temperature, so we'll simplify the thermodynamics. At absolute temperature T, the mean translational kinetic energy of each iron nucleus is $\left(\frac{3}{2}\right)kT$. The velocity is much less than c(= 1), so we can use the nonrelativistic expression for kinetic energy, $K = \left(\frac{1}{2}\right)mv^2$, which gives a mean value for v^2 of $\frac{3kT}{m}$. In the limit of v << 1, time dilation produces a change in frequency by a factor of $\frac{1}{\gamma}$, which differs from unity by approximately $-\frac{v^2}{2}$. The relative time dilation is therefore $-\frac{3kT}{2m}$, or, in metric units, $-\frac{3kT}{2mc^2}$. The vertical scale in figure h contains an arbitrary offset, since Pound and Rebka's measurements were the best absolute measurements to date of the frequency. The predicted slope of $-\frac{3k}{2mc^2}$, however, is not arbitrary. Plugging in 57 atomic mass units for m, we find the slope to be 2.4×10^{-15}, which, as shown in the figure is an excellent approximation (off by only 10%) near room temperature.
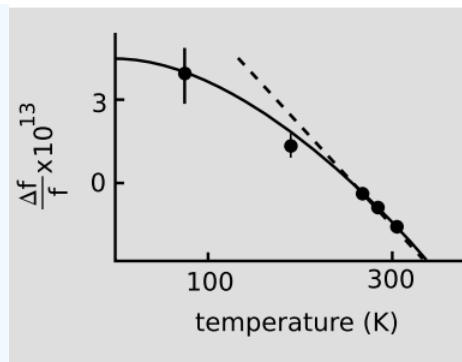
Figure 2.5.8 - The change in the frequency of x-ray photons emitted by ^{57}Fe as a function of temperature, drawn after Pound And Rebka (1960). Dots are experimental measurements. The solid curve is Pound and Rebka's theoretical calculation using the Debye theory of the lattice vibrations with a Debye temperature of 420 degrees C. The dashed line is one with the slope calculated in the text using a simplified treatment of the thermodynamics. There is an arbitrary vertical offset in the experimental data, as well as the theoretical curves.

[12] Phys. Rev. Lett. 4 (1960) 337

[13] Phys. Rev. Lett. 4 (1960) 274

Geodesics and Stationary Action

One way of characterizing geodesics in spacetime is by using an action principle. This is similar to characterizing a geodesic in Euclidean space as a line of minimum length between two points. For a timelike geodesic from event P to event Q, we have a proper time τ. In Lorentz spacetime, this proper time is greater than it would have been for any non-geodesic motion from P to Q. In curved spacetime, we must weaken this statement somewhat. The proper time may not be a global maximum, but it is stationary. Stationarity means that if we vary the curve by some small amount, not moving any part of it by a coordinate distance greater than ϵ, then the change in τ is of order ϵ^2.

For spacelike geodesics in Lorentz spacetime, the proper length is stationary, but small deformations of the curve can either increase or decrease the proper length. The stationary action approach does not work well for lightlike geodesics.

2.6: The Light Cone

Given an event P, we can now classify all the causal relationships in which P can participate. In Newtonian physics, these relationships fell into two classes: P could potentially cause any event that lay in its future, and could have been caused by any event in its past. In a Lorentz spacetime, we have a trichotomy rather than a dichotomy. There is a third class of events that are too far away from P in space, and too close in time, to allow any cause and effect relationship, since causality's maximum velocity is c. Since we're working in units in which $c = 1$, the boundary of this set is formed by the lines with slope ± 1 on a (t, x) plot. This is referred to as the *light cone*, and in the generalization from 1+1 to 3+1 dimensions, it literally becomes a (four-dimensional) cone. The terminology comes from the fact that light happens to travel at c, the maximum speed of cause and effect. If we make a cut through the cone defined by a surface of constant time in P's future, the resulting section is a sphere (analogous to the circle formed by cutting a three-dimensional cone), and this sphere is interpreted as the set of events on which P could have had a causal effect by radiating a light pulse outward in all directions.
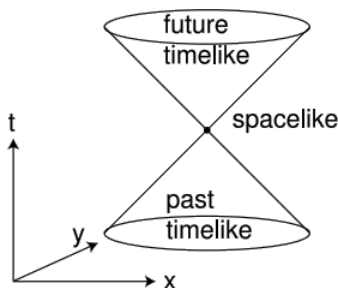
Figure 2.6.1 *: The light cone in 2+1 dimensions.*

Events lying inside one another's light cones are said to have a *timelike* relationship. Events outside each other's light cones are *spacelike* in relation to one another, and in the case where they lie on the surfaces of each other's light cones the term is *lightlike*.

The light cone plays the same role in the Lorentz geometry that the circle plays in Euclidean geometry. The truth or falsehood of propositions in Euclidean geometry remains the same regardless of how we rotate the figures, and this is expressed by Euclid's E3 asserting the existence of circles, which remain invariant under rotation. Similarly, Lorentz boosts preserve light cones and truth of propositions in a Lorentz frame.

Exercise 2.6.1

Self-check: Under what circumstances is the time-ordering of events P and Q preserved under a Lorentz boost?
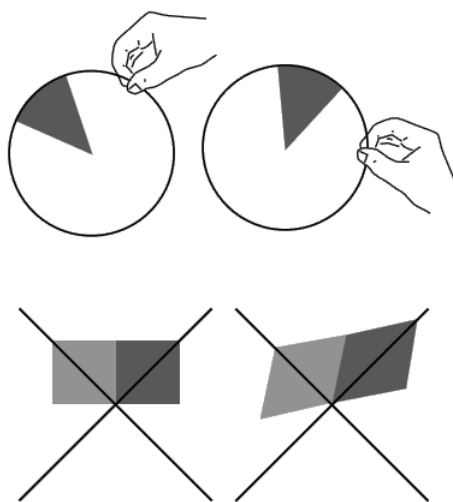
Figure 2.6.2 *: The circle plays a privileged role in Euclidean geometry. When rotated, it stays the same. The pie slice is not invariant as the circle is. A similar privileged place is occupied by the light cone in Lorentz geometry. Under a Lorentz boost, the spacetime parallelograms change, but the light cone doesn't.*

In a uniform Lorentz spacetime, all the light cones line up like soldiers with their axes parallel with one another. When gravity is present, however, this uniformity is disturbed in the vicinity of the masses that constitute the sources. The light cones lying near the sources tip toward the sources. Superimposed on top of this gravitational tipping together, recent observations have demonstrated a systematic tipping-apart effect which becomes significant on cosmological distance scales. The parameter Λ that sets the strength of this effect is known as the *cosmological constant*. The cosmological constant is not related to the presence of any sources (such as negative masses), and can be interpreted instead as a tendency for space to expand over time on its own initiative. In the present era, the cosmological constant has overpowered the gravitation of the universe's mass, causing the expansion of the universe to accelerate.

Exercise 2.6.2: Tipping Light Cones

In the bottom panel of Figure 2.3.3, can an observer look at the properties of the spacetime in her immediate vicinity and tell how much her light cones are tipping, and in which direction? Compare with ⧉ Figure 1.5.9.
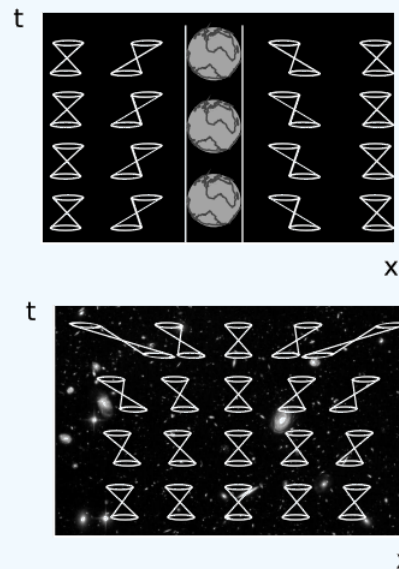
Figure $2.6.3$ - *Light cones tip over for two reasons in general relativity: because of the presence of masses, which have gravitational fields, and because of the cosmological constant. The time and distance scales in the bottom figure are many orders of magnitude greater than those in the top.*

Example 14: A Newtonian Black hole

In the case of a black hole, the light cone tips over so far that the entire future timelike region lies within the black hole. If an observer is present at such an event, then that observer's entire potential future lies within the black hole, not outside it. By expanding on the logical consequences of this statement, we arrive at an example of relativity's proper interpretation as a *theory of causality*, not a theory of objects exerting forces on one another as in Newton's vision of action at a distance, or Lorentz's original ether-drag interpretation of the factor γ, in which length contraction arose from a physical strain imposed on the atoms composing a physical body.

Imagine a black hole from a Newtonian point of view, as proposed in 1783 by geologist John Michell. Setting the escape velocity equal to the speed of light, we find that this will occur for any gravitating spherical body compact enough to have

$$\frac{M}{r} > \frac{c^2}{2G}. \tag{2.6.1}$$

A fully relativistic argument, as given in Section 6.2, agrees on

$$\frac{M}{r} \propto \frac{c^2}{G} \tag{2.6.2}$$

which is fixed by units. The correct unitless factor depends on the definition of r, which is flexible in general relativity.) A flash of light emitted from the surface of such a Newtonian black hole would fall back down like water from a fountain, but it

would nevertheless be possible for physical objects to escape, e.g., if they were lifted out in a bucket dangling from a cable. If the cable is to support its own weight, it must have a tensile strength per unit density of at least $\frac{c^2}{2}$, which is about ten orders of magnitude greater than that of carbon nanotube fibers. (The factor of $\frac{1}{2}$ is not to be taken seriously, since it comes from a nonrelativistic calculation.)
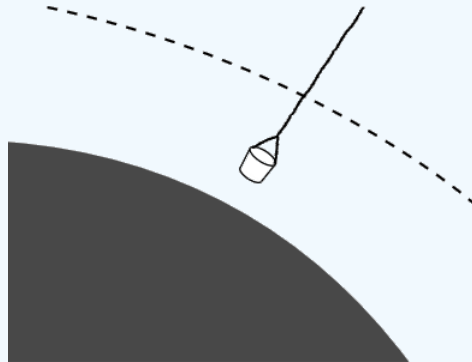
Figure 2.6.4 - *Matter is lifted out of a Newtonian black hole with a bucket. The dashed line represents the point at which the escape velocity equals the speed of light.*

The cause-and-effect interpretation of relativity tells us that this Newtonian picture is incorrect. A physical object that approaches to within a distance r of a concentration of mass M, with $\frac{M}{r}$ sufficiently large, has no causal future lying at larger values of r. The conclusion is that there is a limit on the tensile strength of any substance, imposed purely by general relativity, and we can state this limit without having to know anything about the physical nature of the interatomic forces. A more complete treatment of the tension in the rope is given in example 5. Cf. also problem 4 and section 3.5, as well as some references given in the remark following problem 4.

Velocity Addition

In classical physics, velocities add in relative motion. For example, if a boat moves relative to a river, and the river moves relative to the land, then the boat's velocity relative to the land is found by vector addition. This linear behavior cannot hold relativistically. For example, if a spaceship is moving at 0.60c relative to the earth, and it launches a probe at $0.60c$ relative to itself, we can't have the probe moving at $1.20c$ relative to the earth, because this would be greater than the maximum speed of cause and effect, c. To see how to add velocities relativistically, we start by rewriting the Lorentz transformation as the matrix

$$\begin{pmatrix} \cosh\eta & \sinh\eta \\ \sinh\eta & \cosh\eta \end{pmatrix}, \tag{2.6.3}$$

where

$$\eta = \tanh^{-1} v \tag{2.6.4}$$

is called the *rapidity*. We are guaranteed that the matrix can be written in this form, because its area-preserving property says that the determinant equals 1, and $\cosh^2\eta - \sinh^2\eta = 1$ is an identity of the hyperbolic trig functions. It is now straightforward to verify that multiplication of two matrices of this form gives a third matrix that is also of this form, with

$$\eta = \eta_1 + \eta_2. \tag{2.6.5}$$

In other words, rapidities add linearly; velocities don't. In the example of the spaceship and the probe, the rapidities add as

$$\tanh^{-1} 0.60 + \tanh^{-1} 0.60 = 0.693 + 0.693 = 1.386 \tag{2.6.6}$$

giving the probe a velocity of tanh 1.386 = 0.88 relative to the earth. Any number of velocities can be added in this way,

$$\eta_1 + \eta_2 + \ldots + \eta_n. \tag{2.6.7}$$

Exercise 2.6.3: Rapidity

Interpret the asymptotes of the graph in Figure 2.3.5.
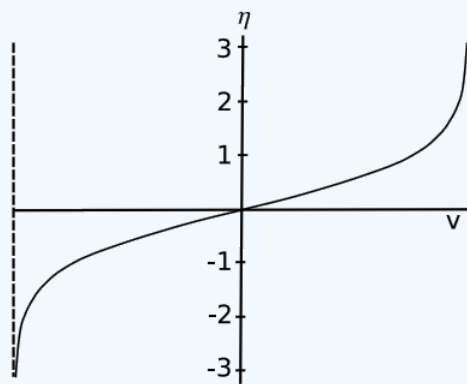
Figure 2.6.5 - *The rapidity, $\eta = \tanh^{-1} v$, as a function of v.*

A difficult philosophical question is whether the time dilation and length contractions predicted by relativity are "real." This depends, of course, on what one means by "real." They are frame-dependent, i.e., observers in different frames of reference disagree about them. But this doesn't tell us much about their reality, since velocities are frame-dependent in Newtonian mechanics, but nobody worries about whether velocities are real. John Bell (1928-1990) proposed the following thought experiment to physicists in the CERN cafeteria, and found that nearly all of them got it wrong. He took this as evidence that their intuitions had been misguided by the standard way of approaching this question of the reality of Lorentz contractions.

Let spaceships A and B accelerate as shown in Figure 2.3.6 along a straight line. Observer C does not accelerate. The accelerations, as judged by C, are constant, and equal for the two ships. Each ship is equipped with a yard-arm, and a thread is tied between the two arms. Does the thread break, due to Lorentz contraction? (We assume that the acceleration is gentle enough that the thread does not break simply because of its own inertia.)
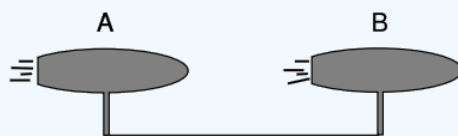
Figure 2.6.6

The popular answer in the CERN cafeteria was that the thread would not break, the reasoning being that Lorentz contraction is a frame-dependent effect, and no such contraction would be observed in A and B's frame. The ships maintain a constant distance from one another, so C merely disagrees with A and B about the length of the thread, as well as other lengths like the lengths of the spaceships.

The error in this reasoning is that the accelerations of A and B were specified to be equal and constant in C's frame, not in A and B's. Bell's interpretation is that the frame-dependence is a distraction, that Lorentz contraction is in some sense a real effect, and that it is therefore immediately clear that the thread must break, without even having to bother going into any other frame. To convince his peers in the cafeteria, however, Bell presumably needed to satisfy them as to the specific errors in their reasoning, and this requires that we consider the frame-dependence explicitly.

We can first see that it is impossible, in general, for different observers to agree about what is meant by constant acceleration. Suppose that A and B agree with C about the constancy of their acceleration. Then A and B experience a voyage in which the rapidities of the stars around them (and of observer C) increase linearly with time. As the rapidity approaches infinity, both C and the stars approach the speed of light. But since A and C agree on the magnitude of their velocity relative to one another, this means that A's velocity as measured by C must approach c, and this contradicts the premise that C observes constant acceleration for both ships. Therefore A and B do not consider their own accelerations to be constant.

A and B do not agree with C about simultaneity, and since they also do not agree that their accelerations are constant, they do not consider their own accelerations to be equal at a given moment of time. Therefore the string changes its length, and this is consistent with Bell's original, simple answer, which did not require comparing different frames of reference. To establish that the string comes under tension, rather than going slack, we can apply the equivalence principle. By the equivalence principle,

any experiments done by A and B give the same results as if they were immersed in a gravitational field. The leading ship B sees A as experiencing a gravitational time dilation. According to B, the slowpoke A isn't accelerating as rapidly as it should, causing the string to break.

These ideas are closely related to the fact that general relativity does not admit any spacetime that can be interpreted as a uniform gravitational field (see problem 7).

Logic

The trichotomous classification of causal relationships has interesting logical implications. In classical Aristotelian logic, every proposition is either true or false, but not both, and given propositions p and q, we can form propositions such as p ∧ q (both p and q) or p ∨ q (either p or q). Propositions about physical phenomena can only be verified by observation. Let p be the statement that a certain observation carried out at event P gives a certain result, and similarly for q at Q. If PQ is spacelike, then the truth or falsehood of p ∧ q cannot be checked by physically traveling to P and Q, because no observer would be able to attend both events. The truth-value of p ∧ q is unknown to any observer in the universe until a certain time, at which the relevant information has been able to propagate back and forth. What if P and Q lie inside two different black holes? Then the truth-value of p ∧ q can never be determined by any observer. Another example is the case in which P and Q are separated by such a great distance that, due to the accelerating expansion of the universe, their future light cones do not overlap.

We conclude that Aristotelian logic cannot be appropriately applied to relativistic observation in this way. Some workers attempting to construct a quantum-mechanical theory of gravity have suggested an even more radically observer-dependent logic, in which different observers may contradict one another on the truth-value of a single proposition p_1, unless they agree in advance on the list p_2, p_3, . . . of all the other propositions that they intend to test as well. We'll return to these questions later.

2.7: Experimental Tests of Lorentz Geometry

We've already seen, in section 1.2, a variety of evidence for the nonclassical behavior of spacetime. We're now in a position to discuss tests of relativity more quantitatively. An up-to-date review of such tests is given by Mattingly.[14]

One such test is that relativity requires the speed of light to be the same in all frames of reference, for the following reasons. Compare with the speed of sound in air. The speed of sound is not the same in all frames of reference, because the wave propagates at a fixed speed relative to the air. An observer at who is moving relative to the air will measure a different speed of sound. Light, on the other hand, isn't a vibration of any physical medium. Maxwell's equations predict a definite value for the speed of light, regardless of the motion of the source. This speed also can't be relative to any medium. If the speed of light isn't fixed relative to the source, and isn't fixed relative to a medium, then it must be fixed relative to anything at all. The only speed in relativity that is equal in all frames of reference is c, so light must propagate at c. We will see later that there is a deeper reason for this; relativity requires that any massless particle propagate at c. The requirement of v = c for massless particles is so intimately hard-wired into the structure of relativity that any violation of it, no matter how tiny, would be of great interest. Essentially, such a violation would disprove Lorentz invariance, i.e., the invariance of the laws of physics under Lorentz transformations. There are two types of tests we could do: (1) test whether photons of all energies travel at the same speed, i.e., whether the vacuum is dispersive; (2) test whether observers in all frames of reference measure the same speed of light.

Dispersion of the Vacuum

Some candidate quantum-mechanical theories of gravity, such as loop quantum gravity, predict a granular structure for spacetime at the Planck scale, $\sqrt{\frac{\hbar G}{c^3}} = 10^{-35}$ m, which one could imagine might lead to deviations from v = 1 that would become more and more significant for photons with wavelengths getting closer and closer to that scale. Lorentz-invariance would then be an approximation valid only at large scales. It turns out that the state of the art in loop quantum gravity is not yet sufficient to say whether or not such an effect should exist.

Presently the best experimental tests of the invariance of the speed of light with respect to wavelength come from astronomical observations of gamma-ray bursts, which are sudden outpourings of high-energy photons, believed to originate from a supernova explosion in another galaxy. One such observation, in 2009,[15] collected photons from such a burst, with a duration of 2 seconds, indicating that the propagation time of all the photons differed by no more than 2 seconds out of a total time in flight on the order of ten billion years, or about one part in 10^{17}! A single superlative photon in the burst had an energy of 31 GeV, and its arrival within the same 2-second time window demonstrates Lorentz invariance over a vast range of photon energies, contradicting heuristic estimates that had been made by some researchers in loop quantum gravity.
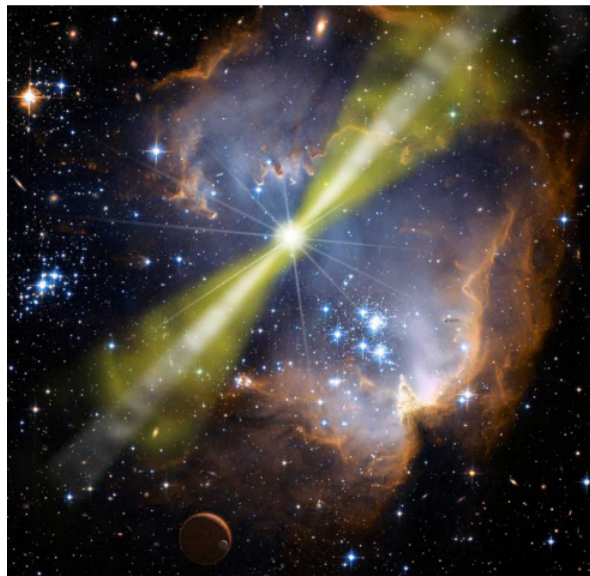
Figure 2.7.1 - An artist's conception of a gamma-ray burst, resulting from a supernova explosion.

Observer-independence of c

The constancy of the speed of light for observers in all frames of reference was originally detected in 1887 when Michelson and Morley set up a clever apparatus to measure any difference in the speed of light beams traveling east-west and north-south. The motion of the earth around the sun at 110,000 km/hour (about 0.01% of the speed of light) is to our west during the day. Michelson and Morley believed that light was a vibration of a physical medium, the ether, so they expected that the speed of light would be a fixed value relative to the ether. As the earth moved through the ether, they thought they would observe an effect on the velocity of light along an east-west line. For instance, if they released a beam of light in a westward direction during the day, they expected that it would move away from them at less than the normal speed because the earth was chasing it through the ether. They were surprised when they found that the expected 0.01% change in the speed of light did not occur.
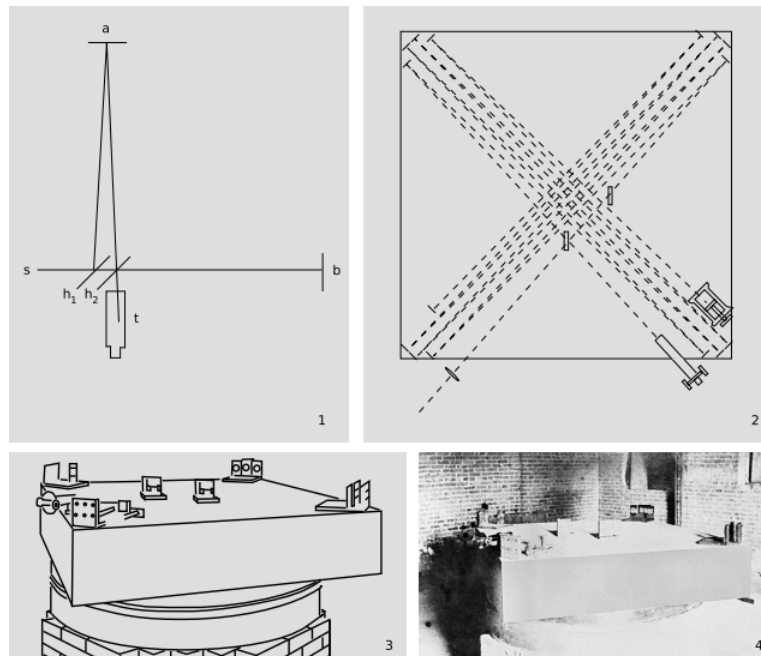
Figure 2.7.2 - The Michelson-Morley experiment, shown in photographs, and drawings from the original 1887 paper. 1. A simplified drawing of the apparatus. A beam of light from the source, s, is partially reflected and partially transmitted by the half-silvered mirror h1. The two half-intensity parts of the beam are reflected by the mirrors at a and b, reunited, and observed in the telescope, t. If the earth's surface was supposed to be moving through the ether, then the times taken by the two light waves to pass through the moving ether would be unequal, and the resulting time lag would be detectable by observing the interference between the waves when they were reunited. 2. In the real apparatus, the light beams were reflected multiple times. The effective length of each arm was increased to 11 meters, which greatly improved its sensitivity to the small expected difference in the speed of light. 3. In an earlier version of the experiment, they had run into problems with its "extreme sensitiveness to vibration," which was "so great that it was impossible to see the interference fringes except at brief intervals . . . even at two o'clock in the morning." They therefore mounted the whole thing on a massive stone floating in a pool of mercury, which also made it possible to rotate it easily. 4. A photo of the apparatus. Note that it is underground, in a room with solid brick walls.

Although the Michelson-Morley experiment was nearly two decades in the past by the time Einstein published his first paper on relativity in 1905, and Einstein did know about it,[16] it's unclear how much it influenced him. Michelson and Morley themselves were uncertain about whether the result was to be trusted, or whether systematic and random errors were masking a real effect from the ether. There were a variety of competing theories, each of which could claim some support from the shaky data. Some physicists believed that the ether could be dragged along by matter moving through it, which inspired variations on the experiment that were conducted on mountaintops in thin-walled buildings, (figure), or with one arm of the apparatus out in the open, and the other surrounded by massive lead walls. In the standard sanitized textbook version of the history of science, every scientist does his experiments without any preconceived notions about the truth, and any disagreement is quickly settled by a definitive experiment. In reality, this period of confusion about the Michelson-Morley experiment lasted for four decades, and a few reputable skeptics, including Miller, continued to believe that Einstein was wrong, and kept trying different variations of the experiment as late as the 1920's. Most of the remaining doubters were convinced by an extremely precise version of the experiment performed by Joos in 1930, although you can still find kooks on the internet who insist that Miller was right, and that there was a vast conspiracy to cover up his results.

Figure 2.7.3 - Dayton Miller thought that the result of the Michelson-Morley experiment could be explained because the ether had been pulled along by the dirt, and the walls of the laboratory. This motivated him to carry out a series of experiments at the top of Mount Wilson, in a building with thin walls.

Before Einstein, some physicists who did believe the negative result of the Michelson-Morley experiment came up with explanations that preserved the ether. In the period from 1889 to 1895, both Lorentz and George FitzGerald suggested that the negative result of the Michelson-Morley experiment could be explained if the earth, and every physical object on its surface, was contracted slightly by the strain of the earth's motion through the ether. Thus although Lorentz developed all the mathematics of Lorentz frames, and got them named after himself, he got the interpretation wrong.

Lorentz Violation by Gravitational Forces

The tests described in earlier both involve the behavior of light, i.e., they test whether or not electromagnetism really has the exact Lorentz-invariant behavior contained implicitly in Maxwell's equations. In the jargon of the field, they test Lorentz invariance in the "photon sector." Since relativity is a theory of gravity, it is natural to ask whether the Lorentz invariance holds for gravitational forces as well as electromagnetic ones. If Lorentz invariance is violated by gravity, then the strength of gravitational forces might depend on the observer's motion through space, relative to some fixed reference frame analogous to that of the ether. Historically, gravitational Lorentz violations have been much more difficult to test, since gravitational forces are so weak, and the first high-precision data were obtained by Nordtvedt and Will in 1957, 70 years after Michelson and Morley. Nordtvedt and Will measured the strength of the earth's gravitational field as a function of time, and found that it did not vary on a 24-hour cycle with the earth's rotation, once tidal effects had been accounted for. Further constraints come from data on the moon's orbit obtained by reflecting laser beams from a mirror left behind by the Apollo astronauts.

A recent high-precision laboratory experiment was done in 2009 by Chung et al.[17] They constructed an interferometer in a vertical plane that is conceptually similar to a Michelson interferometer, except that it uses cesium atoms rather than photons. That is, the light waves of the Michelson-Morley experiment are replaced by quantum-mechanical matter waves. The roles of the half-silvered and fully silvered mirrors are filled by lasers, which kick the atoms electromagnetically. Each atom's wavefunction is split into two parts, which travel by two different paths through spacetime, eventually reuniting and interfering. The result is a measurement of g to about one part per billion. The results, shown in figure d, put a strict limit on violations of Lorentz geometry by gravity.
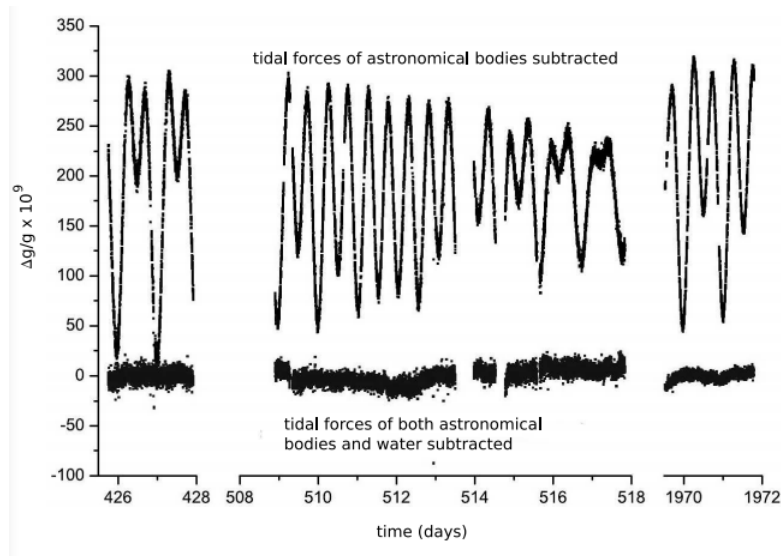
Figure 2.7.4 - The results of the measurement of g by Chung et al. The experiment was done on the Stanford University campus, surrounded by the Pacific ocean and San Francisco Bay, so it was subject to varying gravitational from both astronomical bodies and the rising and falling ocean tides. Once both of these effects are subtracted out of the data, there is no Lorentz-violating variation in g due to the earth's motion through space. Note that the data are broken up into three periods, with gaps of three months and four years separating them. (c) APS, used under the U.S. fair use exception to copyright.
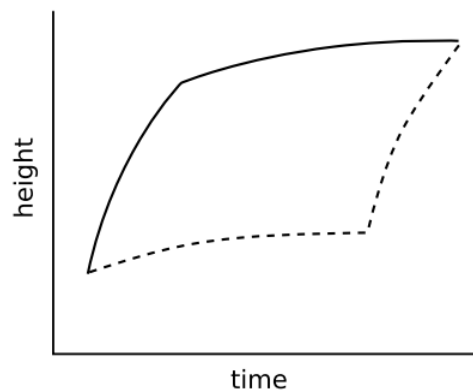
Figure 2.7.5 - The matter interferometer used by Chung et al. Each atom's wavefunction is split into two parts, which travel along two different paths (solid and dashed lines).

References

[14] livingreviews.org/lrr-2005-5

[15] http://arxiv.org/abs/0908.1832

[16] J. van Dongen, http://arxiv.org/abs/0908.1545

[17] arxiv.org/abs/0905.1929

2.8: Three Spatial Dimensions (Part 1)

New and nontrivial phenomena arise when we generalize from 1+1 dimensions to 3+1.

Lorentz Boosts in Three Dimensions

How does a Lorentz boost along one axis, say x, affect the other two spatial coordinates y and z?

First, we can rule out the possibility that such a transformation could have various terms such as $t' = \ldots + (\ldots)y + \ldots$. For example, if the t coefficient was positive for $v > 0$, then the laws of physics would be different from the laws that applied in a universe where the y or t axis was inverted, but this would violate parity or timereversal symmetry. This establishes that observers in the two frames agree on the directions of the y and z axes and on simultaneity along those axes when they coincide.

Now suppose that two observers, in motion relative to one another along the x axis, each carry a stick, represented by line segments AB and CD, oriented along the y axis, such that the bases of the sticks A and C coincide at some time. Due to the vanishing of the types of terms in the transformation referred to above, they agree that B and D are collinear with A (and C) at this time. Then by O3 and O4, either B lies between A and D, D lies between A and B, or B=D. That is, they must agree whether the sticks are equal in length or, if not, then on whose is longer. This would violate L1, isotropy of space, since it would distinguish +x from −x.

Another simple way to obtain this result is as follows. We have already proved that area in the (t, x) plane is preserved. The same proof applies to volume in the spaces (t, x, y) and (t, x, z), hence lengths in the y and z directions are preserved. (The proof does not apply to volume in, e.g., (x, y, z) space, because the x transformation depends on t, and therefore if we are given a region in (x, y, z), we do not have enough information to say how it will change under a Lorentz boost.)

The complete form of the transformation L($v\hat{x}$), a Lorentz boost along the x axis with velocity v, is therefore:

$$t' = \gamma t + v\gamma x$$
$$x' = v\gamma t + \gamma x$$
$$y' = y$$
$$z' = z$$

Based on the trivial nature of this generalization, it might seem as though no qualitatively new considerations would arise in 3+1 dimensions as compared with 1+1. To see that this is not the case, consider Figure 2.8.1. A boost along the x axis tangles up the x and t coordinates. A y-boost mingles y and t. Therefore consecutive boosts along x and y can cause x and y to mix. The result, as we'll see in more detail below, is that two consecutive boosts along non-collinear axes are not equivalent to a single boost; they are equivalent to a boost plus a spatial rotation. The remainder of this section discusses this effect, known as *Thomas precession*, in more detail; it can be omitted on a first reading.
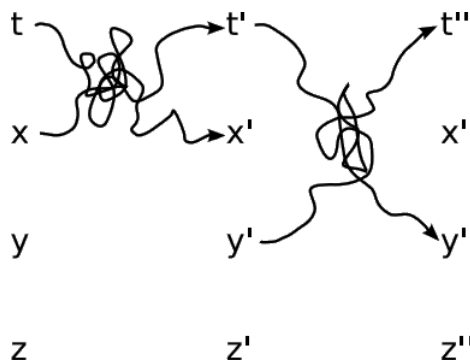
Figure 2.8.1 - A boost along x followed by a boost along y results in tangling up of the x and y coordinates, so the result is not just a boost but a boost plus a rotation.

Exercise 2.8.1

Apply similar reasoning to a Galilean boost.

Gyroscopes and the Equivalence Principle

To see how this mathematical fact would play out as a physical effect, we need to consider how to make a physical manifestation of the concept of a direction in space.

In two space dimensions, we can construct a ring laser, $\frac{b}{1}$, which in its simplest incarnation is a closed loop of optical fiber with a bidirectional laser inserted in one place (Figure $2.8.2$). Coherent light traverses the loop simultaneously in both directions, interfering in a beat pattern, which can be observed by sampling the light at some point along the loop's circumference. If the loop is rotated in its own plane, the interference pattern is altered, because the beamsampling device is in a different place, and the path lengths traveled by the two beams has been altered. This phase shift is called the *Sagnac effect*, after M. Georges Sagnac, who observed the effect in 1913 and interpreted it, incorrectly, as evidence for the existence of the aether.[18] The loop senses its own angular velocity relative to an inertial reference frame. If we transport the loop while always carefully adjusting its orientation so as to prevent phase shifts, then its orientation has been preserved. The atomic clocks used in the Hafele-Keating atomic-clock experiment were subject to the Sagnac effect.
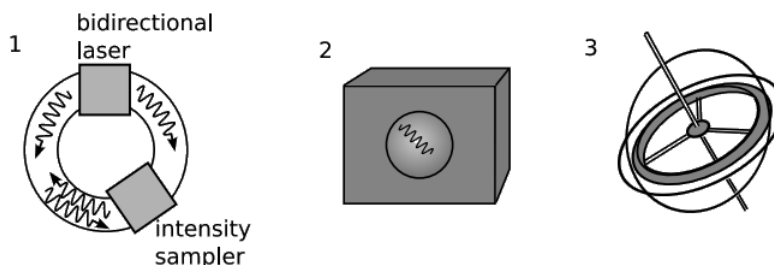
Figure 2.8.2 - Inertial devices for maintaining a direction in space: 1. A ring laser. 2. The photon in a perfectly reflective spherical cavity. 3. A gyroscope.
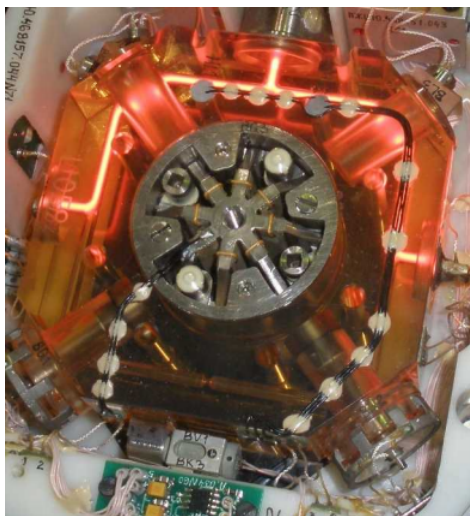
Figure 2.8.3 - A ring laser gyroscope built for use in inertial guidance of aircraft.

In three spatial dimensions, we could build a spherical cavity with a reflective inner surface, and release a photon inside, $\frac{b}{2}$.

In reality, the photon-in-a-cavity is not very practical. The photon would eventually be absorbed or scattered, and it would also be difficult to accurately initialize the device and read it out later. A more practical tool is a gyroscope (Figure $2.8.3$). For example, one of the classic tests of general relativity is the 2007 Gravity Probe B experiment (discussed in detail in section 5.5 and section 6.2), in which four gyroscopes aboard a satellite were observed to precess due to special- and general-relativistic effects.

The gyroscope, however, is not so obviously a literal implementation of our basic concept of a direction. How, then, can we be sure that its behavior is equivalent to that of the photon-in-a-cavity? We could, for example, carry out a complete mathematical development of the angular momentum vector in relativity.[19] The equivalence principle, however, allows us to bypass such technical details. Suppose that we seal the two devices inside black boxes, with identical external control panels for initializing them and reading them out. We initialize them identically, and then transport them along side-by-side world-lines. Nonrelativistically, both the mechanical gyroscope and the photon-gyroscope would maintain absolute, fixed directions in space. Relativistically, they will not necessarily maintain their orientations. For example, we've already seen that there are reasons to

expect that their orientations will change if they are subjected to accelerations that are not all along the same line. Because relativity is a geometrical theory of spacetime, this difference between the classical and relativistic behavior must be determinable from purely geometrical considerations, such as the shape of the world-line. If it depended on something else, then we could conceivably see a disagreement in the outputs of the two instruments, but this would violate the equivalence principle.

> **Note**
>
> This is done, for example, in Misner, Thorne, and Wheeler, *Gravitation*, pp. 157-159.

Suppose there were such a discrepancy. That discrepancy would be a physically measurable property of the spacetime region through which the two gyroscopes had been transported. The effect would have a certain magnitude and direction, so by collecting enough data we could map it out as vector field covering that region of spacetime. This field evidently causes material particles to accelerate, since it has an effect on the mechanical gyroscope. Roughly speaking (the reasoning will be filled in more rigorously in Section 4.4), the fact that this field acts differently on the two gyroscopes is like getting a non-null result from an Eötvös experiment, and it therefore violates the equivalence principle. We conclude that gyroscopes $\frac{b}{2}$ and $\frac{b}{3}$ are equivalent. In other words, there can only be one uniquely defined notion of direction, and the details of how it is implemented are irrelevant.

References

[18] Comptes rendus de l'Acad´emie des science 157 (1913) 708

2.9: Three Spatial Dimensions (Part 2)

Lorentz Boosts Causing Rotations

As a quantitative example, consider the following thought experiment. Put a gyroscope in a box, and send the box around the square path shown in figure d at constant speed. The gyroscope defines a local coordinate system, which according to classical physics would maintain its orientation. At each corner of the square, the box has its velocity vector changed abruptly, as represented by the hammer. We assume that the hits with the hammer are transmitted to the gyroscope at its center of mass, so that they do not result in any torque. Nonrelativistically, if the set of gyroscopes travels once around the square, it should end up at the same place and in the same orientation, so that the coordinate system it defines is identical with the original one.
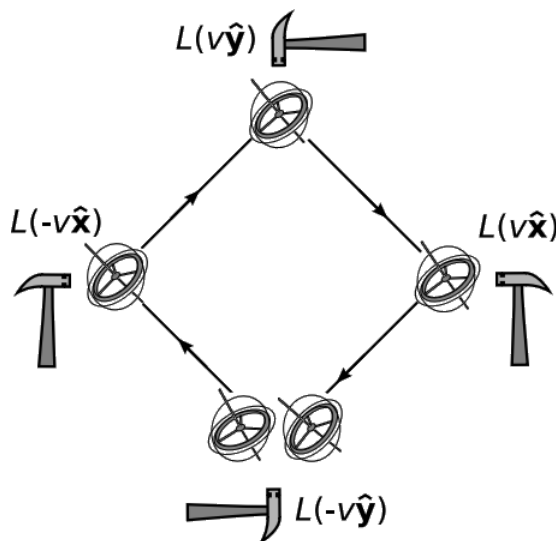
Figure 2.9.4: Nonrelativistically, the gyroscope should not rotate as long as the forces from the hammer are all transmitted to it at its center of mass.

For notation, let L($v\hat{x}$) indicate the boost along the x axis described by the transformation earlier. This is a transformation that changes to a frame of reference moving in the *negative* x direction compared to the original frame. A particle considered to be at rest in the original frame is described in the new frame as moving in the *positive* x direction. Applying such an L to a vector **p**, we calculate L**p**, which gives the coordinates of the event as measured in the new frame. An expression like ML**p** is equivalent by associativity to M(L**p**), i.e., ML represents applying L first, and then M.

In this notation, the hammer strikes can be represented by a series of four Lorentz boosts,

$$T = L(v\hat{x})L(v\hat{y})L(-v\hat{x})L(-v\hat{y}), \tag{2.9.1}$$

where we assume that the square has negligible size, so that all four Lorentz boosts act in a way that preserves the origin of the coordinate systems. (We have no convenient way in our notation L(. . .) to describe a transformation that does not preserve the origin.) The first transformation, L($-v\hat{y}$), changes coordinates measured by the original gyroscope-defined frame to new coordinates measured by the new gyroscope-defined frame, after the box has been accelerated in the positive y direction.

calculation of T matrices (Maxima Software)

The calculation of T is messy, and to be honest, I made a series of mistakes when I tried to crank it out by hand. Calculations in relativity have a reputation for being like this. Figure 2.9.5 shows a page from one of Einstein's notebooks, written in fountain pen around 1913. At the bottom of the page, he wrote "zu umstaendlich," meaning "too involved."

Figure 2.9.5: *A page from one of Einstein's notebooks.*

Luckily we live in an era in which this sort of thing can be handled by computers. Starting at this point in the book, I will take appropriate opportunities to demonstrate how to use the free and open-source computer algebra system Maxima to keep complicated calculations manageable. The following Maxima program calculates a particular element of the matrix T.

```
1   /* For convenience, define gamma in terms of v: */
2   gamma:1/sqrt(1-v*v);
3   /* Define Lx as L(x-hat), Lmx as L(-x-hat), etc.: */
4   Lx:matrix([gamma,    gamma*v,  0],
5             [gamma*v,  gamma,    0],
6             [0,        0,        1]);
7   Ly:matrix([gamma,    0,        gamma*v],
8             [0,        1,        0],
9             [gamma*v,  0,        gamma]);
10  Lmx:matrix([gamma,   -gamma*v, 0],
11             [-gamma*v, gamma,    0],
12             [0,        0,        1]);
13  Lmy:matrix([gamma,   0,        -gamma*v],
14             [0,        1,        0],
15             [-gamma*v, 0,        gamma]);
16  /* Calculate the product of the four matrices: */
17  T:Lx.Ly.Lmx.Lmy;
18  /* Define a column vector along the x axis: */
19  P:matrix([0],[1],[0]);
20  /* Find the result of T acting on this vector,
21     expressed as a Taylor series to second order in v: */
22  taylor(T.P,v,0,2);
```

Statements are terminated by semicolons, and comments are written like /* ... */ On line 2, we see a symbolic definition of the symbol gamma in terms of the symbol v. The colon means "is defined as." Line 2 does not mean, as it would in most programming languages, to take a stored numerical value of v and use it to calculate a numerical value of γ. In fact, v does not have a numerical value defined at this point, nor will it ever have a numerical value defined for it throughout this program. Line 2 simply means that whenever Maxima encounters the symbol gamma, it should take it as an abbreviation for the symbol 1/sqrt(1-v*v). Lines 5-16 define some 3×3 matrices that represent the L transformations. The basis is $\hat{t}, \hat{x}, \hat{y}$. Line 18 calculates the product of the four matrices; the dots represent matrix multiplication. Line 23 defines a vector along the x axis, expressed as a column matrix (three rows of one column each) so that Maxima will know how to operate on it using matrix multiplication by T.

Finally line 26 outputs the result of T acting on P:

```
19                                        [  0 + . . .   ]
20                                        [              ]
21    (%o9)/T/                            [  1 + . . .   ]
22                                        [              ]
23                                        [    2         ]
24                                        [ - v  + . . . ]
```

I've omitted some output generated automatically from the earlier steps in the computation. The (%o9) indicates that this is Maxima's output from the ninth and final step.

In other words,

$$T\begin{pmatrix}0\\1\\0\end{pmatrix} = \begin{pmatrix}0\\1\\-v^2\end{pmatrix} + \dots, \tag{2.9.2}$$

where . . . represents higher-order terms in v. Suppose that we use the initial frame of reference, before T is applied, to determine that a particular reference point, such as a distant star, is along the x axis. Applying T, we get a new vector TP, which we find has a nonvanishing y component approximately equal to $-v^2$. This result is entirely unexpected classically. It tells us that the gyroscope, rather than maintaining its original orientation as it would have done classically, has rotated slightly. It has precessed in the counterclockwise direction in the x−y plane, so that the direction to the star, as measured in the coordinate system defined by the gyroscope, appears to have rotated clockwise. As the box moved clockwise around the square, the gyroscope has apparently rotated by a counterclockwise angle $\chi \approx v^2$ about the z axis. We can see that this is a purely relativistic effect, since for v << 1 the effect is small. For historical reasons discussed later, this phenomenon is referred to as the Thomas precession.

The particular features of this square geometry are not necessary. I chose them so that (1) the boosts would be along the Cartesian axes, so that we would be able to write them down easily; (2) it is clear that the effect doesn't arise from any asymmetric treatment of the spatial axes; and (3) the change in the orientation of the gyroscope can be measured at the same point in space, e.g., by comparing it with a twin gyroscope that stays at home. In general:

A gyroscope transported around a closed loop in flat spacetime changes its orientation compared with one that is not accelerated.

This is a purely relativistic effect, since a Newtonian gyroscope does not change its axis of rotation unless subjected to a torque; if the boosts are accomplished by forces that act at the gyroscope's center of mass, then there is no nonrelativistic explanation for the effect.

The effect can occur in the absence of any gravitational fields. That is, this is a phenomenon of special relativity.

The composition of two or more Lorentz boosts along different axes is not equivalent to a single boost; it is equivalent to a boost plus a spatial rotation.

Lorentz boosts do not commute, i.e., it makes a difference what order we perform them in. Even if there is almost no time lag between the first boost and the second, the order of the boosts matters. If we had applied the boosts in the opposite order, the handedness of the effect would have been reversed.

Exercise 2.9.2

Self-check: If Lorentz boosts did commute, what would be the consequences for the expression L($v\hat{x}$) L($v\hat{y}$) L($-v\hat{x}$) L($-v\hat{y}$)?

The Velocity Disk

Figure 2.5.6 shows a useful way of visualizing the combined effects of boosts and rotations in 2+1 dimensions. The disk depicts all possible states of motion relative to some arbitrarily chosen frame of reference. Lack of motion is represented by the point at the center. A point at distance v from the center represents motion at velocity v in a particular direction in the x − y plane. By drawing little axes at a particular point, we can represent a particular frame of reference: the frame is in motion at some velocity, with its own x and y axes are oriented in a particular way.
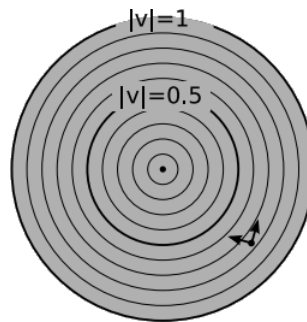
Figure 2.9.6 - *The velocity disk.*

It turns out to be easier to understand the qualitative behavior of our mysterious rotations if we switch from the low-velocity limit to the contrary limit of ultrarelativistic velocities. Suppose we have a rocket-ship with an inertial navigation system consisting of two gyroscopes at right angles to one another. We first accelerate the ship in the y direction, and the acceleration is steady in the sense that it feels constant to observers aboard the ship. Since it is rapidities, not velocities, that add linearly, this means that as an observer aboard the ship reads clock times $\tau_1, \tau_2, \ldots$, all separated by equal intervals $\Delta\tau$, the ship's rapidity changes at a constant rate, $\eta_1, \eta_2, \ldots$.. This results in a series of frames of reference that appear closer and closer together on the diagram as the ship approaches the speed of light, at the edge of the disk. We can start over from the center again and repeat the whole process along the x axis, resulting in a similar succession of frames. In both cases, the boosts are being applied along a single line, so that there is no rotation of the x and y axes.
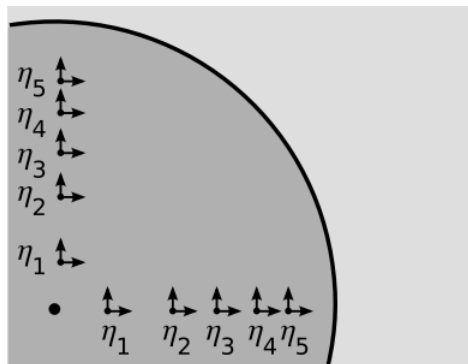
Figure 2.9.7 - *Two excursions in a rocketship: one along the y axis and one along x.*

Now suppose that the ship were to accelerate along a route like the one shown in Figure 2.5.8. It first accelerates along the y axis at a constant rate (again, as judged by its own sensors), until its velocity is very close to the speed of light, A. It then accelerates, again at a self-perceived constant rate and with thrust in a fixed direction as judged by its own gyroscopes, until it is moving at the same ultrarelativistic speed in the x direction, B. Finally, it decelerates in the x direction until it is again at rest, O. This motion traces out a clockwise loop on the velocity disk. The motion in space is also clockwise.
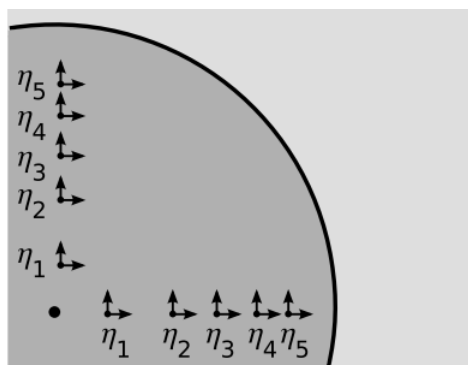
Figure 2.9.8 - *A round-trip involving ultrarelativistic velocities. All three legs are at constant acceleration.*

We might naively think that the middle leg of the trip, from A to B, would be a straight line on the velocity disk, but this can't be the case. First, we know that non-collinear boosts cause rotations. Traveling around a clockwise path causes counterclockwise rotation, and vice-versa. Therefore an observer in the rest frame O sees the ship (and its gyroscopes) as rotating as it moves from A

to B. The ship's trajectory through space is clockwise, so according to O the ship rotates counterclockwise as it goes A to B. The ship is always firing its engines in a fixed direction as judged by its gyroscopes, but according to O the ship is rotating counterclockwise, its thrust is progressively rotating counterclockwise, and therefore its trajectory turns counterclockwise. We conclude that leg AB on the velocity disk is concave, rather than being a straight-line hypotenuse of a triangle OAB.

We can also determine, by the following argument, that leg AB is perpendicular to the edge of the disk where it touches the edge of the disk. In the transformation from frame A to frame O, y coordinates are dilated by a factor of γ, which approaches infinity in the limit we're presently considering. Observers aboard the rocket-ship, occupying frame A, believe that their task is to fire the rocket's engines at an angle of 45 degrees with respect to the y axis, so as to eliminate their velocity with respect to the origin, and simultaneously add an equal amount of velocity in the x direction. This 45-degree angle in frame A, however, is not a 45-degree angle in frame O. From the stern of the ship to its bow we have displacements Δx and Δy, and in the transformation from A to O, Δy is magnified almost infinitely. As perceived in frame O, the ship's orientation is almost exactly antiparallel to the y axis.[21]

> **Note**
>
> Although we will not need any more than this for the purposes of our present analysis, a longer and more detailed discussion by Rhodes and Semon, www.bates.edu/~msemon/RhodesSemonFinal.pdf, Am. J. Phys. 72(7)2004, shows that this type of inertially guided, constant-thrust motion is always represented on the velocity disk by an arc of a circle that is perpendicular to the disk at its edge. (We consider a diameter of the disk to be the limiting case of a circle with infinite radius.)

As the ship travels from A to B, its orientation (as judged in frame O) changes from $-\hat{y}$ to $\hat{x}$. This establishes, in a much more direct fashion, the direction of the Thomas precession: its handedness is contrary to the handedness of the direction of motion. We can now also see something new about the fundamental reason for the effect. It has to do with the fact that observers in different states of motion disagree on spatial angles. Similarly, imagine that you are a two-dimensional being who was told about the existence of a new, third, spatial dimension. You have always believed that the cosine of the angle between two unit vectors **u** and **v** is given by the vector dot product $u_x v_x + u_y v_y$. If you were allowed to explore a two-dimensional projection of a three-dimensional scene, e.g., on the flat screen of a television, it would seem to you as if all the angles had been distorted. You would have no way to interpret the visual conventions of perspective. But once you had learned about the existence of a z axis, you would realize that these angular distortions were happening because of rotations out of the x−y plane. Such rotations really conserve the quantity $u_x v_x + u_y v_y + u_z v_z$; only because you were ignoring the $u_z v_z$ term did it seem that angles were not being preserved. Similarly, the generalization from three Euclidean spatial dimensions to 3+1-dimensional spacetime means that three-dimensional dot products are no longer conserved.
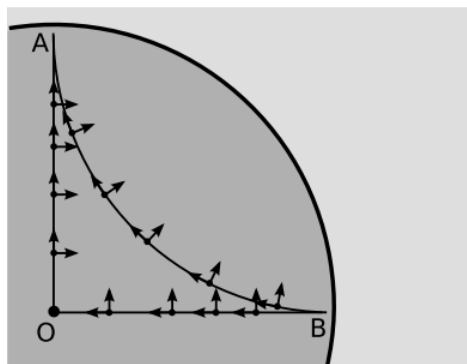
Figure 2.9.9 - *In the limit where A and B are ultrarelativistic velocities, leg AB is perpendicular to the edge of the velocity disk. The result is that the x − y frame determined by the ship's gyroscopes has rotated by 90 degrees by the time it gets home.*

The General Low-v Limit

Let's find the low-v limit of the Thomas precession in general, not just in the highly artificial special case of $\chi \approx v^2$ for the example involving the four hammer hits. To generalize to the case of smooth acceleration, we first note that the rate of precession $\frac{d\chi}{dt}$ must have the following properties.

- It is odd under a reversal of the direction of motion, $\mathbf{v} \rightarrow -\mathbf{v}$. (This corresponds to sending the gyroscope around the square in the opposite direction.)
- It is odd under a reversal of the acceleration due to the second boost, $\mathbf{a} \rightarrow -\mathbf{a}$.

- It is a rotation about the spatial axis perpendicular to the plane of the **v** and **a** vectors, in the opposite direction compared to the handedness of the curving trajectory.
- It is approximately linear in **v** and **a**, for small **v** and **a**.

The only rotationally invariant mathematical operation that has these symmetry properties is the vector cross product, so the rate of precession must be k**a** × **v**, where k > 0 is nearly independent of **v** and **a** for small **v** and **a**.

To pin down the value of k, we need to find a connection between our two results: $\chi \approx v^2$ for the four hammer hits, and $\frac{d\chi}{dt} \approx k\mathbf{a} \times \mathbf{v}$ for smooth acceleration. We can do this by considering the physical significance of areas on the velocity disk. As shown in Figure 2.5.10, the rotation χ due to carrying the velocity around the boundary of a region is additive when adjacent regions are joined together. We can therefore find χ for any region by breaking the region down into elements of area dA and integrating their contributions dχ. What is the relationship between dA and dχ? The velocity disk's structure is nonuniform, in the sense that near the edge of the disk, it takes a larger boost to move a small distance. But we're investigating the low-velocity limit, and in the low-velocity region near the center of the disk, the disk's structure is approximately uniform. We therefore expect that there is an approximately constant proportionality factor relating dA and dχ at low velocities. The example of the hammer corresponds geometrically to a square with area v^2, so we find that this proportionality factor is unity, dA ≈ dχ.
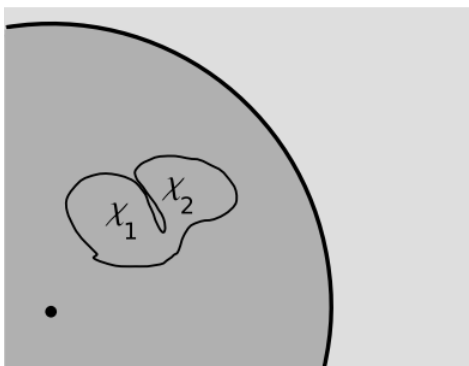
Figure 2.9.10 - *If the crack between the two areas is squashed flat, the two pieces of the path on the interior coincide, and their contributions to the precession cancel out (v → −v, but a → +a, so a × v → −a × v). Therefore the precession χ obtained by going around the outside is equal to the sum $\chi_1 + \chi_2$ of the precessions that would have been obtained by going around the two parts.*

To relate this to smooth acceleration, consider a particle performing circular motion with period T, which has $|\mathbf{a} \times \mathbf{v}| = \frac{2\pi v^2}{T}$. Over one full period of the motion, we have $\chi = \int k|\mathbf{a} \times \mathbf{v}|dt = 2\pi k v^2$, and the particle's velocity vector traces a circle of area A = πv^2 on the velocity disk. Equating A and χ, we find k = $\frac{1}{2}$. The result is that in the limit of low velocities, the rate of rotation is

$$\boldsymbol{\Omega} \approx \frac{1}{2}\mathbf{a} \times \mathbf{v}, \qquad\qquad (2.9.3)$$

where $\boldsymbol{\Omega}$ is the angular velocity vector of the rotation. In the special case of circular motion, this can be written as $\Omega = (\frac{1}{2})v^2\omega$, where $\omega = \frac{2\pi}{T}$ is the angular frequency of the motion.

An Experimental Test: Thomas Precession in Hydrogen

If we want to see this precession effect in real life, we should look for a system in which both v and a are large. An atom is such a system.

The Bohr model, introduced in 1913, marked the first quantitatively successful, if conceptually muddled, description of the atomic energy levels of hydrogen. Continuing to take c = 1, the over-all scale of the energies was calculated to be proportional to $m\alpha^2$, where m is the mass of the electron, and $\alpha = \frac{ke^2}{\hbar} \approx \frac{1}{137}$, known as the fine structure constant, is essentially just a unitless way of expressing the coupling constant for electrical forces. At higher resolution, each excited energy level is found to be split into several sub-levels. The transitions among these close-lying states are in the millimeter region of the microwave spectrum. The energy scale of this fine structure is $\sim m\alpha^4$. This is down by a factor of α^2 compared to the visible-light transitions, hence the name of the constant. Uhlenbeck and Goudsmit showed in 1926 that a splitting on this order of magnitude was to be expected due to the magnetic interaction between the proton and the electron's magnetic moment, oriented along its spin. The effect they calculated, however, was too big by a factor of two.
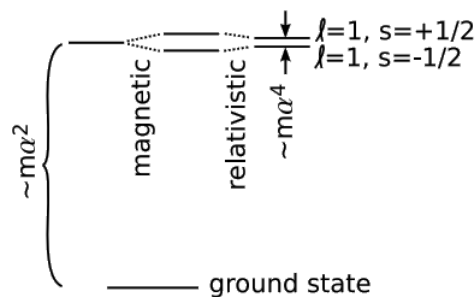
Figure 2.9.11: *States in hydrogen are labeled with their ℓ and s quantum numbers, representing their orbital and spin angular momenta in units of $\hbar$. The state with $s = +\frac{1}{2}$ has its spin angular momentum aligned with its orbital angular momentum, while the $s = -\frac{1}{2}$ state has the two angular momenta in opposite directions. The direction and order of magnitude of the splitting between the two ℓ = 1 states is successfully explained by magnetic interactions with the proton, but the calculated effect is too big by a factor of 2. The relativistic Thomas precession cancels out half of the effect.*

The explanation of the mysterious factor of two had in fact been implicit in a 1916 calculation by Willem de Sitter, one of the first applications of general relativity. De Sitter treated the earth-moon system as a gyroscope, and found the precession of its axis of rotation, which was partly due to the curvature of spacetime and partly due to the type of rotation described earlier in this section. The effect on the motion of the moon was noncumulative, and was only about one meter, which was much too small to be measured at the time. In 1927, however, Llewellyn Thomas applied similar reasoning to the hydrogen atom, with the electron's spin vector playing the role of gyroscope. Since gravity is negligible here, the effect has nothing to do with curvature of spacetime, and Thomas's effect corresponds purely to the special-relativistic part of de Sitter's result. It is simply the rotation described above, with $\Omega = (\frac{1}{2})v^2\omega$. Although Thomas was not the first to calculate it, the effect is known as Thomas precession. Since the electron's spin is $\frac{\hbar}{2}$, the energy splitting is $\pm(\frac{\hbar}{2})\Omega$, depending on whether the electron's spin is in the same direction as its orbital motion, or in the opposite direction. This is less than the atom's gross energy scale ω by a factor of $\frac{v^2}{2}$, which is $\sim \alpha^2$. The Thomas precession cancels out half of the magnetic effect, bringing theory in agreement with experiment.

Uhlenbeck later recalled: "...when I first heard about [the Thomas precession], it seemed unbelievable that a relativistic effect could give a factor of 2 instead of something of order $\frac{v}{c}$ . . . Even the cognoscenti of relativity theory (Einstein included!) were quite surprised."

2.E: Geometry of Flat Spacetime (Exercises)

1. Suppose that we don't yet know the exact form of the Lorentz transformation, but we know based on the Michelson-Morley experiment that the speed of light is the same in all inertial frames, and we've already determined, e.g., by arguments like those earlier, that there can be no length contraction in the direction perpendicular to the motion. We construct a "light clock," consisting simply of two mirrors facing each other, with a light pulse bouncing back and forth between them.

   a. Suppose this light clock is moving at a constant velocity v in the direction perpendicular its own optical arm, which is of length L. Use the Pythagorean theorem to prove that the clock experiences a time dilation given by $\gamma = 1/\sqrt{1 - v^2}$, thereby fixing the time-time portion of the Lorentz transformation.

   b. Why is it significant for the interpretation of special relativity that the result from part a is independent of L?

   c. Carry out a similar calculation in the case where the clock moves with constant acceleration a as measured in some inertial frame. Although the result depends on L, prove that in the limit of small L, we recover the earlier constant-velocity result, with no explicit dependence on a.

> **Note**
>
> Some authors state a "clock postulate" for special relativity, which says that for a clock that is sufficiently small, the rate at which it runs depends only on v, not a (except in the trivial sense that v and a are related by calculus). The result of part c shows that the clock "postulate" is really a theorem, not a statement that is logically independent of the other postulates of special relativity. Although this argument only applies to a particular family of light clocks of various sizes, one can also make any small clock into an acceleration-insensitive clock, by attaching an accelerometer to it and applying an appropriate correction to compensate for the clock's observed sensitivity to acceleration. (It's still necessary for the clock to be small, since otherwise the lack of simultaneity in relativity makes it impossible to describe the whole clock as having a certain acceleration at a certain instant.) Farley at al.[22] have verified the "clock postulate" to within 2% for the radioactive decay of muons with γ ~ 12 being accelerated by magnetic fields at 5×10^{18} m/s². Some people get confused by this acceleration-independent property of small clocks and think that it contradicts the equivalence principle. For a good explanation, see http://math.ucr.edu/home/baez/physics/Relativity/SR/clock.html.
>
> ---
>
> [22] Nuovo Cimento 45 (1966) 281

2. Some of the most conceptually direct tests of relativistic time dilation were carried out by comparing the rates of twin atomic clocks, one left on a mountaintop for a certain amount of time, the other in a nearby valley below.[23] Unlike the clocks in the Hafele-Keating experiment, these are stationary for almost the entire duration of the experiment, so any time dilation is purely gravitational, not kinematic. One could object, however, that the clocks are not really at rest relative to one another, due to the earth's rotation. This is an example of how the distinction between gravitational and kinematic time dilations is frame-dependent, since the effect is purely gravitational in the rotating frame, where the gravitational field is reduced by the fictitious centrifugal force. Show that, in the non-rotating frame, the ratio of the kinematic effect to the gravitational one comes out to be 2.8×10^{-3} at the latitude of Tokyo. This small value indicates that the experiment can be interpreted as a very pure test of the gravitational time dilation effect. To calculate the effect, you will need to use the fact that, as discussed earlier, gravitational redshifts can be interpreted as gravitational time dilations.
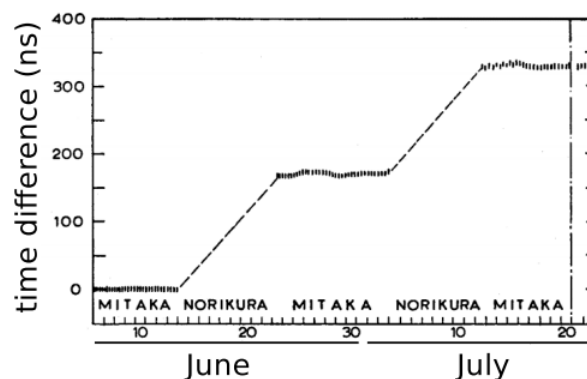
A graph from the paper by Iijima, showing the time difference between the two clocks. One clock was kept at Mitaka Observatory, at 58 m above sea level. The other was moved back and forth between a second observatory, Norikura Corona Station, and the peak of the Norikura volcano, 2876 m above sea level. The plateaus on the graph are data from the periods when the clocks were compared side by side at Mitaka. The difference between one plateau and the next is the gravitational time dilation accumulated during the period when the mobile clock was at the top of Norikura.

> **Note**
>
> L. Briatore and S. Leschiutta, "Evidence for the earth gravitational shift by direct atomic-time-scale comparison," Il Nuovo Cimento B, 37B (2): 219 (1977). Iijima et al., "An experiment for the potential blue shift at the Norikura Corona Station," Annals of the Tokyo Astronomical Observatory, Second Series, Vol. XVII, 2 (1978) 68.

3. (a) In Figure 2.5.10, we showed that the Thomas precession is proportional to area on the velocity disk. Use a similar argument to show that the Sagnac effect is proportional to the area enclosed by the loop. (b) Verify this more directly in the special case of a circular loop. (c) Show that a light clock of the type described in problem 1 is insensitive to rotation with constant angular velocity. (d) Connect these results to the commutativity and transitivity assumptions in the Einstein clock synchronization procedure described later.

4. Example 14 discusses relativistic bounds on the properties of matter, using the example of pulling a bucket out of a black hole. Derive a similar bound by considering the possibility of sending signals out of the black hole using longitudinal vibrations of a cable, as in the child's telephone made of two tin cans connected by a piece of string.

> **Note**
>
> Surprisingly subtle issues can arise in such calculations; see A.Y. Shiekh, Can. J. Phys. 70, 458 (1992). For a quantitative treatment of a dangling rope in relativity, see Greg Egan, "The Rindler Horizon," http://gregegan.customer.netspace.net.au/SCIENCE/Rindler/RindlerHorizon.html.

5. The Maxima program demonstrates how to multiply matrices and find Taylor series. Apply this technique to the following problem. For successive Lorentz boosts along the same axis with rapidities η_1 and η_2, find the matrix representing the combined Lorentz transformation, in a Taylor series up to the first nonclassical terms in each matrix element. A mixed Taylor series in two variables can be obtained simply by nesting taylor functions. The taylor function will happily work on matrices, not just scalars.

3: Differential Geometry

Differential geometry uses the techniques of differential calculus, integral calculus, linear algebra and multilinear algebra to study problems in geometry.

Thumbnail: A triangle immersed in a saddle-shape plane (a hyperbolic paraboloid), as well as two diverging ultraparallel lines. (Public Domain; LucasVB).

3.1: Introduction to Differential Geometry

General relativity is described mathematically in the language of *differential geometry*. Let's take those two terms in reverse order.

The *geometry* of spacetime is non-Euclidean, not just in the sense that the 3+1-dimensional geometry of Lorentz frames is different than that of 4 interchangeable Euclidean dimensions, but also in the sense that parallels do not behave in the way described by E5 or A1-A3. In a Lorentz frame, which describes space without any gravitational fields, particles whose world-lines are initially parallel will continue along their parallel world-lines forever. But in the presence of gravitational fields, initially parallel world-lines of free-falling particles will in general diverge, approach, or even cross. Thus, neither the existence nor the uniqueness of parallels can be assumed. We can't describe this lack of parallelism as arising from the curvature of the world-lines, because we're using the world-lines of free-falling particles as our definition of a "straight" line. Instead, we describe the effect as coming from the curvature of spacetime itself. The Lorentzian geometry is a description of the case in which this curvature is negligible.

What about the word *differential*? The equivalence principle states that even in the presence of gravitational fields, local Lorentz frames exist. How local is "local?" If we use a microscope to zoom in on smaller and smaller regions of spacetime, the Lorentzian approximation becomes better and better. Suppose we want to do experiments in a laboratory, and we want to ensure that when we compare some physically observable quantity against predictions made based on the Lorentz geometry, the resulting discrepancy will not be too large. If the acceptable error is ϵ, then we should be able to get the error down that low if we're willing to make the size of our laboratory no bigger than δ. This is clearly very similar to the Weierstrass style of defining limits and derivatives in calculus. In calculus, the idea expressed by differentiation is that every smooth curve can be approximated locally by a line; in general relativity, the equivalence principle tells us that curved spacetime can be approximated locally by flat spacetime. But consider that no practitioner of calculus habitually solves problems by filling sheets of scratch paper with epsilons and deltas. Instead, she uses the Leibniz notation, in which dy and dx are interpreted as infinitesimally small numbers. You may be inclined, based on your previous training, to dismiss infinitesimals as neither rigorous nor necessary. In 1966, Abraham Robinson demonstrated that concerns about rigor had been unfounded; we'll come back to this point in section 3.3. Although it is true that any calculation written using infinitesimals can also be carried out using limits, the following example shows how much more well suited the infinitesimal language is to differential geometry.

Example 1: Areas on a sphere

The area of a region S in the Cartesian plane can be calculated as $\int_S$ dA, where dA = dx dy is the area of an infinitesimal rectangle of width dx and height dy. A curved surface such as a sphere does not admit a global Cartesian coordinate system in which the constant coordinate curves are both uniformly spaced and perpendicular to one another. For example, lines of longitude on the earth's surface grow closer together as one moves away from the equator. Letting θ be the angle with respect to the pole, and ϕ the azimuthal angle, the approximately rectangular patch bounded by $\theta, \theta + d\theta, \phi$, and $\phi + d\phi$ has width r sin θ dθ and height r dϕ, giving dA = r^2 sin θ dθ dϕ. If you look at the corresponding derivation in an elementary calculus textbook that strictly eschews infinitesimals, the technique is to start from scratch with Riemann sums. This is extremely laborious, and moreover must be carried out again for every new case. In differential geometry, the curvature of the space varies from one point to the next, and clearly we don't want to reinvent the wheel with Riemann sums an infinite number of times, once at each point in space.

3.2: Tangent Vectors

It's not immediately clear what a vector means in the context of curved spacetime. The freshman physics notion of a vector carries all kinds of baggage, including ideas like rotation of vectors and a magnitude that is positive for nonzero vectors. We also used to assume the ability to represent vectors as arrows, i.e., geometrical figures of finite size that could be transported to other places — but in a curved geometry, it is not in general possible to transport a figure to another location without distorting its shape, so there is no notion of congruence. For this reason, it's better to visualize vectors as tangents to the underlying space, as in Figure 3.2.1. Intuitively, we want to think of these vectors as arrows that are infinitesimally small, so that they fit on the curved surface without having to be bent. In the pictures, we simply scale them up to make them visible without an infinitely powerful microscope, and this scaling only makes them *appear* to rise out of the space in which they live.
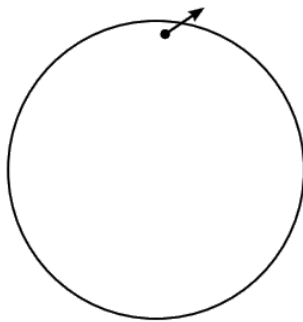
Figure 3.2.1: A vector can be thought of as lying in the plane tangent to a certain point.

A more formal definition of the notion of a tangent vector is given later.

3.3: Affine Notions and Parallel Transport

The Affine Parameter in Curved Spacetime: A Rough Sketch

We want to be able to measure things in curved spacetime. There turn out to be two complementary systems of measurement we can apply: *affine measure* and *metric measure*. Affine measure in a flat geometry was introduced in Section 2.1. Surprisingly, it turns out to be quite easy to generalize this to the curved case. Our construction of the affine parameter with a scaffolding of parallelograms depended on the existence and uniqueness of parallels expressed by axiom A1, so we might imagine that there was no point in trying to generalize the construction to curved spacetime. But the equivalence principle tells us that spacetime is locally affine to some approximation. Concretely, clock-time is one example of an affine parameter, and the curvature of spacetime clearly can't prevent us from building a clock and releasing it on a free-fall trajectory.

More generally, we can use the fact that every segment of a geodesic is geometrically similar to every other segment. For example, consider an arc of the earth's equator spanning one degree of longitude. That arc could be slid along the equator to a different location, then expanded to cover 3 degrees of longitude. The two arcs are similar.

> ### Example 2: Geodesics are special
>
> The following three non-examples show that this is a special property of geodesics.
>
> The property is not enjoyed by a non-geodesic curve. A segment of a pentagon that encompasses one of the vertices is not similar to some other segment that is straight.
>
> Another non-example involving non-geodesics is the curve that we get in 1+1-dimensional spacetime by joining together the positive x axis and the positive t axis. We can never take a one-year segment of the t axis and, through any combination of boosts and rotations, make it coincide with a one-light-year piece of the x axis. The original segment is timelike, and any boost or rotation will preserve its timelike character.
>
> Furthermore, it is not true in general, when curvature exists, that we can take any geometrical figure, transport it wherever we like, and also scale it as we like. For example, Euclidean geometry is a good approximation on small portions of the Earth's spherical surface, so a roadmap can be made in the shape of a rectangle with four right-angle corners. However, it is not possible to scale up such a rectangle; to map a large portion of the world, we have to introduce distortions of the type used in map projections.

Because geodesics have this special property, we can slide any portion of a geodesic to anywhere else on the geodesic and employ it as a standard of measure. This gives us a complete system of measurement along that geodesic, and it works regardless of whether the geodesic is timelike, lightlike, or spacelike. But as in flat geometry, affine measurement does not allow us to compare lengths along one geodesic to lengths along another.

The Affine Parameter in More Detail

When we originally defined affine measure in Section 2.1, for a flat space, we did it through the explicit construction of a scaffolding. An important example of the differential, i.e., local, nature of our geometry is the generalization of the scaffolding construction from to a context broader than affine geometry.

To generalize the recipe for the construction (Figure 3.2.1), the first obstacle is the ambiguity of the instruction to construct parallelogram $(0, 1, q_0, q_1)$, which requires us to draw $1q_1$ parallel to $0q_0$. Suppose we construe this as an instruction to make the two segments initially parallel, i.e., parallel as they depart the line at 0 and 1. By the time they get to q_0 and q_1, they may be converging or diverging.
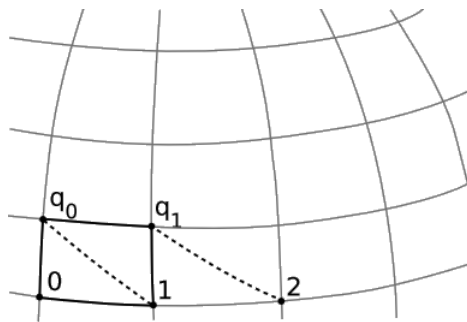
Figure 3.3.1 - Construction of an affine parameter in curved spacetime.

Because parallelism is only approximate here, there will be a certain amount of error in the construction of the affine parameter. One way of detecting such an error is that lattices constructed with different initial distances will get out of step with one another. For example, we can define $\frac{1}{2}$ as before by requiring that the lattice constructed with initial segment 0 $\frac{1}{2}$ line up with the original lattice at 1. We will find, however, that they do *not* quite line up at other points, such as 2. Let's use this discrepancy $\epsilon = 2 - 2'$ as a numerical measure of the error. It will depend on both δ_1, the distance 01, and on δ_2, the distance between 0 and q_0. Since ϵ vanishes for either $\delta_1 = 0$ or $\delta_2 = 0$, and since the equivalence principle guarantees smooth behavior on small scales, the leading term in the error will in general be proportional to the product $\delta_1 \delta_2$. In the language of infinitesimals, we can replace δ_1 and δ_2 with infinitesimally short distances, which for simplicity we assume to be equal, and which we call $d\lambda$. Then the affine parameter λ is defined as $\lambda = \int d\lambda$, where the error of order $d\lambda^2$ is, as usual, interpreted as the negligible discrepancy between the integral and its approximation as a Riemann sum.

Parallel Transport

If you were alert, you may have realized that I cheated you at a crucial point in this construction. We were to make $1q_1$ and $0q_0$ "initially parallel" as they left 01. How should we even define this idea of "initially parallel?" We could try to do it by making angles q_001 and q_112 equal, but this doesn't quite work, because it doesn't specify whether the angle is to the left or the right on the two-dimensional plane of the page. In three or more dimensions, the issue becomes even more serious. The construction workers building the lattice need to keep it all in one plane, but how do they do that in curved spacetime?

A mathematician's answer would be that our geometry lacks some additional structure called a *connection*, which is a rule that specifies how one locally flat neighborhood is to be joined seamlessly onto another locally flat neighborhood nearby. If you've ever bought two maps and tried to tape them together to make a big map, you've formed a connection. If the maps were on a large enough scale, you also probably noticed that this was impossible to do perfectly, because of the curvature of the earth.

Physically, the idea is that in flat spacetime, it is possible to construct inertial guidance systems like the ones discussed earlier. Since they are possible in flat spacetime, they are also possible in locally flat neighborhoods of spacetime, and they can then be carried from one neighborhood to another. In three space dimensions, a gyroscope's angular momentum vector maintains its direction, and we can orient other vectors, such as $1q_1$, relative to it. Suppose for concreteness that the construction of the affine parameter above is being carried out in three space dimensions. We place a gyroscope at 0, orient its axis along $0q_0$, slide it along the line to 1, and then construct $1q_1$ along that axis.

In 3+1 dimensions, a gyroscope only does part of the job. We now have to maintain the direction of a four-dimensional vector. Four-vectors will not be discussed in detail until section 4.2, but similar devices can be used to maintain their orientations in spacetime. These physical devices are ways of defining a mathematical notion known as *parallel transport*, which allows us to take a vector from one point to another in space. In general, specifying a notion of parallel transport is equivalent to specifying a connection.

Parallel transport is path-dependent, as shown in Figure 3.2.2.
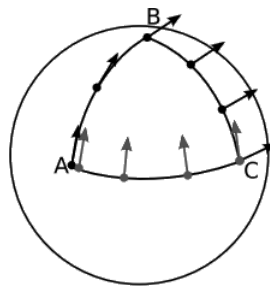
Figure 3.3.2 - Parallel transport is path-dependent. On the surface of this sphere, parallel-transporting a vector along ABC gives a different answer than transporting it along AC.

Affine Parameters Defined Only along Geodesics

In the context of flat spacetime, the affine parameter was defined only along lines, not arbitrary curves, and could not be compared between lines running in different directions. In curved spacetime, the same limitation is present, but with "along lines" replaced by "along geodesics." Figure 3.2.3 shows what goes wrong if we try to apply the construction to a world-line that isn't a geodesic. One definition of a geodesic is that it's the course we'll end up following if we navigate by keeping a fixed bearing relative to an inertial guidance device such as gyroscope; that is, the tangent to a geodesic, when parallel-transported farther along the geodesic, is still tangent. A non-geodesic curve lacks this property, and the effect on the construction of the affine parameter is that the segments nq_n drift more and more out of alignment with the curve.
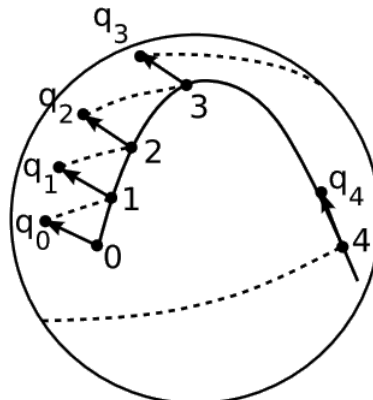
Figure 3.2.2. Parallel transport doesn't preserve the vectors' angle relative to the curve, as it would with a geodesic. The errors in the construction blow up in a way that wouldn't happen if the curve had been a geodesic. The fourth dashed parallel flies off wildly around the back of the sphere, wrapping around and meeting the curve at a point, 4, that is essentially random.

3.4: Models

A typical first reaction to the phrase "curved spacetime" — or even "curved space," for that matter — is that it sounds like nonsense. How can featureless, empty space itself be curved or distorted? The concept of a distortion would seem to imply taking all the points and shoving them around in various directions as in a Picasso painting, so that distances between points are altered. But if space has no identifiable dents or scratches, it would seem impossible to determine which old points had been sent to which new points, and the distortion would have no observable effect at all. Why should we expect to be able to build differential geometry on such a logically dubious foundation? Indeed, historically, various mathematicians have had strong doubts about the logical self-consistency of both non-Euclidean geometry and infinitesimals. And even if an authoritative source assures you that the resulting system is self-consistent, its mysterious and abstract nature would seem to make it difficult for you to develop any working picture of the theory that could play the role that mental sketches of graphs play in organizing your knowledge of calculus.

Figure 3.4.1 - Tullio Levi-Civita (1873- 1941) worked on models of number systems possessing infinitesimals and on differential geometry. He invented the tensor notation, which Einstein learned from his textbook. He was appointed to prestigious endowed chairs at Padua and the University of Rome, but was fired in 1938 because he was a Jew and an anti-fascist.

Models provide a way of dealing with both the logical issues and the conceptual ones. Figure 3.2.1 "pops" off of the page, presenting a strong psychological impression of a curved surface rendered in perspective. This suggests finding an actual mathematical object, such as a curved surface, that satisfies all the axioms of a certain logical system, such as non-Euclidean geometry. Note that the model may contain extrinsic elements, such as the existence of a third dimension, that are not connected to the system being modeled.

Let's focus first on consistency. In general, what can we say about the self-consistency of a mathematical system? To start with, we can never prove anything about the consistency or lack of consistency of something that is not a well-defined formal system, e.g., the Bible. Even Euclid's *Elements*, which was a model of formal rigor for thousands of years, is loose enough to allow considerable ambiguity. If you're inclined to scoff at the silly Renaissance mathematicians who kept trying to prove the parallel postulate E5 from postulates E1-E4, consider the following argument. Suppose that we replace E5 with E5', which states that parallels *don't* exist: given a line and a point not on the line, no line can ever be drawn through the point and parallel to the given line. In the new system of plane geometry E' consisting of E1-E4 plus E5', we can prove a variety of theorems, and one of them is that there is an upper limit on the area of any figure. This imposes a limit on the size of circles, and that appears to contradict E3, which says we can construct a circle with any radius.

We therefore conclude that E' lacks self-consistency. Oops! As your high school geometry text undoubtedly mentioned in passing, E' is a perfectly respectable system called elliptic geometry. So what's wrong with this supposed proof of its lack of self-consistency? The issue is the exact statement of E3. E3 does not say that we can construct a circle given any real number as its radius. Euclid could not have intended any such interpretation, since he had no notion of real numbers. To Euclid, geometry was primary, and numbers were geometrically constructed objects, being represented as lengths, angles, areas, and volumes. A literal translation of Euclid's statement of the axiom is "To describe a circle with any center and distance."[1] "Distance" means a line segment. There is therefore no contradiction in E', because E' has a limit on the lengths of line segments.

Now suppose that such ambiguities have been eliminated from the system's basic definitions and axioms. In general, we expect it to be easier to prove an inconsistent system's inconsistency than to demonstrate the consistency of a consistent one. In the former case, we can start cranking out theorems, and if we can find a way to prove both proposition P and its negation ¬P, then obviously something is wrong with the system. One might wonder whether such a contradiction could remain contained within one corner of the system, like nuclear waste. It can't. Aristotelian logic allows proof by contradiction: if we prove both P and ¬P based on certain assumptions, then our assumptions must have been wrong. If we can prove both P and ¬P *without* making any assumptions, then proof by contradiction allows us to establish the truth of *any* randomly chosen proposition. Thus a single contradiction is sufficient, in Aristotelian logic, to invalidate the entire system. This goes by the Latin rubric *ex falso quodlibet,* meaning "from a falsehood, whatever you please." Thus any contradiction proves the inconsistency of the entire system.

Proving consistency is harder. If you're mathematically sophisticated, you may be tempted to leap directly to **Gödel's theorem**, and state that nobody can ever prove the self-consistency of a mathematical system. This would be a misapplication of Gödel. Gödel's theorem only applies to mathematical systems that meet certain technical criteria, and some of the interesting systems we're dealing with don't meet those criteria; in particular, Gödel's theorem doesn't apply to Euclidean geometry, and Euclidean geometry was proved self-consistent by Tarski and his students around 1950. Furthermore, we usually don't require an absolute proof of self-consistency. Usually we're satisfied if we can prove that a certain system, such as elliptic geometry, is at least as self-consistent as another system, such as Euclidean geometry. This is called *equiconsistency*. The general technique for proving equiconsistency of two theories is to show that a model of one can be constructed within the other.

Suppose, for example, that we construct a geometry in which the space of points is the surface of a sphere, and lines are understood to be the geodesics, i.e., the great circles whose centers coincide at the sphere's center. This geometry, called spherical geometry, is useful in cartography and navigation. It is non-Euclidean, as we can demonstrate by exhibiting at least one proposition that is false in Euclidean geometry. For example, construct a triangle on the earth's surface with one corner at the north pole, and the other two at the equator, separated by 90 degrees of longitude. The sum of its interior angles is 270 degrees, contradicting Euclid, book I, proposition 32. Spherical geometry must therefore violate at least one of the axioms E1-E5, and indeed it violates both E1 (because no unique line is determined by two antipodal points such as the north and south poles) and E5 (because parallels don't exist at all).

A closely related construction gives a model of elliptic geometry, in which E1 holds, and only E5 is thrown overboard. To accomplish this, we model a point using a diameter of the sphere,[2] and a line as the set of all diameters lying in a certain plane. This has the effect of identifying antipodal points, so that there is now no violation of E1. Roughly speaking, this is like lopping off half of the sphere, but making the edges wrap around. Since this model of elliptic geometry is embedded within a Euclidean space, all the axioms of elliptic geometry can now be proved as theorems in Euclidean geometry. If a contradiction arose from them, it would imply a contradiction in the axioms of Euclidean geometry. We conclude that elliptic geometry is equiconsistent with Euclidean geometry. This was known long before Tarski's 1950 proof of Euclidean geometry's self-consistency, but since nobody was losing any sleep over hidden contradictions in Euclidean geometry, mathematicians stopped wasting their time looking for contradictions in elliptic geometry.

> **Note**
>
> The term "elliptic" may be somewhat misleading here. The model is still constructed from a sphere, not an ellipsoid.

> **Example 3: Infinitesimals**
>
> Consider the following axiomatically defined system of numbers:
>
> 1. It is a field, i.e., it has addition, subtraction, multiplication, and division with the usual properties.
> 2. It is an ordered geometry in the sense of O1-O4, and the ordering relates to addition and multiplication in the usual way.
> 3. Existence of infinitesimals: There exists a positive number d such that $d < 1$, $d < 1/2$, $d < 1/3$, . . .
>
> A model of this system can be constructed within the real number system by defining d as the identity function $d(x) = x$ and forming the set of functions of the form $f(d) = P(d)/Q(d)$, where P and Q are polynomials with real coefficients. The ordering of functions f and g is defined according to the sign of $\lim_{x \to 0^+} f(x) - g(x)$. Axioms 1-3 can all be proved from the real-number axioms. Therefore this system, which includes infinitesimals, is equiconsistent with the reals. More elaborate constructions can extend this to systems that have more of the properties of the reals, and a browser-based calculator that implements such a system is available at lightandmatter.com/calc/inf. Abraham Robinson extended this in 1966 to all of analysis, and thus there is nothing intrinsically nonrigorous about doing analysis in the style of Gauss and Euler, with symbols like dx representing infinitesimally small quantities.[3]

Besides proving consistency, these models give us insight into what's going on. The model of elliptic geometry suggests an insight into the reason that there is an upper limit on lengths and areas: it is because the space wraps around on itself. The model of infinitesimals suggests a fact that is not immediately obvious from the axioms: the infinitesimal quantities compose a hierarchy, so that for example 7d is in finite proportion to d, while d^2 is like a "lesser flea" in Swift's doggerel: "Big fleas have little fleas/ On their backs to ride 'em,/ and little fleas have lesser fleas,/And so, ad infinitum."

Spherical and elliptic geometry are not valid models of a general-relativistic spacetime, since they are locally Euclidean rather than Lorentzian, but they still provide us with enough conceptual guidance to come up with some ideas that might never have occurred to us otherwise:

- In spherical geometry, we can have a two-sided polygon called a lune that encloses a nonzero area. In general relativity, a lune formed by the world-lines of two particles represents motion in which the particles separate but are later reunited, presumably because of some mass between them that created a gravitational field. An example is gravitational lensing.
- Both spherical models wraps around on themselves, so that they are not topologically equivalent to infinite planes. We therefore form a conjecture there may be a link between curvature, which is a local property, and topology, which is global. Such a connection is indeed observed in relativity. For example, cosmological solutions of the equations of general relativity come in two flavors. One type has enough matter in it to produce more than a certain critical amount of curvature, and this type is topologically closed. It describes a universe that has finite spatial volume, and that will only exist for a finite time before it recontracts in a Big Crunch. The other type, corresponding to the universe we actually inhabit, has infinite spatial volume, will exist for infinite time, and is topologically open.
- There is a distance scale set by the size of the sphere, with its inverse being a measure of curvature. In general relativity, we expect there to be a similar way to measure curvature numerically, although the curvature may vary from point to point.
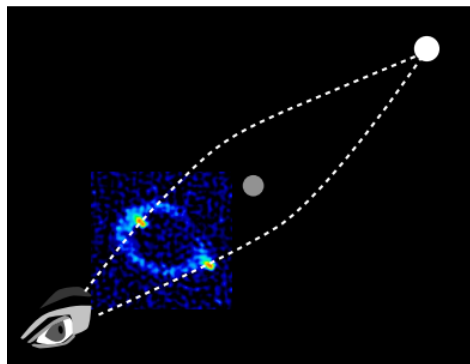
Figure 3.4.2 - An Einstein's ring is formed when there is a chance alignment of a distant source with a closer gravitating body. Here, a quasar, MG1131+0456, is seen as a ring due to focusing of light by an unknown object, possibly a supermassive black hole. Because the entire arrangement lacks perfect axial symmetry, the ring is nonuniform; most of its brightness is concentrated in two lumps on opposite sides. This type of gravitational lensing is direct evidence for the curvature of space predicted by gravitational lensing. The two geodesics form a lune, which is a figure that cannot exist in Euclidean geometry.

Exercise 3.4.1

Prove from the axioms E' that elliptic geometry, unlike spherical geometry, cannot have a lune with two distinct vertices. Convince yourself nevertheless, using the spherical model of E', that it is possible in elliptic geometry for two lines to enclose a region of space, in the sense that from any point P in the region, a ray emitted in any direction must intersect one of the two lines. Summarize these observations with a characterization of lunes in elliptic geometry versus lunes in spherical geometry.

References

[1] Heath, pp. 195-202

3.5: Intrinsic Quantities

Models can be dangerous, because they can tempt us to impute physical reality to features that are purely extrinsic, i.e., that are only present in that particular model. This is as opposed to intrinsic features, which are present in all models, and which are therefore logically implied by the axioms of the system itself. The existence of lunes is clearly an intrinsic feature of non-Euclidean geometries, because intersection of lines was defined before any model has even been proposed.

Example 4: Curvature in elliptic geometry

What about curvature? In the spherical model of elliptic geometry, the size of the sphere is an inverse measure of curvature. Is this a valid intrinsic quantity, or is it extrinsic? It seems suspect, because it is a feature of the model. If we try to define "size" as the radius R of the sphere, there is clearly reason for concern, because this seems to refer to the center of the sphere, but existence of a three-dimensional Euclidean space inside and outside the surface is clearly an extrinsic feature of the model. There is, however, a way in which a creature confined to the surface can determine R, by constructing geodesic and an affine parameter along that geodesic, and measuring the distance λ accumulated until the geodesic returns to the initial point. Since antipodal points are identified, λ equals half the circumference of the sphere, not its whole circumference, so R = $\frac{\lambda}{\pi}$, by wholly intrinsic methods.

Example 5: Extrinsic curvature

Euclid's axioms E1-E5 refer to explicit constructions. If a two-dimensional being can physically verify them all as descriptions of the two-dimensional space she inhabits, then she knows that her space is Euclidean, and that propositions such as the Pythagorean theorem are physically valid in her universe. But the diagram in a/1 illustrating illustrating the proof of the Pythagorean theorem in Euclid's Elements (proposition I.47) is equally valid if the page is rolled onto a cylinder, 2, or formed into a wavy corrugated shape, 3. These types of curvature, which can be achieved without tearing or crumpling the surface, are extrinsic rather than intrinsic. Of the curved surfaces in Figure 3.4.1, only the sphere, 4, has intrinsic curvature; the diagram can't be plastered onto the sphere without folding or cutting and pasting.
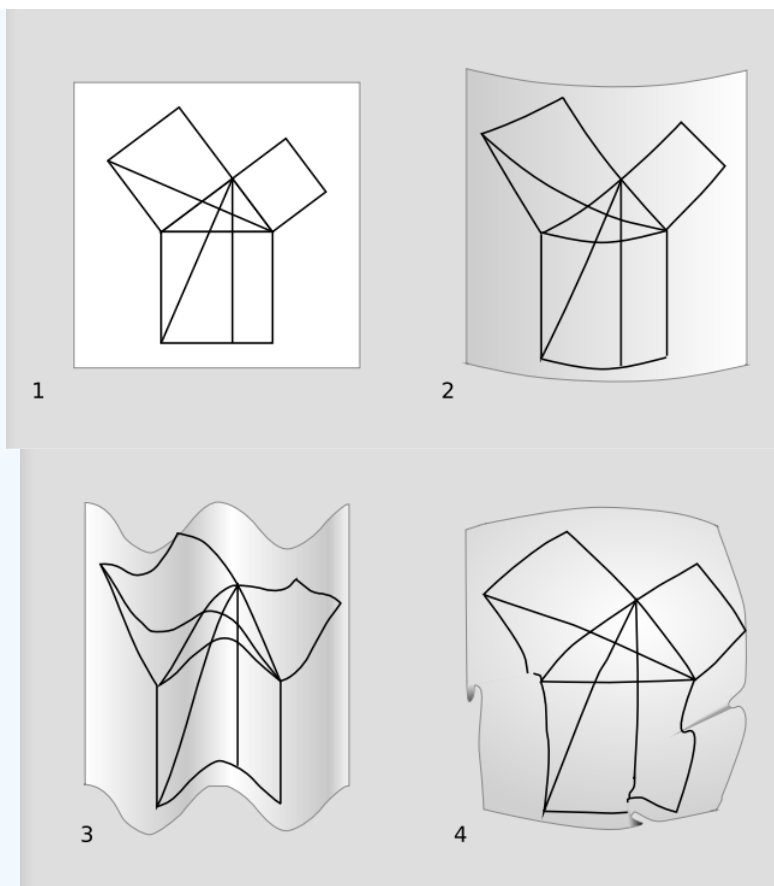
Figure 3.5.1

Exercise 3.5.1

Self-check: How would the ideas of example 5 apply to a cone?

Example 5 shows that it can be difficult to sniff out bogus extrinsic features that seem intrinsic, and example 4 suggests the desirability of developing methods of calculation that never refer to any extrinsic quantities, so that we never have to worry whether a symbol like R staring up at us from a piece of paper is intrinsic. This is why it is unlikely to be helpful to a student of general relativity to pick up a book on differential geometry that was written without general relativity specifically in mind. Such books have a tendency to casually mix together intrinsic and extrinsic notation. For example, a vector cross product $\mathbf{a} \times \mathbf{b}$ refers to a vector poking out of the plane occupied by $\mathbf{a}$ and $\mathbf{b}$, and the space outside the plane may be extrinsic; it is not obvious how to generalize this operation to the 3+1 dimensions of relativity (since the cross product is a three-dimensional beast), and even if it were, we could not be assured that it would have any intrinsically well defined meaning.

Coordinate Independence

To see how to proceed in creating a manifestly intrinsic notation, consider the two types of intrinsic observations that are available in general relativity:

- 1. We can tell whether events and world-lines are *incident*: whether or not two lines intersect, two events coincide, or an event lies on a certain line.

Incidence measurements, for example detection of gravitational lensing, are global, but they are the *only* global observations we can do.[4] If we were limited entirely to incidence, spacetime would be described by the austere system of projective geometry, a geometry without parallels or measurement. In projective geometry, all propositions are essentially statements about combinatorics, e.g., that it is impossible to plant seven trees so that they form seven lines of three trees each.

> **Note**
>
> Einstein referred to incidence measurements as "determinations of spacetime coincidences." For his presentation of this idea, see Appendix A.

But:

- 2. We can also do measurements in local Lorentz frames.

This gives us more power, but not as much as we might expect. Suppose we define a coordinate such as t or x. In Newtonian mechanics, these coordinates would form a predefined background, a preexisting stage for the actors. In relativity, on the other hand, consider a completely arbitrary change of coordinates of the form x → x' = f(x), where f is a smooth one-to-one function. For example, we could have x → x + px^3 + q sin(rx) (with p and q chosen small enough so that the mapping is always one-to-one). Since the mapping is one-to-one, the new coordinate system preserves all the incidence relations. Since the mapping is smooth, the new coordinate system is still compatible with the existence of local Lorentz frames. The difference between the two coordinate systems is therefore entirely extrinsic, and we conclude that a manifestly intrinsic notation should avoid any explicit reference to a coordinate system. That is, if we write a calculation in which a symbol such as x appears, we need to make sure that nowhere in the notation is there any hidden assumption that x comes from any particular coordinate system. For example, the equation should still be valid if the generic symbol x is later taken to represent the distance r from some center of symmetry. This coordinate-independence property is also known as general covariance, and this type of smooth change of coordinates is also called a diffeomorphism.

> **Example 6:**
>
> As an exotic example of a change of coordinates, take a torus and label it with coordinates (θ, φ), where $\theta + 2\pi$ is taken to be the same as θ, and similarly for φ. Now subject it to the coordinate transformation T defined by $\theta \rightarrow \theta + \varphi$, which is like opening the torus, twisting it by a full circle, and then joining the ends back together. T is known as the "Dehn twist," and it is different from most of the coordinate transformations we do in relativity because it can't be done smoothly, i.e., there is no continuous function f(x) on $0 \leq x \leq 1$ such that every value of f is a smooth coordinate transformation, f(0) is the identity transformation, and f(1) = T.

Frames Moving at c?

A good application of these ideas is to the question of what the world would look like in a frame of reference moving at the speed of light. This question has a long and honorable history. As a young student, Einstein tried to imagine what an electromagnetic wave would look like from the point of view of a motorcyclist riding alongside it. We now know, thanks to Einstein himself, that it really doesn't make sense to talk about such observers.

The most straightforward argument is based on the positivist idea that concepts only mean something if you can define how to measure them operationally. If we accept this philosophical stance (which is by no means compatible with every concept we ever discuss in physics), then we need to be able to physically realize this frame in terms of an observer and measuring devices. But we can't. It would take an infinite amount of energy to accelerate Einstein and his motorcycle to the speed of light.

Since arguments from positivism can often kill off perfectly interesting and reasonable concepts, we might ask whether there are other reasons not to allow such frames. There are. Recall that we placed two technical conditions on coordinate transformations: they are supposed to be smooth and one-to-one. The smoothness condition is related to the inability to boost Einstein's motorcycle into the speed-of-light frame by any continuous, classical process. (Relativity is a classical theory.) But independent of that, we have a problem with the one-to-one requirement. Figure 3.4.2 shows what happens if we do a series of Lorentz boosts to higher and higher velocities. It should be clear that if we could do a boost up to a velocity of c, we would have effected a coordinate transformation that was not one-to-one. Every point in the plane would be mapped onto a single lightlike line.
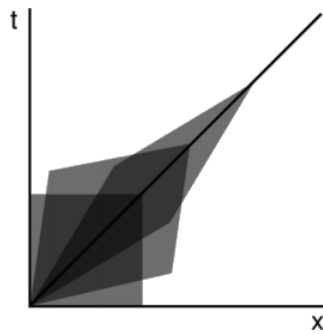
Figure 3.5.2 - A series of Lorentz boosts acts on a square.

This page titled 3.5: Intrinsic Quantities is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Benjamin Crowell via source content that was edited to the style and standards of the LibreTexts platform.

3.6: The Metric (Part 1)

Consider a coordinate x defined along a certain curve, which is not necessarily a geodesic. For concreteness, imagine this curve to exist in two spacelike dimensions, which we can visualize as the surface of a sphere embedded in Euclidean 3-space. These concrete features are not strictly necessary, but they drive home the point that we should not expect to be able to define x so that it varies at a steady rate with elapsed distance; for example, we know that it will not be possible to define a two-dimensional Cartesian grid on the surface of a sphere. In the figure, the tick marks are therefore not evenly spaced. This is perfectly all right, given the coordinate invariance of general relativity. Since the incremental changes in x *are* equal, I've represented them below the curve as little vectors of equal length. They are the wrong length to represent distances along the curve, but this wrongness is an inevitable fact of life in relativity.
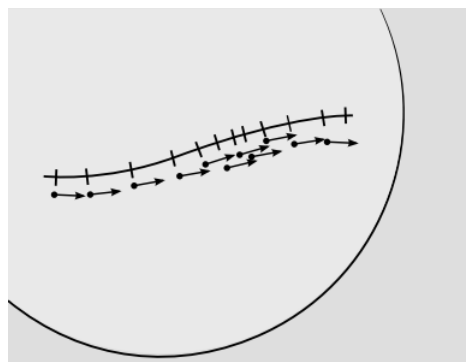
Figure $3.6.1$: The tick marks on the line define a coordinate measured along the line. It is not possible to set up such a coordinate system globally so that the coordinate is uniform everywhere. The arrows represent changes in the value of the coordinate; since the changes in the coordinate are all equal, the arrows are all the same length.

Now suppose we want to integrate the arc length of a segment of this curve. The little vectors are infinitesimal. In the integrated length, each little vector should contribute some amount, which is a scalar. This scalar is not simply the magnitude of the vector, $ds \neq \sqrt{d\mathbf{x} \cdot d\mathbf{x}}$, since the vectors are the wrong length. Figure $3.6.1$ is clearly reminiscent of the geometrical picture of vectors and dual vectors developed earlier. But the purely affine notion of vectors and their duals is not enough to define the length of a vector in general; it is only sufficient to define a length relative to other lengths along the same geodesic. When vectors lie along different geodesics, we need to be able to specify the additional conversion factor that allows us to compare one to the other. The piece of machinery that allows us to do this is called a *metric*.

Fixing a metric allows us to define the proper scaling of the tick marks relative to the arrows at a given point, i.e., in the birdtracks notation it gives us a natural way of taking a displacement vector such as $\rightarrow$s, with the arrow pointing into the symbol, and making a corresponding dual vector s$\rightarrow$, with the arrow coming out. This is a little like cloning a person but making the clone be of the opposite sex. Hooking them up like s$\rightarrow$s then tells us the squared magnitude of the vector. For example, if $\rightarrow$dx is an infinitesimal timelike displacement, then dx$\rightarrow$dx is the squared time interval dx^2 measured by a clock traveling along that displacement in spacetime. (Note that in the notation dx^2, it's clear that dx is a scalar, because unlike $\rightarrow$dx and dx$\rightarrow$ it doesn't have any arrow coming in or out of it.) Figure $3.6.2$ shows the resulting picture.
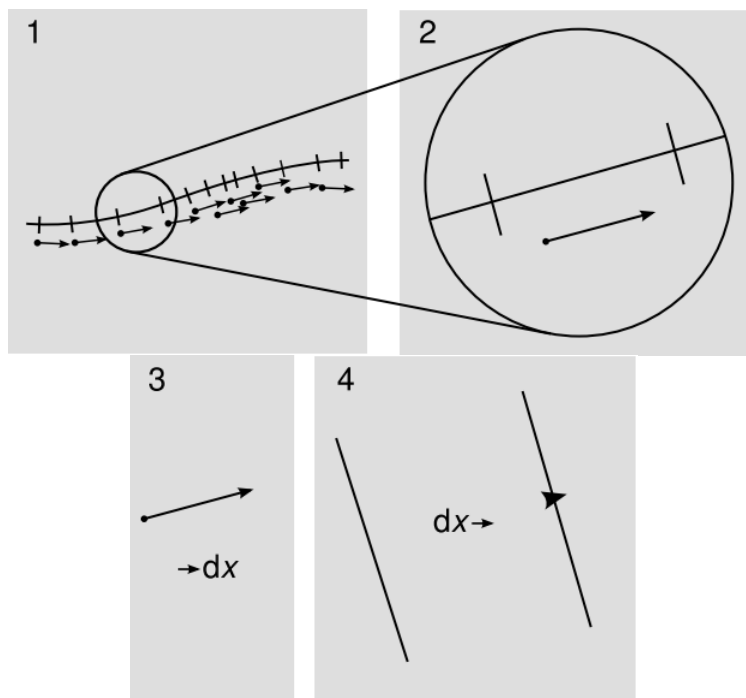
Figure 3.6.2: The vectors →dx and dx→ are duals of each other.

In the abstract index notation introduced earlier, the vectors →dx and dx→ are written dx^a and dx_a. When a specific coordinate system has been fixed, we write these with concrete, Greek indices, dx^μ and dx_μ. In an older and conceptually incompatible notation and terminology due to Sylvester (1853), one refers to dx_μ as a *contravariant* vector, and dx^μ as *covariant*. The confusing terminology is summarized in Appendix C.

The assumption that a metric exists is nontrivial. There is no metric in Galilean spacetime, for example, since in the limit c → ∞ the units used to measure timelike and spacelike displacements are not comparable. Assuming the existence of a metric is equivalent to assuming that the universe holds at least one physically manipulable clock or ruler that can be moved over long distances and accelerated as desired. In the distant future, large and causally isolated regions of the cosmos may contain only massless particles such as photons, which cannot be used to build clocks (or, equivalently, rulers); the physics of these regions will be fully describable without a metric. If, on the other hand, our world contains not just zero or one but two or more clocks, then the metric hypothesis requires that these clocks maintain a consistent relative rate when accelerated along the same world-line. This consistency is what allows us to think of relativity as a theory of space and time rather than a theory of clocks and rulers. There are other relativistic theories of gravity besides general relativity, and some of these violate this hypothesis.

Given a dx^μ, how do we find its dual dx_μ, and vice versa? In one dimension, we simply need to introduce a real number g as a correction factor. If one of the vectors is shorter than it should be in a certain region, the correction factor serves to compensate by making its dual proportionately longer. The two possible mappings (*covariant* to *contravariant* and *contravariant* to *covariant*) are accomplished with factors of g and 1/g. The number g is the metric, and it encodes all the information about distances. For example, if ϕ represents longitude measured at the arctic circle, then the metric is the only source for the datum that a displacement dϕ corresponds to 2540 km per radian.

Now let's generalize to more than one dimension. Because globally Cartesian coordinate systems can't be imposed on a curved space, the constant-coordinate lines will in general be neither evenly spaced nor perpendicular to one another. If we construct a local set of basis vectors lying along the intersections of the constant coordinate surfaces, they will not form an orthonormal set. We would like to have an expression of the form

$$ds^2 = \sum dx^\mu dx_\mu \tag{3.6.1}$$

for the squared arc length, and using the Einstein summation notation this becomes

$$ds^2 = dx^\mu dx_\mu \tag{3.6.2}$$

The Euclidean Metric

For Cartesian coordinates in a Euclidean plane, where one doesn't normally bother with the distinction between covariant and contravariant vectors, this expression for ds^2 is simply the *Pythagorean theorem*, summed over two values of μ for the two coordinates:

$$ds^2 = dx^\mu dx_\mu = dx^2 + dy^2 \tag{3.6.3}$$

The symbols dx, ds^0, dx^0, and dx_0 are all synonyms, and likewise for dy, ds^1, dx^1, and dx_1. (Because notations such as ds^1 force the reader to keep track of which digits have been assigned to which letters, it is better practice to use notation such as dy or ds^y; the latter notation could in principle be confused with one in which y was a variable taking on values such as 0 or 1, but in reality we understand it from context, just as we understand that the d's in $\frac{dy}{dx}$ are not referring to some variable d that stands for a number.)

In the non-Euclidean case, the Pythagorean theorem is false; dx^μ and dx_μ are no longer synonyms, so their product is no longer simply the square of a distance. To see this more explicitly, let's write the expression so that only the covariant quantities occur. By local flatness, the relationship between the covariant and contravariant vectors is linear, and the most general relationship of this kind is given by making the metric a symmetric matrix $g_{\mu\nu}$. Substituting $dx_\mu = g_{\mu\nu}x^\nu$, we have

$$ds^2 = g_{\mu\nu}dx^\mu dx^\nu \tag{3.6.4}$$

where there are now implied sums over both μ and ν. Notice how implied sums occur only when the repeated index occurs once as a superscript and once as a subscript; other combinations are ungrammatical.

> ### Exercise 3.6.1
>
> Self-check: Why does it make sense to demand that the metric be symmetric?

Earlier we encountered the distinction among scalars, vectors, and dual vectors. These are specific examples of *tensors*, which can be expressed in the birdtracks notation as objects with m arrows coming in and n coming out, or. In index notation, we have m superscripts and n subscripts. A scalar has m = n = 0. A dual vector has (m, n) = (0, 1), a vector (1, 0), and the metric (0, 2). We refer to the number of indices as the rank of the tensor. Tensors are discussed in more detail, and defined more rigorously, in chapter 4. For our present purposes, it is important to note that just because we write a symbol with subscripts or superscripts, that doesn't mean it deserves to be called a tensor. This point can be understood in the more elementary context of Newtonian scalars and vectors. For example, we can define a Euclidean "vector" **u** = (m, T, e), where m is the mass of the moon, T is the temperature in Chicago, and e is the charge of the electron. This creature **u** doesn't deserve to be called a vector, because it doesn't behave as a vector under rotation. The general philosophy is that a tensor is something that has certain properties under changes of coordinates. For example, we've already seen earlier the different scaling behavior of tensors with ranks (1, 0), (0, 0), and (0, 1).

When discussing the symmetry of rank-2 tensors, it is convenient to introduce the following notation:

$$T_{(ab)} = \frac{1}{2}\left(T_{ab} + T_{ba}\right) \tag{3.6.5}$$

$$T_{[ab]} = \frac{1}{2}\left(T_{ab} - T_{ba}\right) \tag{3.6.6}$$

Any T_{ab} can be split into symmetric and antisymmetric parts. This is similar to writing an arbitrary function as a sum of and odd function and an even function. The metric has only a symmetric part: $g_{(ab)} = g_{ab}$, and $g_{[ab]} = 0$. This notation is generalized to ranks greater than 2 later.

> ### Exercise 3.6.2
>
> Self-check: Characterize an antisymmetric rank-2 tensor in two dimensions.

> ### Example 7: A change of scale
>
> - Suppose we start by describing the Euclidean plane with a certain set of Cartesian coordinates, but then want to change to a new set of coordinates that are rescaled compared to the original ones. How is the effect of this rescaling represented in g?.
> - If we change our units of measurement so that $x_\mu \to \alpha x^\mu$, while demanding that ds^2 come out the same, then we need $g_{\mu\nu} \to \alpha^{-2} g_{\mu\nu}$.

Comparing with section 2.1, we deduce the general rule that a tensor of rank (m, n) transforms under scaling by picking up a factor of α^{m-n}.

This whole notion of scaling and units in general relativity turns out to be nontrivial and interesting. See section 5.11, for a more detailed discussion.

Example 8: polar coordinates

Consider polar coordinates (r, θ) in a Euclidean plane. The constant-coordinate curves happen to be orthogonal everywhere, so the off-diagonal elements of the metric $g_{r\theta}$ and $g_{\theta r}$ vanish. Infinitesimal coordinate changes dr and dθ correspond to infinitesimal displacements dr and r dθ in orthogonal directions, so by the Pythagorean theorem, $ds^2 = dr^2 + r^2\ d\theta^2$, and we read off the elements of the metric $g_{rr} = 1$ and $g_{\theta\theta} = r^2$.

Notice how in example 8 we started from the generally valid relation $ds^2 = g_{\mu\nu}dx^\mu dx^\nu$, but soon began writing down facts like $g_{\theta\theta} = r^2$ that were only valid in this particular coordinate system. To make it clear when this is happening, we maintain the distinction between abstract Latin indices and concrete Greek indices introduced earlier. For example, we can write the general expression for squared differential arc length with Latin indices,

$$ds^2 = g_{ij}dx^i dx^j, \tag{3.6.7}$$

because it holds regardless of the coordinate system, whereas the vanishing of the off-diagonal elements of the metric in Euclidean polar coordinates has to be written as $g_{\mu\nu} = 0$ for $\mu \neq \nu$, since it would in general be false if we used a different coordinate system to describe the same Euclidean plane.

Example 9: Oblique Cartesian coordinates
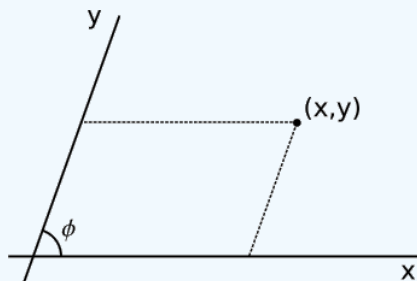
Figure 3.6.3

- Oblique Cartesian coordinates are like normal Cartesian coordinates in the plane, but their axes are at at an angle $\phi \neq \frac{\pi}{2}$ to one another. Find the metric in these coordinates. The space is globally Euclidean.
- Since the coordinates differ from Cartesian coordinates only in the angle between the axes, not in their scales, a displacement dx^i along either axis, i = 1 or 2, must give ds = dx, so for the diagonal elements we have $g_{11} = g_{22} = 1$. The metric is always symmetric, so $g_{12} = g_{21}$. To fix these off-diagonal elements, consider a displacement by ds in the direction perpendicular to axis 1. This changes the coordinates by $dx^1 = -ds \cot\phi$ and $dx^2 = ds \cos\phi$. We then have

$$\begin{aligned} ds^2 &= g_{ij}dx^i dx^j \\ &= ds^2(\cot^2\phi + csc^2\phi - 2g_{12}\cot\phi\csc\phi) \\ g_{12} &= \cos\phi. \end{aligned}$$

Example 10: area

In one dimension, g is a single number, and lengths are given by ds = $\sqrt{g}$ dx. The square root can also be understood through example 7, in which we saw that a uniform rescaling x $\to \alpha$x is reflected in $g_{\mu\nu} \to \alpha^{-2}g_{\mu\nu}$.

In two-dimensional Cartesian coordinates, multiplication of the width and height of a rectangle gives the element of area $dA = \sqrt{g_{11}g_{22}}dx^1 dx^2$. Because the coordinates are orthogonal, g is diagonal, and the factor of $\sqrt{g_{11}g_{22}}$ is identified as the square root of its determinant, so dA = $\sqrt{|g|}$ dx^1 dx^2. Note that the scales on the two axes are not necessarily the same, $g_{11} \neq g_{22}$.

The same expression for the element of area holds even if the coordinates are not orthogonal. In example 9, for instance, we have $\sqrt{|g|} = \sqrt{1 - \cos^2 \phi} = \sin \phi$, which is the right correction factor corresponding to the fact that dx^1 and dx^2 form a parallelepiped rather than a rectangle.

Example 11: area of a sphere

For coordinates (θ, ϕ) on the surface of a sphere of radius r, we have, by an argument similar to that of example 8 $g_{\theta\theta} = r^2, g_{\phi\phi} = r^2 \sin^2 \theta, g_{\theta\phi} = 0$. The area of the sphere is

$$
\begin{aligned}
A &= \int dA \\
&= \int \int \sqrt{|g|} \, d\theta \, d\phi \\
&= r^2 \int \int \sin \theta \, d\theta \, d\phi \\
&= 4\pi r^2
\end{aligned}
$$

Example 12: inverse of the metric

- Relate g^{ij} to g_{ij}.
- The notation is intended to treat covariant and contravariant vectors completely symmetrically. The metric with lower indices g_{ij} can be interpreted as a change-of-basis transformation from a contravariant basis to a covariant one, and if the symmetry of the notation is to be maintained, g^{ij} must be the corresponding inverse matrix, which changes from the covariant basis to the contravariant one. The metric must always be invertible.

In the one-dimensional case, the metric at any given point was simply some number g, and we used factors of g and 1/g to convert back and forth between covariant and contravariant vectors. Example 12 makes it clear how to generalize this to more dimensions:

$$
\begin{aligned}
x_a &= g_{ab} x^b \\
x^a &= g^{ab} x_b
\end{aligned}
$$

This is referred to as raising and lowering indices. There is no need to memorize the positions of the indices in these rules; they are the only ones possible based on the grammatical rules, which are that summation only occurs over top-bottom pairs, and upper and lower indices have to match on both sides of the equals sign. This whole system, introduced by Einstein, is called "index-gymnastics" notation.

Example 13: Raising and lowering indices on a rank-two tensor

In physics we encounter various examples of matrices, such as the moment of inertia tensor from classical mechanics. These have two indices, not just one like a vector. Again, the rules for raising and lowering indices follow directly from grammar. For example,

$$A^a_b = g^{ac} A_{cb} \tag{3.6.8}$$

and

$$A_{ab} = g_{ac} g_{bd} A^{cd}. \tag{3.6.9}$$

Example 14: A matrix operating on a vector

The row and column vectors from linear algebra are the covariant and contravariant vectors in our present terminology. (The convention is that covariant vectors are row vectors and contravariant ones column vectors, but I don't find this worth memorizing.) What about matrices? A matrix acting on a column vector gives another column vector, $\mathbf{q} = U\mathbf{p}$ Translating this into indexgymnastics notation, we have

$$q^a = U^{\cdots} \ldots p^b, \tag{3.6.10}$$

where we want to figure out the correct placement of the indices on U. Grammatically, the only possible placement is

$$q^a = U^a_b p^b. \tag{3.6.11}$$

This shows that the natural way to represent a column-vector-to-column-vector linear operator is as a rank-2 tensor with one upper index and one lower index.

In birdtracks notation, a rank-2 tensor is something that has two arrows connected to it. Our example becomes → q =→ U → p. That the result is itself an upper-index vector is shown by the fact that the right-hand-side taken as a whole has a single external arrow coming into it.

The distinction between vectors and their duals may seem irrelevant if we can always raise and lower indices at will. We can't always do that, however, because in many perfectly ordinary situations there is no metric. See example 6.

The Lorentz Music

In a locally Euclidean space, the Pythagorean theorem allows us to express the metric in local Cartesian coordinates in the simple form

$$g_\mu\mu = +1, g\mu\nu = 0 \tag{3.6.12}$$

i.e.,

$$g = diag(+1, +1, \ldots, +1). \tag{3.6.13}$$

This is not the appropriate metric for a locally Lorentz space. The axioms of Euclidean geometry E3 (existence of circles) and E4 (equality of right angles) describe the theory's invariance under rotations, and the Pythagorean theorem is consistent with this, because it gives the same answer for the length of a vector even if its components are reexpressed in a new basis that is rotated with respect to the original one. In a Lorentzian geometry, however, we care about invariance under Lorentz boosts, which do not preserve the quantity $t^2 + x^2$. It is not circles in the (t, x) plane that are invariant, but light cones, and this is described by giving g_{tt} and g_{xx} opposite signs and equal absolute values. A lightlike vector (t, x), with t = x, therefore has a magnitude of exactly zero

$$s^2 = g_{tt}t^2 + g_{xx}x^2 = 0, \tag{3.6.14}$$

and this remains true after the Lorentz boost (t, x) → (γt, γx). It is a matter of convention which element of the metric to make positive and which to make negative. In this book, I'll use g_{tt} = +1 and g_{xx} = −1, so that g = diag(+1, −1). This has the advantage that any line segment representing the timelike world-line of a physical object has a positive squared magnitude; the forward flow of time is represented as a positive number, in keeping with the philosophy that relativity is basically a theory of how causal relationships work. With this sign convention, spacelike vectors have positive squared magnitudes, timelike ones negative. The same convention is followed, for example, by Penrose. The opposite version, with g = diag(−1, +1) is used by authors such as Wald and Misner, Thorne, and Wheeler.

Our universe does not have just one spatial dimension, it has three, so the full metric in a Lorentz frame is given by

$$g = diag(+1, −1, −1, −1). \tag{3.6.15}$$

Example 15: Mixed covariant-contravariant form of the metric

In example 13, we saw how to raise and lower indices on a rank-two tensor, and example 14 showed that it is sometimes natural to consider the form in which one index is raised and one lowered. The metric itself is a rank-two tensor, so let's see what happens when we compute the mixed form g^a_b from the lower-index form. In general, we have

$$A^a_b = g^{ac} A_{cb}, \tag{3.6.16}$$

and substituting g for A gives

$$g^a_b = g^{ac} g_{cb}. \tag{3.6.17}$$

But we already know that g⋯ is simply the inverse matrix of g.. (example 12), which means that g^a_b is simply the identity matrix. That is, whereas a quantity like g_{ab} or g^{ab} carries all the information about our system of measurement at a given point, g^a_b carries no information at all. Where g_{ab} or g^{ab} can have both positive and negative elements, elements that have units, and off-diagonal elements, g^a_b is just a generic symbol carrying no information other than the dimensionality of the space.

> The metric tensor is so commonly used that it is simply left out of birdtrack diagrams. Consistency is maintained because because $g^a{}_b$ is the identity matrix, so $\rightarrow g \rightarrow$ is the same as $\rightarrow \rightarrow$.

3.7: The Metric (Part 2)

Isometry, Inner Products, and the Erlangen Program

In Euclidean geometry, the dot product of vectors **a** and **b** is given by

$$g_{xx}a_xb_x + g_{yy}a_yb_y + g_{zz}a_zb_z = a_xb_x + a_yb_y + a_zb_z, \tag{3.7.1}$$

and in the special case where **a** = **b** we have the squared magnitude. In the tensor notation,

$$a^\mu b_\nu = a^1b_1 + a^2b_2 + a^3b_3. \tag{3.7.2}$$

Like magnitudes, dot products are invariant under rotations. This is because knowing the dot product of vectors **a** and **b** entails knowing the value of

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{a}|\cos\theta_{\mathbf{ab}} \tag{3.7.3}$$

and Euclid's E4 (equality of right angles) implies that the angle $\theta_{\mathbf{ab}}$ is invariant. The same axioms also entail invariance of dot products under translation; Euclid waits only until the second proposition of the *Elements* to prove that line segments can be copied from one location to another. This seeming triviality is actually false as a description of physical space, because it amounts to a statement that space has the same properties everywhere.

The set of all transformations that can be built out of successive translations, rotations, and reflections is called the group of isometries. It can also be defined as the group that preserves dot products, or the group that preserves congruence of triangles.

> ### GRoups
>
> In mathematics, a group is defined as a binary operation that has an identity, inverses, and associativity. For example, addition of integers is a group. In the present context, the members of the group are not numbers but the transformations applied to the Euclidean plane. The group operation on transformations T_1 and T_2 consists of finding the transformation that results from doing one and then the other, i.e., composition of functions.

In Lorentzian geometry, we usually avoid the Euclidean term dot product and refer to the corresponding operation by the more general term *inner product*. In a specific coordinate system we have

$$a^\mu b_\nu = a^0b_0 - a^1b_1 - a^2b_2 - a^3b_3. \tag{3.7.4}$$

The inner product is invariant under Lorentz boosts, and also under the Euclidean isometries. The group found by making all possible combinations of continuous transformations from these two sets is called the Poincaré group. The Poincaré group is not the symmetry group of all of spacetime, since curved spacetime has different properties in different locations. The equivalence principle tells us, however, that space can be approximated locally as being flat, so the Poincaré group is locally valid, just as the Euclidean isometries are locally valid as a description of geometry on the Earth's curved surface.

> ### CPT Symmetry
>
> The discontinuous transformations of spatial reflection and time reversal are not included in the definition of the Poincaré group, although they do preserve inner products. General relativity has symmetry under spatial reflection (called P for parity), time reversal (T), and charge inversion (C), but the standard model of particle physics is only invariant under the composition of all three, CPT, not under any of these symmetries individually.

> ### Example 16: The triangle inequality
>
> In Euclidean geometry, the triangle inequality |**b** + **c**| < |**b**| + |**c**| follows from
>
> $$(|\mathbf{b}| + |\mathbf{c}|)^2 - (\mathbf{b}+\mathbf{c}) \cdot (\mathbf{b}+\mathbf{c}) = 2(|\mathbf{b}||\mathbf{c}| - \mathbf{b}\cdot\mathbf{c}) \geq 0. \tag{3.7.5}$$
>
> The reason this quantity always comes out positive is that for two vectors of fixed magnitude, the greatest dot product is always achieved in the case where they lie along the same direction.
>
> In Lorentzian geometry, the situation is different. Let **b** and **c** be timelike vectors, so that they represent possible world-lines. Then the relation **a** = **b** + **c** suggests the existence of two observers who take two different paths from one event to another. A

goes by a direct route while B takes a detour. The magnitude of each timelike vector represents the time elapsed on a clock carried by the observer moving along that vector. The triangle equality is now reversed, becoming $|\mathbf{b} + \mathbf{c}| > |\mathbf{b}| + |\mathbf{c}|$. The difference from the Euclidean case arises because inner products are no longer necessarily maximized if vectors are in the same direction. E.g., for two lightlike vectors, $b^i c_j$ vanishes entirely if $\mathbf{b}$ and $\mathbf{c}$ are parallel. For timelike vectors, parallelism actually minimizes the inner product rather than maximizing it.

Proof

Let $\mathbf{b}$ and $\mathbf{c}$ be parallel and timelike, and directed forward in time. Adopt a frame of reference in which every spatial component of each vector vanishes. This entails no loss of generality, since inner products are invariant under such a transformation. Since the time-ordering is also preserved under transformations in the Poincaré group, each is still directed forward in time, not backward. Now let $\mathbf{b}$ and $\mathbf{c}$ be pulled away from parallelism, like opening a pair of scissors in the $x - t$ plane. This reduces $b_t c_t$, while causing $b_x c_x$ to become negative. Both effects increase the inner product.

□

In his 1872 inaugural address at the University of Erlangen, Felix Klein used the idea of groups of transformations to lay out a general classification scheme, known as the Erlangen program, for all the different types of geometry. Each geometry is described by the group of transformations, called the principal group, that preserves the truth of geometrical statements. Euclidean geometry's principal group consists of the isometries combined with arbitrary changes of scale, since there is nothing in Euclid's axioms that singles out a particular distance as a unit of measurement. In other words, the principal group consists of the transformations that preserve similarity, not just those that preserve congruence. Affine geometry's principal group is the transformations that preserve parallelism; it includes shear transformations, and there is therefore no invariant notion of angular measure or congruence. Unlike Euclidean and affine geometry, elliptic geometry does not have scale invariance. This is because there is a particular unit of distance that has special status; as we saw in example 4, a being living in an elliptic plane can determine, by entirely intrinsic methods, a distance scale R, which we can interpret in the hemispherical model as the radius of the sphere. General relativity breaks this symmetry even more severely. Not only is there a scale associated with curvature, but the scale is different from one point in space to another.

Einstein's Carousel

Non-Euclidean Geometry Observed in the Rotating Frame

The following example was historically important, because Einstein used it to convince himself that general relativity should be described by non-Euclidean geometry.[8] Its interpretation is also fairly subtle, and the early relativists had some trouble with it.

Note

The example is described in Einstein's paper "The Foundation of the General Theory of Relativity." An excerpt, which includes the example, is given in Appendix A.

Suppose that observer A is on a spinning carousel while observer B stands on the ground. B says that A is accelerating, but by the equivalence principle A can say that she is at rest in a gravitational field, while B is free-falling out from under her. B measures the radius and circumference of the carousel, and finds that their ratio is 2π. A carries out similar measurements, but when she puts her meter-stick in the azimuthal direction it becomes Lorentz-contracted by the factor $\gamma = (1 - \omega^2 r^2)^{-1/2}$, so she finds that the ratio is greater than 2π. In A's coordinates, the spatial geometry is non-Euclidean, and the metric differs from the Euclidean one found in example 8.
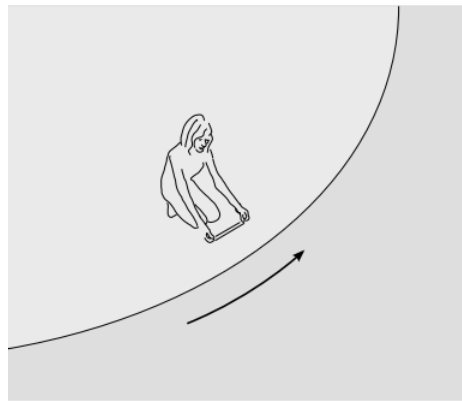
Figure 3.7.4: Observer A, rotating with the carousel, measures an azimuthal distance with a ruler.

Observer A feels a force that B considers to be fictitious, but that, by the equivalence principle, A can say is a perfectly real gravitational force. According to A, an observer like B is free-falling away from the center of the disk under the influence of this gravitational field. A also observes that the spatial geometry of the carousel is non-Euclidean. Therefore it seems reasonable to conjecture that gravity can be described by non-Euclidean geometry, rather than as a physical force in the Newtonian sense.

At this point, you know as much about this example as Einstein did in 1912, when he began using it as the seed from which general relativity sprouted, collaborating with his old schoolmate, mathematician Marcel Grossmann, who knew about differential geometry. The remainder of this subsection, which you may want to skip on a first reading, goes into more detail on the interpretation and mathematical description of the rotating frame of reference. Even more detailed treatments are given by Grøn[9] and Dieks.[10]

Ehrenfest's Paradox

Ehrenfest[11] described the following paradox. Suppose that observer B, in the lab frame, measures the radius of the disk to be r when the disk is at rest, and r' when the disk is spinning. B can also measure the corresponding circumferences C and C'. Because B is in an inertial frame, the spatial geometry does not appear non-Euclidean according to measurements carried out with his meter sticks, and therefore the Euclidean relations $C = 2\pi r$ and $C' = 2\pi r'$ both hold. The radial lines are perpendicular to their own motion, and they therefore have no length contraction, $r = r'$, implying $C = C'$. The outer edge of the disk, however, is everywhere tangent to its own direction of motion, so it is Lorentz contracted, and therefore $C' < C$. The resolution of the paradox is that it rests on the incorrect assumption that a rigid disk can be made to rotate. If a perfectly rigid disk was initially not rotating, one would have to distort it in order to set it into rotation, because once it was rotating its outer edge would no longer have a length equal to 2π times its radius. Therefore if the disk is perfectly rigid, it can never be rotated. As discussed earlier, relativity does not allow the existence of infinitely rigid or infinitely strong materials. If it did, then one could violate causality. If a perfectly rigid disk existed, vibrations in the disk would propagate at infinite velocity, so tapping the disk with a hammer in one place would result in the transmission of information at $v > c$ to other parts of the disk, and then there would exist frames of reference in which the information was received before it was transmitted. The same applies if the hammer tap is used to impart rotational motion to the disk.

Figure 3.7.5: Einstein and Ehrenfest.

The Metric in the Rotating Frame

What if we try to get around these problems by applying torque uniformly all over the disk, so that the rotation starts smoothly and simultaneously everywhere? We then run into issues identical to the ones raised by Bell's spaceship paradox. In fact, Ehrenfest's paradox is nothing more than Bell's paradox wrapped around into a circle. The same question of time synchronization comes up.

To spell this out mathematically, let's find the metric according to observer A by applying the change of coordinates $\theta' = \theta - \omega t$. First we take the Euclidean metric of example 8 and rewrite it as a (globally) Lorentzian metric in spacetime for observer B,

$$ds^2 = dt^2 - dr^2 - r^2 d\theta^2. \tag{3.7.6}$$

Applying the transformation into A's coordinates, we find

$$ds^2 = (1 - \omega^2 r^2)dt^2 - dr^2 - r^2 d\theta'^2 - 2\omega r^2 d\theta' dt. \tag{3.7.7}$$

Recognizing ωr as the velocity of one frame relative to another, and $(1 - \omega^2 r^2)^{-1/2}$ as γ, we see that we do have a relativistic time dilation effect in the dt^2 term. But the dr^2 and $d\theta'^2$ terms look Euclidean. Why don't we see any Lorentz contraction of the length scale in the azimuthal direction?

The answer is that coordinates in general relativity are arbitrary, and just because we can write down a certain set of coordinates, that doesn't mean they have any special physical interpretation. The coordinates (t, r, θ') do not correspond physically to the quantities that A would measure with clocks and meter-sticks. The tip-off is the $d\theta'$ dt cross-term. Suppose that A sends two cars driving around the circumference of the carousel, one clockwise and one counterclockwise, from the same point. If (t, r, θ') coordinates corresponded to clock and meter-stick measurements, then we would expect that when the cars met up again on the far side of the disk, their dashboards would show equal values of the arc length rθ' on their odometers and equal proper times ds on their clocks. But this is not the case, because the sign of the dθ' dt term is opposite for the two world-lines. The same effect occurs if we send beams of light in both directions around the disk, and this is the Sagnac effect.

This is a symptom of the fact that the coordinate t is not properly synchronized between different places on the disk. We already know that we should not expect to be able to find a universal time coordinate that will match up with every clock, regardless of the clock's state of motion. Suppose we set ourselves a more modest goal. Can we find a universal time coordinate that will match up with every clock, provided that the clock is at rest relative to the rotating disk?

The Spatial Metric and Synchronization of Clocks

A trick for improving the situation is to eliminate the $d\theta'$ dt cross-term by completing the square in the metric (Equation 3.7.7). The result is

$$ds^2 = (1 - \omega^2 r^2)\left[dt + \frac{\omega r^2}{1 - \omega^2 r^2}d\theta'\right]^2 - dr^2 - \frac{r^2}{1 - \omega^2 r^2}d\theta'^2. \tag{3.7.8}$$

The interpretation of the quantity in square brackets is as follows. Suppose that two observers situate themselves on the edge of the disk, separated by an infinitesimal angle $d\theta'$. They then synchronize their clocks by exchanging light pulses. The time of flight, measured in the lab frame, for each light pulse is the solution of the equation ds^2 = 0, and the only difference between the clockwise result dt$_1$ and the counterclockwise one dt$_2$ arises from the sign of dθ'. The quantity in square brackets is the same in both cases, so the amount by which the clocks must be adjusted is

$$dt = \frac{(dt_2 - dt_1)}{2}, \tag{3.7.9}$$

or

$$dt = \frac{\omega r^2}{1 - \omega^2 r^2}d\theta'. \tag{3.7.10}$$

Substituting this into the metric, we are left with the purely spatial metric

$$ds^2 = -dr^2 - \frac{r^2}{1 - \omega^2 r^2} d\theta'^2. \tag{3.7.11}$$

The factor of $(1 - \omega^2 r^2)^{-1} = \gamma^2$ in the $d\theta'^2$ term is simply the expected Lorentz-contraction factor. In other words, the circumference is, as expected, greater than $2\pi r$ by a factor of γ.

Does the metric (Equation 3.7.11) represent the same non-Euclidean spatial geometry that A, rotating with the disk, would determine by meterstick measurements? Yes and no. It *can* be interpreted as the one that A would determine by radar measurements. That is, if A measures a round-trip travel time dt for a light signal between points separated by coordinate distances dr and $d\theta'$, then A can say that the spatial separation is $\frac{dt}{2}$, and such measurements will be described correctly by Equation 3.7.11. Physical meter-sticks, however, present some problems. Meter-sticks rotating with the disk are subject to Coriolis and centrifugal forces, and this problem can't be avoided simply by making the meter-sticks infinitely rigid, because infinitely rigid objects are forbidden by relativity. In fact, these forces will inevitably be strong enough to destroy any meter stick that is brought out to r = $\frac{1}{\omega}$, where the speed of the disk becomes equal to the speed of light.

It might appear that we could now define a global coordinate

$$T = t + \frac{\omega r^2}{1 - \omega^2 r^2} \theta', \tag{3.7.12}$$

interpreted as a time coordinate that was synchronized in a consistent way for all points on the disk. The trouble with this interpretation becomes evident when we imagine driving a car around the circumference of the disk, at a speed slow enough so that there is negligible time dilation of the car's dashboard clock relative to the clocks tied to the disk. Once the car gets back to its original position, θ' has increased by $2pi$, so it is no longer possible for the car's clock to be synchronized with the clocks tied to the disk. We conclude that it is not possible to synchronize clocks in a rotating frame of reference; if we try to do it, we will inevitably have to have a discontinuity somewhere. This problem is present even locally, as demonstrated by the possibility of measuring the Sagnac effect with apparatus that is small compared to the disk. The only reason we were able to get away with time synchronization in order to establish the metric in Equation 3.7.11 is that all the physical manifestations of the impossibility of synchronization, e.g., the *Sagnac effect*, are proportional to the area of the region in which synchronization is attempted. Since we were only synchronizing two nearby points, the area enclosed by the light rays was zero.

Example 17: GPS

As a practical example, the GPS system is designed mainly to allow people to find their positions relative to the rotating surface of the earth (although it can also be used by space vehicles). That is, they are interested in their (r, θ', ϕ) coordinates. The frame of reference defined by these coordinates is referred to as ECEF, for Earth-Centered, Earth-Fixed.

The system requires synchronization of the atomic clocks carried aboard the satellites, and this synchronization also needs to be extended to the (less accurate) clocks built into the receiver units. It is impossible to carry out such a synchronization globally in the rotating frame in order to create coordinates (T, r, θ', ϕ). If we tried, it would result in discontinuities (see problem 8).

Instead, the GPS system handles clock synchronization in coordinates (t, r, θ', ϕ), as in Equation 3.7.7. These are known as the Earth-Centered Inertial (ECI) coordinates. The t coordinate in this system is not the one that users at neighboring points on the earth's surface would establish if they carried out clock synchronization using electromagnetic signals. It is simply the time coordinate of the nonrotating frame of reference tied to the earth's center. Conceptually, we can imagine this time coordinate as one that is established by sending out an electromagnetic "tick-tock" signal from the earth's center, with each satellite correcting the phase of the signal based on the propagation time inferred from its own r. In reality, this is accomplished by communication with a master control station in Colorado Springs, which communicates with the satellites via relays at Kwajalein, Ascension Island, Diego Garcia, and Cape Canaveral.

Example 18: Einstein's goof, in the rotating frame

Example 10 recounted Einstein's famous mistake in predicting that a clock at the pole would experience a time dilation relative to a clock at the equator, and the empirical test of this fact by Alley et al. using atomic clocks. The perfect cancellation of gravitational and kinematic time dilations might seem fortuitous, but it fact it isn't. When we transform into the frame rotating along with the earth, there is no longer any kinematic effect at all, because neither clock is moving. In this frame, the surface of the earth's oceans is an equipotential, so the gravitational time dilation vanishes as well, assuming both clocks are at sea level.

> In the transformation to the rotating frame, the metric picks up a $d\theta' \, dt$ term, but since both clocks are fixed to the earth's surface, they have $d\theta' = 0$, and there is no Sagnac effect.

Impossibility of Rigid Rotation, even with External Forces

The determination of the spatial metric with rulers at rest relative to the disk is appealing because of its conceptual simplicity compared to complicated procedures involving radar, and this was presumably why Einstein presented the concept using ruler measurements in his 1916 paper laying out the general theory of relativity.[12] In an effort to recover this simplicity, we could propose using external forces to compensate for the centrifugal and Coriolis forces to which the rulers would be subjected, causing them to stay straight and maintain their correct lengths. Something of this kind is carried out with the large mirrors of some telescopes, which have active systems that compensate for gravitational deflections and other effects. The first issue to worry about is that one would need some way to monitor a ruler's length and straightness. The monitoring system would presumably be based on measurements with beams of light, in which case the physical rulers themselves would become superfluous.

> **Note**
>
> The paper is reproduced in the back of the book, and the relevant part is in Appendix A.

In addition, we would need to be able to manipulate the rulers in order to place them where we wanted them, and these manipulations would include angular accelerations. If such a thing was possible, then it would also amount to a loophole in the resolution of the Ehrenfest paradox. Could Ehrenfest's rotating disk be accelerated and decelerated with help from external forces, which would keep it from contorting into a potato chip? The problem we run into with such a strategy is one of clock synchronization. When it was time to impart an angular acceleration to the disk, all of the control systems would have to be activated simultaneously. But we have already seen that global clock synchronization cannot be realized for an object with finite area, and therefore there is a logical contradiction in this proposal. This makes it impossible to apply rigid angular acceleration to the disk, but not necessarily the rulers, which could in theory be one-dimensional.

References

[9] Relativistic description of a rotating disk, Am. J. Phys. 43 (1975) 869

[10] Space, Time, and Coordinates in a Rotating World, www.phys.uu.nl/igg/dieks

[11] P. Ehrenfest, Gleichf¨ormige Rotation starrer K¨orper und Relativitätstheorie, Z. Phys. 10 (1909) 918, available in English translation at en.wikisource.org.

3.8: The Metric in General Relativity

So far we've considered a variety of examples in which the metric is predetermined. This is not the case in general relativity. For example, Einstein published general relativity in 1915, but it was not until 1916 that Schwarzschild found the metric for a spherical, gravitating body such as the sun or the earth.

When masses are present, finding the metric is analogous to finding the electric field made by charges, but the interpretation is more difficult. In the electromagnetic case, the field is found on a preexisting background of space and time. In general relativity, there is no preexisting geometry of spacetime. The metric tells us how to find distances in terms of our coordinates, but the coordinates themselves are completely arbitrary. So what does the metric even mean? This was an issue that caused Einstein great distress and confusion, and at one point, in 1914, it even led him to publish an incorrect, dead-end theory of gravity in which he abandoned coordinate-independence.

With the benefit of hindsight, we can consider these issues in terms of the general description of measurements in relativity given in section 3.4:

1. We can tell whether events and world-lines are incident.
2. We can do measurements in local Lorentz frames.

The Hole Argument

The main factor that led Einstein to his false start is known as the hole argument. Suppose that we know about the distribution of matter throughout all of spacetime, including a particular region of finite size — the "hole" — which contains no matter. By analogy with other classical field theories, such as electromagnetism, we expect that the metric will be a solution to some kind of differential equation, in which matter acts as the source term. We find a metric $g(x)$ that solves the field equations for this set of sources, where x is some set of coordinates. Now if the field equations are coordinate-independent, we can introduce a new set of coordinates x', which is identical to x outside the hole, but differs from it on the inside. If we reexpress the metric in terms of these new coordinates as $g'(x')$, then we are guaranteed that $g'(x')$ is also a solution. But furthermore, we can substitute x for x', and $g'(x)$ will still be a solution. For outside the hole there is no difference between the primed and unprimed quantities, and inside the hole there is no mass distribution that has to match the metric's behavior on a point-by-point basis.
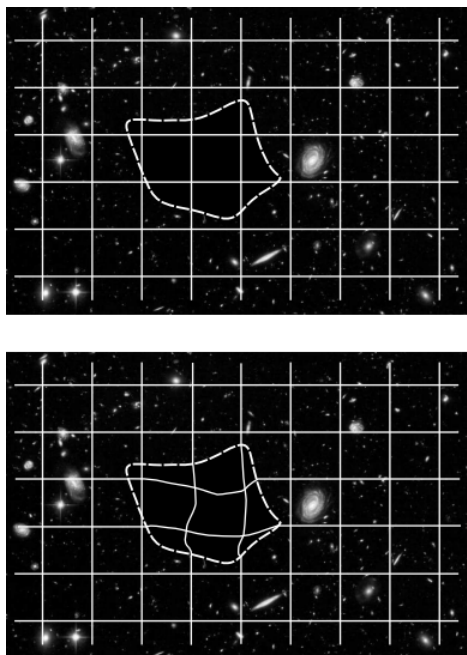
Figure 3.8.1 - Einstein's hole argument.

We conclude that in any coordinate-invariant theory, it is impossible to uniquely determine the metric inside such a hole. Einstein initially decided that this was unacceptable, because it showed a lack of determinism; in a classical theory such as general relativity, we ought to be able to predict the evolution of the fields, and it would seem that there is no way to predict the metric inside the hole. He eventually realized that this was an incorrect interpretation. The only type of global observation that general relativity lets

relations. For example, if two test particles sent into the region collide at a point **x** inside the hole, then changing the point's name
to **x′** doesn't change the observable fact that they collided.

A Machian Paradox

Another type of argument that made Einstein suffer is also resolved by a correct understanding of measurements, this time the use
of measurements in local Lorentz frames. The earth is in hydrostatic equilibrium, and its equator bulges due to its rotation. Suppose
that the universe was empty except for two planets, each rotating about the line connecting their centers.[13] Since there are no stars
or other external points of reference, the inhabitants of each planet have no external reference points against which to judge their
rotation or lack of rotation. They can only determine their rotation, Einstein said, relative to the other planet. Now suppose that one
planet has an equatorial bulge and the other doesn't. This seems to violate determinism, since there is no cause that could produce
the differing effect. The people on either planet can consider themselves as rotating and the other planet as stationary, or they can
describe the situation the other way around. Einstein believed that this argument proved that there could be no difference between
the sizes of the two planets' equatorial bulges.
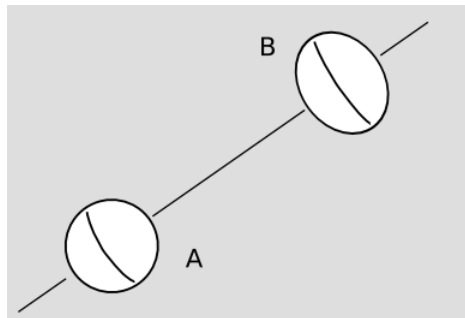
Figure 3.8.2 - A paradox? Planet A has no equatorial bulge, but B does. What cause produces this effect? Einstein reasoned that the
cause couldn't be B's rotation, because each planet rotates relative to the other.

> **Note**
>
> The example is described in Einstein's paper "The Foundation of the General Theory of Relativity." An excerpt, which
> includes the example, is given in Appendix A.

The flaw in Einstein's argument was that measurements in local Lorentz frames do allow one to make a distinction between
rotation and a lack of rotation. For example, suppose that scientists on planet A notice that their world has no equatorial bulge,
while planet B has one. They send a space probe with a clock to B, let it stay on B's surface for a few years, and then order it to
return. When the clock is back in the lab, they compare it with another clock that stayed in the lab on planet A, and they find that
less time has elapsed according to the one that spent time on B's surface. They conclude that planet B is rotating more quickly than
planet A, and that the motion of B's surface was the cause of the observed time dilation. This resolution of the apparent paradox
depends specifically on the Lorentzian form of the local geometry of spacetime; it is not available in, e.g., Cartan's curved-
spacetime description of Newtonian gravity.

Einstein's original, incorrect use of this example sprang from his interest in the ideas of the physicist and philosopher Ernst Mach.
Mach had a somewhat ill-defined idea that since motion is only a well-defined notion when we speak of one object moving relative
to another object, the inertia of an object must be caused by the influence of all the other matter in the universe. Einstein referred to
this as Mach's principle. Einstein's false starts in constructing general relativity were frequently related to his attempts to make his
theory too "Machian." Section 8.3 discusses an alternative, more Machian theory of gravity proposed by Brans and Dicke in 1951.

3.9: Interpretation of Coordinate Independence

This section discusses some of the issues that arise in the interpretation of coordinate independence. It can be skipped on a first reading.

Is Coordinate Independence Obvious?

One often hears statements like the following from relativists: "Coordinate independence isn't really a physical principle. It's merely an obvious statement about the relationship between mathematics and the physical universe. Obviously the universe doesn't come equipped with coordinates. We impose those coordinates on it, and the way in which we do so can never be dictated by nature." The impressionable reader who is tempted to say, "Ah, yes, that is obvious," should consider that it was far from obvious to Newton ("Absolute, true and mathematical time, of itself, and from its own nature flows equably without regard to anything external . . . "), nor was it obvious to Einstein. Levi-Civita nudged Einstein in the direction of coordinate independence in 1912. Einstein tried hard to make a coordinate-independent theory, but for reasons described in section 3.6, he convinced himself that that was a dead end. In 1914-15 he published theories that were not coordinate independent, which you will hear relativists describe as "obvious" dead ends because they lack any geometrical interpretation. It seems to me that it takes a highly refined intuition to regard as intuitively "obvious" an issue that Einstein struggled with like Jacob wrestling with Elohim.

Is Coordinate Independence Trivial?

It has also been alleged that coordinate independence is trivial. To gauge the justice of this complaint, let's distinguish between two reasons for caring about coordinate independence:

1. Coordinate independence tells us that when we solve problems, we should avoid writing down any equations in notation that isn't manifestly intrinsic, and avoid interpreting those equations as if the coordinates had intrinsic meaning. Violating this advice doesn't guarantee that you've made a mistake, but it makes it much harder to tell whether or not you have.
2. Coordinate independence can be used as a criterion for judging whether a particular theory is likely to be successful.

Nobody questions the first justification. The second is a little trickier. Laying out the general theory systematically in a 1916 paper,[14] Einstein wrote "The general laws of nature are to be expressed by equations which hold good for all the systems of coordinates, that is, are covariant with respect to any substitutions whatever (generally covariant)." In other words, he was explaining why, with hindsight, his 1914-1915 coordinate-dependent theory had to be a dead end.

The only trouble with this is that Einstein's way of posing the criterion didn't quite hit the nail on the head mathematically. As Hilbert famously remarked, "Every boy in the streets of Göttingen understands more about four-dimensional geometry than Einstein. Yet, in spite of that, Einstein did the work and not the mathematicians." What Einstein had in mind was that a theory like Newtonian mechanics not only lacks coordinate independence, but would also be impossible to put into a coordinate-independent form without making it look hopelessly complicated and ugly, like putting lipstick on a pig. But Kretschmann showed in 1917 that any theory could be put in coordinate independent form, and Cartan demonstrated in 1923 that this could be done for Newtonian mechanics in a way that didn't come out particularly ugly. Physicists today are more apt to pose the distinction in terms of "background independence" (meaning that a theory should not be phrased in terms of an assumed geometrical background) or lack of a "prior geometry" (meaning that the curvature of spacetime should come from the solution of field equations rather than being imposed by fiat). But these concepts as well have resisted precise mathematical formulation.[15] My feeling is that this general idea of coordinate independence or background independence is like the equivalence principle: a crucial conceptual principle that doesn't lose its importance just because we can't put it in a mathematical box with a ribbon and a bow. For example, string theorists take it as a serious criticism of their theory that it is not manifestly background independent, and one of their goals is to show that it has a background independence that just isn't obvious on the surface.

Coordinate Independence as a Choice of Gauge

It is instructive to consider coordinate independence from the point of view of a field theory. Newtonian gravity can be described in three equivalent ways: as a gravitational field $\mathbf{g}$, as a gravitational potential ϕ, or as a set of gravitational field lines. The field lines are never incident on one another, and locally the field satisfies Poisson's equation. The electromagnetic field has polarization properties different from those of the gravitational field, so we describe it using either the two fields ($\mathbf{E}, \mathbf{B}$), a pair of potentials,[16] or two sets of field lines. There are similar incidence conditions and local field equations (Maxwell's equations). Gravitational fields in relativity have polarization properties unknown to Newton, but the situation is qualitatively similar to the two foregoing cases. Now consider the analogy between electromagnetism and relativity. In electromagnetism, it is the fields that are directly

observable, so we expect the potentials to have some extrinsic properties. We can, for example, redefine our electrical ground, $\Phi \to \Phi + C$, without any observable consequences. As discussed in more detail in Section 5.6, it is even possible to modify the electromagnetic potentials in an entirely arbitrary and nonlinear way that changes from point to point in spacetime. This is called a gauge transformation. In relativity, the gauge transformations are the smooth coordinate transformations. These gauge transformations distort the field lines without making them cut through one another.
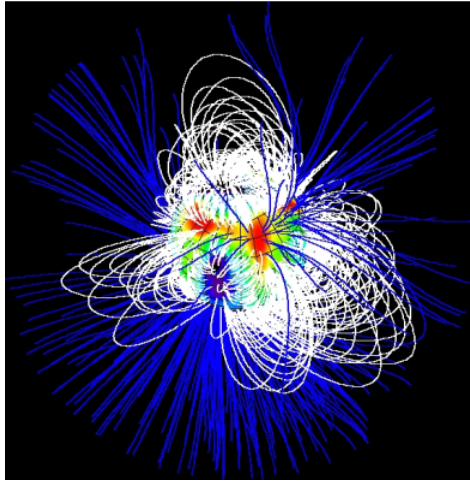
Figure 3.9.1: Since magnetic field lines can never intersect, a magnetic field pattern contains coordinate-independent information in the form of the knotting of the lines. This figure shows the magnetic field pattern of the star SU Aurigae, as measured by Zeeman-Doppler imaging (Petit at al.). White lines represent magnetic field lines that close upon themselves in the immediate vicinity of the star; blue lines are those that extend out into the interstellar medium.

> **Note**
>
> There is the familiar electrical potential ϕ, measured in volts, but also a vector potential $\mathbf{A}$, which you may or may not have encountered. Briefly, the electric field is given not by $-\nabla\phi$ but by $-\nabla\phi - \frac{\partial A}{\partial t}$, while the magnetic field is the curl of $\mathbf{A}$. This is introduced at greater length in section 4.2.

References

[15] Giulini, "Some remarks on the notions of general covariance and background independence," arxiv.org/abs/gr-qc/0603087v1

3.E: Differential Geometry (Exercises)

1. Consider a spacetime that is locally exactly like the standard Lorentzian spacetime described in ch. 2, but that has a global structure differing in the following way from the one we have implicitly assumed. This spacetime has global property G: Let two material particles have world-lines that coincide at event A, with some nonzero relative velocity; then there may be some event B in the future light-cone of A at which the particles' world-lines coincide again. This sounds like a description of something that we would expect to happen in curved spacetime, but let's see whether that is necessary. We want to know whether this violates the flat-space properties L1-L5 in Appendix C, if those properties are taken as local.

   a. Demonstrate that it does not violate them, by using a model in which space "wraps around" like a cylinder.
   b. Now consider the possibility of interpreting L1-L5 as *global* statements. Do spacetimes with property G always violate L3 if L3 is taken globally?

2. Usually in relativity we pick units in which c = 1. Suppose, however, that we want to use SI units. The convention is that coordinates are written with upper indices, so that, fixing the usual Cartesian coordinates in 1+1 dimensions of spacetime, an infinitesimal displacement between two events is notated (ds^t, ds^x). In SI units, the two components of this vector have different units, which may seem strange but is perfectly legal. Describe the form of the metric, including the units of its elements. Describe the lower-index vector ds_a.

3. (a) Explain why the following expressions ain't got good grammar: U_{aa}, $x_a y^a$, $p^a - q^a$. (Recall our notational convention that Latin indices represent abstract indices, so that it would not make sense, for example, to interpret U_{aa} as U's ath diagonal element rather than as an implied sum.)
   (b) Which of these could also be nonsense in terms of units?

4. Suppose that a mountaineer describes her location using coordinates (θ, ϕ, h), representing colatitude, longitude, and altitude. Infer the units of the components of ds^a and of the elements of g_{ab} and g^{ab}. Given that the units of mechanical work should be newton-meters (example 5), infer the components of a force vector F_a and its upper-index version F^a.

5. Generalize Figure 2.1.8 (2) to three dimensions.

6. Suppose you have a collection of pencils, some of which have been sharpened more times than others so that they they're shorter. You toss them all on the floor in random orientations, and you're then allowed to slide them around but not to rotate them. Someone asks you to make up a definition of whether or not a given set of three pencils "cancels." If all pencils are treated equally (i.e., order doesn't matter), and if we respect the rotational invariance of Euclidean geometry, then you will be forced to reinvent vector addition and define cancellation of pencils **p**, **q**, and **r** as **p** + **q** + **r** = 0. Do something similar with "pencil" replaced by "an oriented pairs of lines as in Figure 2.1.8 (2).

7. Describe the quantity g^a_a. (Note the repeated index.)

8. Example 17 discusses the discontinuity that would result if one attempted to define a time coordinate for the GPS system that was synchronized globally according to observers in the rotating frame, in the sense that neighboring observers could verify the synchronization by exchanging electromagnetic signals. Calculate this discontinuity at the equator, and estimate the resulting error in position that would be experienced by GPS users.

9. Resolve the following paradox.
   Equation [3] claims to give the metric obtained by an observer on the surface of a rotating disk. This metric is shown to lead to a non-Euclidean value for the ratio of the circumference of a circle to its radius, so the metric is clearly non-Euclidean. Therefore a local observer should be able to detect violations of the Pythagorean theorem.
   And yet this metric was originally derived by a series of changes of coordinates, starting from the Euclidean metric in polar coordinates, as derived in example 8. Section 3.4 argued that the intrinsic measurements available in relativity are not capable of detecting an arbitrary smooth, one-to-one change of coordinates. This contradicts our earlier conclusion that there are locally detectable violations of the Pythagorean theorem.

10. This problem deals with properties of the metric [3].

    a. A pulse of collimated light is emitted from the center of the disk in a certain direction. Does the spatial track of the pulse form a geodesic of this metric?
    b. Characterize the behavior of the geodesics near r = $\frac{1}{\omega}$.
    c. An observer at rest with respect to the surface of the disk proposes to verify the non-Euclidean nature of the metric by doing local tests in which right triangles are formed out of laser beams, and violations of the Pythagorean theorem are detected. Will this work?

11. In the early decades of relativity, many physicists were in the habit of speaking as if the Lorentz transformation described what an observer would actually "see" optically, e.g., with an eye or a camera. This is not the case, because there is an additional effect due to optical aberration: observers in different states of motion disagree about the direction from which a light ray originated. This is analogous to the situation in which a person driving in a convertible observes raindrops falling from the sky at an angle, even if an observer on the sidewalk sees them as falling vertically. In 1959, Terrell and Penrose independently provided correct analyses,[17] showing that in reality an object may appear contracted, expanded, or rotated, depending on whether it is approaching the observer, passing by, or receding. The case of a sphere is especially interesting. Consider the following four cases:

   A. The sphere is not rotating. The sphere's center is at rest. The observer is moving in a straight line.
   B. The sphere is not rotating, but its center is moving in a straight line. The observer is at rest.
   C. The sphere is at rest and not rotating. The observer moves around it in a circle whose center coincides with that of the sphere.
   D. The sphere is rotating, with its center at rest. The observer is at rest.

   Penrose showed that in case A, the outline of the sphere is still seen to be a circle, although regions on the sphere's surface appear distorted.
   What can we say about the generalization to cases B, C, and D?

12. This problem involves a relativistic particle of mass m which is also a wave, as described by quantum mechanics. Let c = 1 and $\hbar$ = 1 throughout. Starting from the de Broglie relations E = ω and p = k, where k is the wavenumber, find the dispersion relation connecting ω to k. Calculate the group velocity, and verify that it is consistent with the usual relations p = mγv and E = mγ for m > 0. What goes wrong if you instead try to associate v with the phase velocity?

References

[17] James Terrell, "Invisibility of the Lorentz Contraction," Physical Review 116 (1959) 1045. Roger Penrose, "The Apparent Shape of a Relativistically Moving Sphere," Proceedings of the Cambridge Philosophical Society 55 (1959) 139.

CHAPTER OVERVIEW

4: Tensors

We now have enough machinery to be able to calculate quite a bit of interesting physics, and to be sure that the results are actually meaningful in a relativistic context. The strategy is to identify relativistic quantities that behave as Lorentz scalars and Lorentz vectors, and then combine them in various ways. The notion of a tensor has been introduced earlier. A Lorentz scalar is a tensor of rank 0, and a Lorentz vector is a rank 1 tensor.

Thumbnail: Standard configuration of coordinate systems; for a Lorentz boost in the x-direction. (Public Domain; Gerd Kortemeyer).

4.1: Lorentz Scalars

A Lorentz scalar is a quantity that remains invariant under both spatial rotations and Lorentz boosts. Mass is a Lorentz scalar.[1] Electric charge is also a Lorentz scalar, as demonstrated to extremely high precision by experiments measuring the electrical neutrality of atoms and molecules to a relative precision of better than 10^{-20}; the electron in a hydrogen atom has typically velocities of about 1/100, and those in heavier elements such as uranium are highly relativistic, so any violation of Lorentz invariance would give the atoms a nonvanishing net electric charge.

> **Note**
>
> Some older books define mass as transforming according to m → γm, which can be made to give a self-consistent theory, but is ugly.

The time measured by a clock traveling along a particular world-line from one event to another is something that all observers will agree upon; they will simply note the mismatch with their own clocks. It is therefore a Lorentz scalar. This clock-time as measured by a clock attached to the moving body in question is often referred to as proper time, "proper" being used here in the somewhat archaic sense of "own" or "self," as in "The Vatican does not lie within Italy proper." Proper time, which we notate τ , can only be defined for timelike world-lines, since a lightlike or spacelike world-line isn't possible for a material clock.

More generally, when we express a metric as ds^2 = . . ., the quantity ds is a Lorentz scalar. In the special case of a timelike world-line, ds and dτ are the same thing. (In books that use a − + ++ metric, one has ds = − $d\tau$.)

Even more generally, affine parameters, which exist independent of any metric at all, are scalars. As a trivial example, if τ is a particular object's proper time, then τ is a valid affine parameter, but so is 2τ +7. Less trivially, a photon's proper time is always zero, but one can still define an affine parameter along its trajectory. We will need such an affine parameter, for example, in section 6.2, when we calculate the deflection of light rays by the sun, one of the early classic experimental tests of general relativity.

Another example of a Lorentz scalar is the pressure of a perfect fluid, which is often assumed as a description of matter in cosmological models.

> **Example 1: Infinitesimals and the clock "postulate"**
>
> At the beginning of chapter 3, I motivated the use of infinitesimals as useful tools for doing differential geometry in curved spacetime. Even in the context of special relativity, however, infinitesimals can be useful. One way of expressing the proper time accumulated on a moving clock is
>
> $$\begin{aligned} s &= \int ds \\ &= \int \sqrt{g_{ij}dx^i dx^j} \\ &= \int \sqrt{1 - \left(\frac{dx}{dt}\right)^2 - \left(\frac{dy}{dt}\right)^2 - \left(\frac{dz}{dt}\right)^2}\, dt, \end{aligned}$$
>
> which only contains an explicit dependence on the clock's velocity, not its acceleration. This is an example of the clock "postulate" referred to in the remark at the end of homework problem 1. Note that the clock postulate only applies in the limit of a small clock. This is represented in the above equation by the use of infinitesimal quantities like dx.

4.2: Four-vectors (Part 1)

The Velocity and Acceleration Four-vectors

Our basic Lorentz vector is the spacetime displacement dx^i. Any other quantity that has the same behavior as dx^i under rotations and boosts is also a valid Lorentz vector. Consider a particle moving through space, as described in a Lorentz frame. Since the particle may be subject to nongravitational forces, the Lorentz frame cannot be made to coincide (except perhaps momentarily) with the particle's rest frame. If dx^i is not lightlike, then the corresponding infinitesimal proper time interval dτ is nonzero. As with Newtonian three-vectors, dividing a four-vector by a Lorentz scalar produces another quantity that transforms as a four-vector, so dividing the infinitesimal displacement by a nonzero infinitesimal proper time interval, we have the four-velocity vector

$$v^i = \frac{dx^i}{d\tau} \tag{4.2.1}$$

whose components in a Lorentz coordinate system are

$$(\gamma, \gamma u^1, \gamma u^2, \gamma u^3) \tag{4.2.2}$$

where (u^1, u^2, u^3) is the ordinary three-component velocity vector as defined in classical mechanics. The four-velocity's squared magnitude $v^i v_i$ is always exactly 1, even though the particle is not moving at the speed of light. (If it were moving at the speed of light, we would have $d\tau = 0$, and v would be undefined.)

When we hear something referred to as a "vector," we usually take this is a statement that it not only transforms as a vector, but also that it adds as a vector. But we have already seen in Section 2.3 that even collinear velocities in relativity do not add linearly; therefore they clearly cannot add linearly when dressed in the clothing of four-vectors. We've also seen in Section 2.5 that the combination of non-collinear boosts is noncommutative, and is generally equivalent to a boost plus a spatial rotation; this is also not consistent with linear addition of four-vectors. At the risk of beating a dead horse, a four-velocity's squared magnitude is always 1, and this is not consistent with being able to add four-velocity vectors.

Example 2: A zero velocity vector?

Suppose an object has a certain four-velocity v^i in a certain frame of reference. Can we transform into a different frame in which the object is at rest, and its four-velocity is zero?

Solution

No. In general, the Lorentz transformation preserves the magnitude of vectors, so it can never transform a vector with a zero magnitude into one with nonzero magnitude. Since this is a material object (not a ray of light) we can transform into a frame in which the object is at rest, but an object at rest does not have a vanishing four-velocity. It has a four-velocity of (1, 0, 0, 0).

Example 2 suggests a nice way of thinking about velocity vectors, which is that every velocity vector represents a potential observer. An observer is a material object, and therefore has a timelike velocity vector. This observer writes her own velocity vector as (1, 0, 0, 0), i.e., as the unit vector in the timelike direction. Often when we see an expression involving a velocity vector, we can interpret it as describing a measurement taken by a specific observer.

Example 3: Orthogonality as simultaneity

In a space where the inner product can be negative, orthogonality doesn't mean what our euclidean intuition thinks it means. For example, a lightlike vector can be orthogonal to itself — a situation that never occurs in a euclidean space. Suppose we have a timelike vector **t** and a spacelike one **x**. What would it mean for **t** and **x** to be orthogonal, with **t** · **x** = 0?

Solution

Since **t** is timelike, we can make a unit vector $\hat{\mathbf{t}} = \frac{\mathbf{t}}{|\mathbf{t}|}$ out of it, and interpret $\hat{\mathbf{t}}$ as the velocity vector of some hypothetical observer. We then know that in that observer's frame, $\hat{\mathbf{t}}$ is simply a unit vector along the time axis. It now becomes clear that **x** must be parallel to the x axis, i.e., it represents a displacement between two events that this observer considers to be simultaneous.

This is an example of the idea that expressions involving velocity vectors can be interpreted as measurements taken by a certain observer. The expression **t** · **x** = 0 can be interpreted as meaning that according to an observer whose world-line is

---

> tangent to **t, x** represents a relationship of simultaneity.

The four-acceleration is found by taking a second derivative with respect to proper time. Its squared magnitude is only approximately equal to minus the squared magnitude of the Newtonian acceleration three-vector, in the limit of small velocities.

Example 4: Constant acceleration

Suppose a spaceship moves so that the acceleration is judged to be the constant value a by an observer on board. Find the motion x(t) as measured by an observer in an inertial frame.

Solution

Let τ stand for the ship's proper time, and let dots indicate derivatives with respect to τ. The ship's velocity has magnitude 1, so $$\dot{t}^2 - \dot{x}^2 = 1 \ldotp$$

An observer who is instantaneously at rest with respect to the ship judges is to have a four-acceleration (0, a, 0, 0) (because the low-velocity limit applies). The observer in the (t, x) frame agrees on the magnitude of this vector, so

$$\ddot{t}^2 - \ddot{x}^2 = -a^2. \tag{4.2.3}$$

The solution of these differential equations is $t = \frac{1}{a}\sinh a\tau, \ x = \frac{1}{a}\cosh a\tau$, and eliminating τ gives

$$x = \frac{1}{a}\sqrt{1 + a^2 t^2}. \tag{4.2.4}$$

As t approaches infinity, $\frac{dx}{dt}$ approaches the speed of light.

The Momentum Four-vector

Definition for a Material Particle

If we hope to find something that plays the role of momentum in relativity, then the momentum three-vector probably needs to be generalized to some kind of four-vector. If so, then the law of conservation of momentum will be valid regardless of one's frame of reference, which is necessary.[2]

Note

We are not guaranteed that this is the right way to proceed, since the converse is not true: some three-vectors such as the electric and magnetic fields are embedded in rank-2 tensors in more complicated ways than this. See section 4.2.

If we are to satisfy the correspondence principle then the relativistic definition of momentum should probably look as much as possible like the nonrelativistic one. Earlier, we defined the velocity four-vector in the case of a particle whose dx^i is not lightlike. Let's assume for the moment that it makes sense to think of mass as a scalar. As with Newtonian three-vectors, multiplying a Lorentz scalar by a four-vector vector produces another quantity that transforms as a four-vector. We therefore conjecture that the four-momentum of a material particle can be defined as p^i = mv^i, which in Lorentz coordinates is $(m\gamma, m\gamma v^1, m\gamma v^2, m\gamma v^3)$ There is no *a priori* guarantee that this is right, but it's the most reasonable thing to guess. It needs to be checked against experiment, and also for consistency with the other parts of our theory.

The spacelike components look like the classical momentum vector multiplied by a factor of γ, the interpretation being that to an observer in this frame, the moving particle's inertia is increased relative to its value in the particle's rest frame. Such an effect is indeed observed experimentally. This is why particle accelerators are so big and expensive. As the particle approaches the speed of light, γ diverges, so greater and greater forces are needed in order to produce the same acceleration. In relativistic scattering processes with material particles, we find empirically that the four-momentum we've defined is conserved, which confirms that our conjectures above are valid, and in particular that the quantity we're calling m can be treated as a Lorentz scalar, and this is what all physicists do today. The reader is cautioned, however, that up until about 1950, it was common to use the word "mass" for the combination mγ (which is what occurs in the Lorentz-coordinate form of the momentum vector), while referring to m as the "rest mass." This archaic terminology is only used today in some popular-level books and low-level school textbooks.

Equivalence of Mass and Energy

The momentum four-vector has locked within it the reason for Einstein's famous $E = mc^2$, which in our relativistic units becomes simply $E = m$. To see why, consider the experimentally measured inertia of a physical object made out of atoms. The subatomic particles are all moving, and many of the velocities, e.g., the velocities of the electrons, are quite relativistic. This has the effect of increasing the experimentally determined inertial mass of the whole object, by a factor of γ averaged over all the particles — even though the masses of the individual particles are invariant Lorentz scalars. (This same increase must also be observed for the gravitational mass, based on the equivalence principle as verified by Eötvös experiments.)

Now if the object is heated, the velocities will increase on the average, resulting in a further increase in its mass. Thus, a certain amount of heat energy is equivalent to a certain amount of mass. But if heat energy contributes to mass, then the same must be true for other forms of energy. For example, suppose that heating leads to a chemical reaction, which converts some heat into electromagnetic binding energy. If one joule of binding energy did not convert to the same amount of mass as one joule of heat, then this would allow the object to spontaneously change its own mass, and then by conservation of momentum it would have to spontaneously change its own velocity, which would clearly violate the principle of relativity. We conclude that mass and energy are equivalent, both inertially and gravitationally. In relativity, neither is separately conserved; the conserved quantity is their sum, referred to as the mass-energy, E. An alternative derivation, by Einstein, is given in example 16.

Energy is the Timelike Component of the Four-momentum

The Lorentz transformation of a zero vector is always zero. This means that the momentum four-vector of a material object can't equal zero in the object's rest frame, since then it would be zero in all other frames as well. So for an object of mass m, let its momentum four-vector in its rest frame be (f(m), 0, 0, 0), where f is some function that we need to determine, and f can depend only on m since there is no other property of the object that can be dynamically relevant here. Since conservation laws are additive, f has to be f(m) = km for some universal constant k. In where c = 1, k is unitless. Since we want to recover the appropriate Newtonian limit for massive bodies, and since vt = 1 in that limit, we need k = 1. Transforming the momentum four-vector from the particle's rest frame into some other frame, we find that the timelike component is no longer m. We interpret this as the relativistic mass-energy, E.

Since the momentum four-vector was obtained from the magnitude-1 velocity four-vector through multiplication by m, its squared magnitude $p^i p_i$ is equal to the square of the particle's mass. Writing p for the magnitude of the momentum three-vector, and E for the mass-energy, we find the useful relation $m^2 = E^2 - p^2$. We take this to be the relativistic *definition* the mass of any particle, including one whose dx^i is lightlike.

Particles Traveling at c

The definition of four-momentum as $p^i = mv^i$ only works for particles that move at less than c. For those that move at c, the four-velocity is undefined. As we'll see in example 6, this class of particles is exactly those that are massless. As shown in section 1.5, the three-momentum of a light wave is given by p = E. The fact that this momentum is nonzero implies that for light $p^i = mv^i$ represents an indeterminate form. The fact that this momentum equals E is consistent with our definition of mass as $m^2 = E^2 - p^2$.

Mass is Not Additive

Since the momentum four-vector p^a is additive, and our definition of mass as $p^a p_a$ depends on the vector in a nonlinear way, it follows that mass is not additive (even for particles that are not interacting but are simply considered collectively).

> ### Example 5: Mass of two light waves
>
> Let the momentum of a certain light wave be $(p_t, p_x) = (E, E)$, and let another such wave have momentum (E, −E). The total momentum is (2E, 0). Thus this pair of massless particles has a collective mass of 2E.

> ### Example 6: Massless particles travel at c
>
> We demonstrate this by showing that if we suppose the opposite, then there are two different consequences, either of which would be physically unacceptable.
>
> When a particle does have a nonvanishing mass, we have
>
> $$\lim_{\frac{E}{m} \to \infty} |v| = \lim_{\frac{E}{m} \to \infty} \frac{|p|}{E} = 1. \tag{4.2.5}$$

Thus if we had a massless particle with $|v| \neq 1$, its behavior would be different from the limiting behavior of massive particles. But this is physically unacceptable because then we would have a magic method for detecting arbitrarily small masses such as $10^{-10000000000}$ kg. We don't actually know that the photon, for example, is exactly massless; see example 13.

Furthermore, suppose that a massless particle had $|v| < 1$ in the frame of some observer. Then some other observer could be at rest relative to the particle. In such a frame, the particle's three-momentum **p** is zero by symmetry, since there is no preferred direction for it. Then $E^2 = p^2 + m^2$ is zero as well, so the particle's entire energy-momentum four-vector is zero. But a four-vector that vanishes in one frame also vanishes in every other frame. That means we're talking about a particle that can't undergo scattering, emission, or absorption, and is therefore undetectable by any experiment. This is physically unacceptable because we don't consider phenomena (e.g., invisible fairies) to be of physical interest if they are undetectable even in principle.

Example 7: Gravitational redshifts

Since a photon's energy E is equivalent to a certain gravitational mass m, photons that rise or fall in a gravitational field must lose or gain energy, and this should be observed as a redshift or blueshift in the frequency. We expect the change in gravitational potential energy to be $E\Delta\phi$, giving a corresponding opposite change in the photon's energy, so that $\frac{\Delta E}{E} = \Delta\varphi$. In metric units, this becomes $\frac{\Delta E}{E} = \frac{\Delta\varphi}{c^2}$, and in the field near the Earth's surface we have $\frac{\Delta E}{E} = \frac{gh}{c^2}$. This is the same result that was found in section 1.5 based only on the equivalence principle, and verified experimentally by Pound and Rebka as described in section 1.5.

Example 8: Constraints on polarization

We observe that electromagnetic waves are always polarized transversely, never longitudinally. Such a constraint can only apply to a wave that propagates at c. If it applied to a wave that propagated at less than c, we could move into a frame of reference in which the wave was at rest. In this frame, all directions in space would be equivalent, and there would be no way to decide which directions of polarization should be permitted. For a wave that propagates at c, there is no frame in which the wave is at rest (see section 3.4).

Example 9: Relativistic work-energy theorem

In Einstein's original 1905 paper on relativity, he assumed without providing any justification that the Newtonian work-energy relation W = Fd was valid relativistically. One way of justifying this is that we can construct a simple machine with a mechanical advantage A and a reduction of motion by $\frac{1}{A}$, with these ratios being exact relativistically.[3] One can then calculate, as Einstein did,

$$W = \int \frac{dp}{dt}dx = \int \frac{dp}{dv}\frac{dx}{dt}dv = m(\gamma - 1), \qquad (4.2.6)$$

which is consistent with our result for E as a function of γ if we equate it to $E(\gamma) - E(1)$.

[3] For an explicit example, see bit.ly/1aUXIa8.

Example 10: The Dirac sea

A great deal of physics can be derived from the T.H. White's principle that "whatever is not forbidden in compulsory" — originally intended for ants but applied to particles by Gell-Mann. In quantum mechanics, any process that is not forbidden by a conservation law is supposed to occur. The relativistic relation $E = \pm\sqrt{p^2 + m^2}$ has two roots, a positive one and a negative one. The positive-energy and negative-energy states are separated by a no-man's land of width 2m, so no continuous classical process can lead from one side to the other. But quantum-mechanically, if an electron exists with energy $E = +\sqrt{p^2 + m^2}$, it should be able to make a quantum leap into a state with $E = -\sqrt{p^2 + m^2}$, emitting the energy difference of 2E in the form of photons. Why doesn't this happen? One explanation is that the states with E < 0 are all already occupied. This is the "Dirac sea," which we now interpret as being full of electrons. A vacancy in the sea manifests itself as an antielectron.

Example 11: Massive neutrinos

Neutrinos were long thought to be massless, but are now believed to have masses in the eV range. If they had been massless, they would always have had to propagate at the speed of light. Although they are now thought to have mass, that mass is six orders of magnitude less than the MeV energy scale of the nuclear reactions in which they are produced, so all neutrinos observed in experiments are moving at velocities very close to the speed of light.

Example 12: No radioactive decay of massless particles

A photon cannot decay into an electron and a positron, $\gamma \rightarrow e^+ + e^-$, in the absence of a charged particle to interact with. To see this, consider the process in the frame of reference in which the electron-positron pair has zero total momentum. In this frame, the photon must have had zero (three-)momentum, but a photon with zero momentum must have zero energy as well. This means that conservation of relativistic four-momentum has been violated: the timelike component of the four-momentum is the mass-energy, and it has increased from 0 in the initial state to at least $2mc^2$ in the final state.

To demonstrate the consistency of the theory, we can arrive at the same conclusion by a different method. Whenever a particle has a small mass (small compared to its energy, say), it must travel at close to c. It must therefore have a very large time dilation, and will take a very long time to undergo radioactive decay. In the limit as the mass approaches zero, the time required for the decay approaches infinity. Another way of saying this is that the rate of radioactive decay must be fixed in terms of proper time, but there is no such thing as proper time for a massless particle. Thus it is not only this specific process that is forbidden, but any radioactive decay process involving a massless particle.

There are various loopholes in this argument. The question is investigated more thoroughly by Fiore and Modanese.[4]

---

[4] http://arxiv.org/abs/hep-th/9508018

Example 13: Massive photons

Continuing in the same vein as example 11, we can consider the possibility that the photon has some nonvanishing mass. A 2003 experiment by Luo et al.[5] has placed a limit of about 10^{-54} kg on this mass. This is incredibly small, but suppose that future experimental work using improved techniques shows that the mass is less than this, but actually nonzero. A naive reaction to this scenario is that it would shake relativity to its core, since relativity is based upon the assumption that the speed of light is a constant, whereas for a massive particle it need not be constant. But this is a misinterpretation of the role of c in relativity. As should be clear from the approach taken in section 2.2, c is primarily a geometrical property of spacetime, not a property of light.

In reality, such a discovery would be more of a problem for particle physicists than for relativists, as we can see by the following sketch of an argument. Imagine two charged particles, at rest, interacting via an electrical attraction. Quantum mechanics describes this as an exchange of photons. Since the particles are at rest, there is no source of energy, so where do we get the energy to make the photons? The Heisenberg uncertainty principle, $\Delta E \Delta t \gtrsim h$, allows us to steal this energy, provided that we give it back within a time Δt. This time limit imposes a limit on the distance the photons can travel, but by using photons of low enough energy, we can make this distance limit as large as we like, and there is therefore no limit on the range of the force. But suppose that the photon has a mass. Then there is a minimum mass-energy mc^2 required in order to create a photon, the maximum time is h/mc^2, and the maximum range is h/mc. Refining these crude arguments a little, one finds that exchange of zeromass particles gives a force that goes like $\frac{1}{r^2}$, while a nonzero mass results in $\frac{e^{-\mu r}}{r^2}$, where $\mu^{-1} = \frac{\hbar}{mc}$. For the photon, the best current mass limit corresponds to $\mu^{-1} \gtrsim 10^{11}$ m, so the deviation from $\frac{1}{r^2}$ would be difficult to measure in earthbound experiments.

Now Gauss's law is a specific characteristic of $\frac{1}{r^2}$ fields. It would be violated slightly if photons had mass. We would have to modify Maxwell's equations, and it turns out[6] that the necessary change to Gauss's law would be of the form $\nabla \cdot \mathbf{E} = (\ldots)\rho - (\ldots)\mu^2\Phi$, where Φ is the electrical potential, and $(\ldots)$ indicates factors that depend on the choice of units. This tells us that Φ, which in classical electromagnetism can only be measured in terms of differences between different points in space, can now be measured in absolute terms. Gauge symmetry has been broken. But gauge symmetry is indispensible in creating well-behaved relativistic field theories, and this is the reason that, in general, particle physicists have a hard time with forces arising from the exchange of massive particles. The hypothetical Higgs particle, which may be observed at the Large Hadron Collider in the near future, is essentially a mechanism for wriggling out of this difficulty in the case of the massive W

and Z particles that are responsible for the weak nuclear force; the mechanism cannot, however, be extended to allow a massive photon.

[5] Luo et al., "New Experimental Limit on the Photon Rest Mass with a Rotating Torsion Balance," Phys. Rev. Lett. 90 (2003) 081801. The interpretation of such experiments is difficult, and this paper attracted a series of comments. A weaker but more universally accepted bound is 8×10^{-52} kg, Davis, Goldhaber, and Nieto, Phys. Rev. Lett. 35 (1975) 1402.

[6] Goldhaber and Nieto, "Terrestrial and Extraterrestrial Limits on The Photon Mass," Rev. Mod. Phys. 43 (1971) 277

Example 14: Dust and radiation in cosmological models

In cosmological models, one needs an equation of state that relates the pressure P to the mass-energy density ρ. The pressure is a Lorentz scalar. The mass-energy density is not (since massenergy is just the timelike component of a particular vector), but in a coordinate system without any net flow of mass, we can approximate it as one.

The early universe was dominated by radiation. A photon in a box contributes a pressure on each wall that is proportional to $|p^{\mu}|$, where μ is a spacelike index. In thermal equilibrium, each of these three degrees of freedom carries an equal amount of energy, and since momentum and energy are equal for a massless particle, the average momentum along each axis is equal to $\frac{1}{3}$ E. The resulting equation of state is $P = \frac{1}{3}\rho$. As the universe expanded, the wavelengths of the photons expanded in proportion to the stretching of the space they occupied, resulting in $\lambda \propto a^{-1}$, where a is a distance scale describing the universe's intrinsic curvature at a fixed time. Since the number density of photons is diluted in proportion to a^{-3}, and the mass per photon varies as a^{-1}, both ρ and P vary as a^{-4}.

Cosmologists refer to noninteracting, nonrelativistic materials as "dust," which could mean many things, including hydrogen gas, actual dust, stars, galaxies, and some forms of dark matter. For dust, the momentum is negligible compared to the mass-energy, so the equation of state is P = 0, regardless of ρ. The massenergy density is dominated simply by the mass of the dust, so there is no red-shift scaling of the a^{-1} type. The mass-energy density scales as a^{-3}. Since this is a less steep dependence on a than the a^{-4}, there was a point, about a thousand years after the Big Bang, when matter began to dominate over radiation. At this point, the rate of expansion of the universe made a transition to a qualitatively different behavior resulting from the change in the equation of state.

In the present era, the universe's equation of state is dominated by neither dust nor radiation but by the cosmological constant (see section 8.1). Figure 4.2.1 shows the evolution of the size of the universe for the three different regimes. Some of the simpler cases are derived starting in section 8.2.
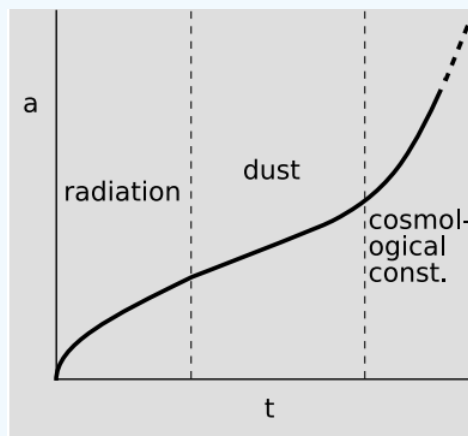
Figure 4.2.1

4.3: Four-vectors (Part 2)

The Frequency Vector and the Relativistic Doppler Shift

In the spirit of index-gymnastics notation, frequency is to time as the wavenumber k = $\frac{1}{\lambda}$ is to space, so when treating waves relativistically it is natural to conjecture that there is a four-frequency f_a made by assembling (f, **k**), which behaves as a Lorentz vector. This is correct, since we already know that ∂_a transforms as a covariant vector, and for a scalar wave of the form

$$A = A_o \exp[2\pi i f_a x^a] \tag{4.3.1}$$

the partial derivative operator is identical to multiplication by $2\pi f_a$.

As an application, consider the relativistic Doppler shift of a light wave. For simpicity, let's restrict ourselves to one spatial dimension. For a light wave, $f = k$, so the frequency vector in 1+1 dimensions is simply (f, f). Putting this through a Lorentz transformation, we find

$$f' = (1+v)\gamma f = \sqrt{\frac{1+v}{1-v}} f, \tag{4.3.2}$$

where the second form displays more clearly the symmetic form of the relativistic relationship, such that interchanging the roles of source and observer is equivalent to flipping the sign of v. That is, the relativistic version only depends on the *relative* motion of the source and the observer, whereas the Newtonian one also depends on the source's motion relative to the medium (i.e., relative to the preferred frame in which the waves have the "right" velocity). In Newtonian mechanics, we have f' = (1 + v)f for a moving observer. Relativistically, there is also a time dilation of the oscillation of the source, providing an additional factor of γ.

This analysis is extended to 3+1 dimensions in problem 11.

> ### Example 15: Ives-Stilwell experiments
>
> The relativistic Doppler shift differs from the nonrelativistic one by the time-dilation factor γ, so that there is still a shift even when the relative motion of the source and the observer is perpendicular to the direction of propagation. This is called the transverse Doppler shift. Einstein suggested this early on as a test of relativity. However, such experiments are difficult to carry out with high precision, because they are sensitive to any error in the alignment of the 90-degree angle. Such experiments were eventually performed, with results that confirmed relativity,[7] but one-dimensional measurements provided both the earliest tests of the relativistic Doppler shift and the most precise ones to date. The first such test was done by Ives and Stilwell in 1938, using the following trick. The relativistic expression
>
> $$S_v = \sqrt{\frac{(1+v)}{(1-v)}} \tag{4.3.3}$$
>
> for the Doppler shift has the property that
>
> $$S_v S_{-v} = 1 \tag{4.3.4}$$
>
> which differs from the nonrelativistic result of
>
> $$(1+v)(1-v) = 1 - v^2. \tag{4.3.5}$$
>
> One can therefore accelerate an ion up to a relativistic speed, measure both the forward Doppler shifted frequency f_f and the backward one f_b, and compute $\sqrt{f_f f_b}$. According to relativity, this should exactly equal the frequency f_o measured in the ion's rest frame.
>
> In a particularly exquisite modern version of the Ives-Stilwell idea,[8] Saathoff et al. circulated Li$^+$ ions at v = .064 in a storage ring. An electron-cooler technique was used in order to reduce the variation in velocity among ions in the beam. Since the identity $S_v S_{-v} = 1$ is independent of v, it was not necessary to measure v to the same incredible precision as the frequencies; it was only necessary that it be stable and well-defined. The natural line width was 7 MHz, and other experimental effects broadened it further to 11 MHz. By curve-fitting the line, it was possible to achieve results good to a few tenths of a MHz. The resulting frequencies, in units of MHz, were:
>
> $$f_f = 582490203.44 \pm .09$$

$$f_b = 512671442.9 \pm 0.5$$

$$\sqrt{f_f f_b} = 546466918.6 \pm 0.3$$

$$f_o = 546466918.8 \pm 0.4 \text{ (from previous experimental work)}$$

The spectacular agreement with theory has made this experiment a lightning rod for anti-relativity kooks.

If one is searching for small deviations from the predictions of special relativity, a natural place to look is at high velocities. IvesStilwell experiments have been performed at velocities as high as 0.84, and they confirm special relativity.[9]

---

[7] See, e.g., Hasselkamp, Mondry, and Scharmann, Zeitschrift f¨ur Physik A: Hadrons and Nuclei 289 (1979) 151.

[8] G. Saathoff et al., "Improved Test of Time Dilation in Relativity," Phys. Rev. Lett. 91 (2003) 190403. A publicly available description of the experiment is given in Saathoff's PhD thesis, www.mpi-hd.mpg.de/ato/homes/saathoff/ diss-saathoff.pdf.

[9] MacArthur et al., Phys. Rev. Lett. 56 (1986) 282 (1986)

2

Earlier, we showed that the celebrated $E = mc^2$ follows directly from the form of the Lorentz transformation. An alternative derivation was given by Einstein in one of his classic 1905 papers laying out the theory of special relativity; the paper is short, and is reproduced in English translation in Appendix A of this book. Having laid the groundwork of four-vectors and relativistic Doppler shifts, we can give an even shorter version of Einstein's argument. The discussion is also streamlined by restricting the discussion to 1+1 dimensions and by invoking photons.

Suppose that a lantern, at rest in the lab frame, is floating weightlessly in outer space, and simultaneously emits two pulses of light in opposite directions, each with energy $\frac{E}{2}$ and frequency f. By symmetry, the momentum of the pulses cancels, and the lantern remains at rest. An observer in motion at velocity v relative to the lab sees the frequencies of the beams shifted to f' = $(1 \pm v)\gamma f$. The effect on the energies of the beams can be found purely classically, by transforming the electric and magnetic fields to the moving frame, but as a shortcut we can apply the quantum-mechanical relation $E_{ph} = hf$ for the energies of the photons making up the beams. The result is that the moving observer finds the total energy of the beams to be not E but $(\frac{E}{2})(1 + v)\gamma + (\frac{E}{2})(1 - v)\gamma = E\gamma$.

Both observers agree that the lantern had to use up some of the energy stored in its fuel in order to make the two pulses. But the moving observer says that in addition to this energy E, there was a further energy $E(\gamma - 1)$. Where could this energy have come from? It must have come from the kinetic energy of the lantern. The lantern's velocity remained constant throughout the experiment, so this decrease in kinetic energy seen by the moving observer must have come from a decrease in the lantern's inertial mass — hence the title of Einstein's paper, "Does the inertia of a body depend upon its energy content?"

To figure out how much mass the lantern has lost, we have to decide how we can even define mass in this new context. In Newtonian mechanics, we had $K = (\frac{1}{2})mv^2$, and by the correspondence principle this must still hold in the low-velocity limit. Expanding $E(\gamma - 1)$ in a Taylor series, we find that it equals $E(\frac{v^2}{2}) + \ldots$, and in the low-velocity limit this must be the same as $\Delta K = (\frac{1}{2})\Delta mv^2$, so $\Delta m = E$. Reinserting factors of c to get back to nonrelativistic units, we have $E = \Delta mc^2$.

A Non-example: Electric and Magnetic Fields

It is fairly easy to see that the electric and magnetic fields cannot be the spacelike parts of two four-vectors. Consider the arrangement shown in Figure 4.2.2 (1). We have two infinite trains of moving charges superimposed on the same line, and a single charge alongside the line. Even though the line charges formed by the two trains are moving in opposite directions, their currents don't cancel. A negative charge moving to the left makes a current that goes to the right, so in frame 1, the total current is twice that contributed by either line charge.
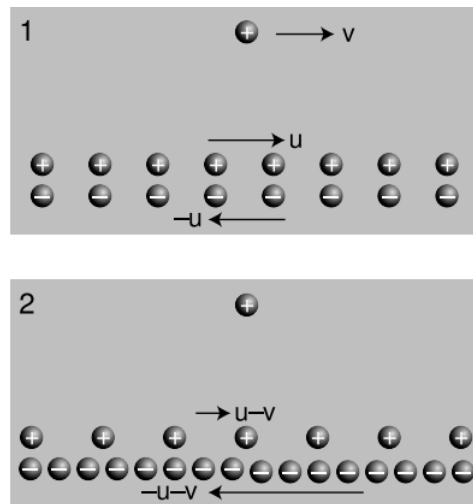
Figure 4.3.2 - Magnetism is a purely relativistic effect.

In frame 1 the charge densities of the two line charges cancel out, and the electric field experienced by the lone charge is therefore zero. Frame 2 shows what we'd see if we were observing all this from a frame of reference moving along with the lone charge. Both line charges are in motion in both frames of reference, but in frame 1, the line charges were moving at equal speeds, so their Lorentz contractions were equal, and their charge densities canceled out. In frame 2, however, their speeds are unequal. The positive charges are moving more slowly than in frame 1, so in frame 2 they are less contracted. The negative charges are moving more quickly, so their contraction is greater now. Since the charge densities don't cancel, there is an electric field in frame 2, which points into the wire, attracting the lone charge.

We appear to have a logical contradiction here, because an observer in frame 2 predicts that the charge will collide with the wire, whereas in frame 1 it looks as though it should move with constant velocity parallel to the wire. Experiments show that the charge does collide with the wire, so to maintain the Lorentz-invariance of electromagnetism, we are forced to invent a new kind of interaction, one between moving charges and other moving charges, which causes the acceleration in frame 2. This is the magnetic interaction, and if we hadn't known about it already, we would have been forced to invent it. That is, magnetism is a purely relativistic effect. The reason a relativistic effect can be strong enough to stick a magnet to a refrigerator is that it breaks the delicate cancellation of the extremely large electrical interactions between electrically neutral objects.

Although the example shows that the electric and magnetic fields do transform when we change from one frame to another, it is easy to show that they do not transform as the spacelike parts of a relativistic four-vector. This is because transformation between frames 1 and 2 is along the axis parallel to the wire, but it affects the components of the fields perpendicular to the wire. The electromagnetic field actually transforms as a rank-2 tensor.

The Electromagnetic Potential Four-vector

An electromagnetic quantity that *does* transform as a four-vector is the potential. In section 3.7, I mentioned the fact, which may or may not already be familiar to you, that whereas the Newtonian gravitational field's polarization properties allow it to be described using a single scalar potential ϕ or a single vector field $\mathbf{g} = -\nabla\phi$, the pair of electromagnetic fields (E, B) needs a pair of potentials, $\mathbf{\Phi}$ and $\mathbf{A}$. It's easy to see that $\mathbf{\Phi}$ can't be a Lorentz scalar. Electric charge q is a scalar, so if $\mathbf{\Phi}$ were a scalar as well, then the product q$\mathbf{\Phi}$ would be a scalar. But this is equal to the energy of the charged particle, which is only the timelike component of the energy-momentum four-vector, and therefore not a Lorentz scaler itself. This is a contradiction, so $\mathbf{\Phi}$ is not a scalar.

To see how to fit $\mathbf{\Phi}$ into relativity, consider the nonrelativistic quantum mechanical relation q$\mathbf{\Phi}$ = hf for a charged particle in a potential $\mathbf{\Phi}$. Since f is the timelike component of a four-vector in relativity, we need $\mathbf{\Phi}$ to be the timelike component of some four vector, Ab. For the spacelike part of this four-vector, let's write $\mathbf{A}$, so that $A_b = (\mathbf{\Phi}, \mathbf{A})$. We can see by the following argument that this mysterious A must have something to do with the magnetic field.

Consider the example of Figure 4.2.3 from a quantum-mechanical point of view. The charged particle q has wave properties, but let's say that it can be well approximated in this example as following a specific trajectory. This is like the ray approximation to wave optics. A light ray in classical optics follows Fermat's principle, also known as the principle of least time, which states that the ray's path from point A to point B is one that extremizes the optical path length (essentially the number of oscillations). The reason for this is that the ray approximation is only an approximation. The ray actually has some width, which we can visualize as a

bundle of neighboring trajectories. Only if the trajectory follows Fermat's principle will the interference among the neighboring paths be constructive. The classical optical path length is found by integrating $\mathbf{k} \cdot ds$, where $\mathbf{k}$ is the wavenumber. To make this relativistic, we need to use the frequency four-vector to form $f_b \, dx^b$, which can also be expressed as $f_b v^b \, d\tau = \gamma(f - \mathbf{k} \cdot \mathbf{v}) \, d\tau$. If the charge is at rest and there are no magnetic fields, then the quantity in parentheses is $f = \frac{E}{h} = \left(\frac{q}{h}\right)\Phi$. The correct relativistic generalization is clearly $f_b = \left(\frac{q}{h}\right)A_b$.
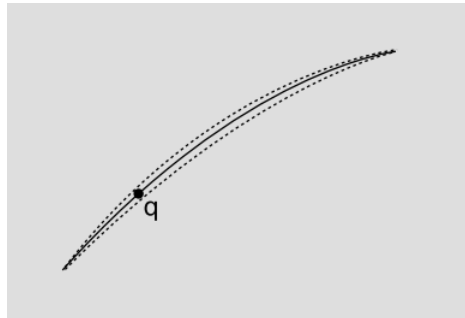
Figure 4.3.3 - The charged particle follows a trajectory that extremizes $\int f_b \, dx^b$ compared to other nearby trajectories. Relativistically, the trajectory should be understood as a world-line in 3+1-dimensional spacetime.

Since A_b's spacelike part, $\mathbf{A}$, results in the velocity-dependent effects, we conclude that $\mathbf{A}$ is a kind of potential that relates to the magnetic field, in the same way that the potential Φ relates to the electric field. $\mathbf{A}$ is known as the vector potential, and the relation between the potentials and the fields is

$$\mathbf{E} = -\nabla\Phi - \frac{\partial \mathbf{A}}{\partial t}$$
$$\mathbf{B} = \nabla \mathbf{A}.$$

An excellent discussion of the vector potential from a purely classical point of view is given in the classic *Feynman Lectures*.[10] Figure 4.2.4 shows an example.
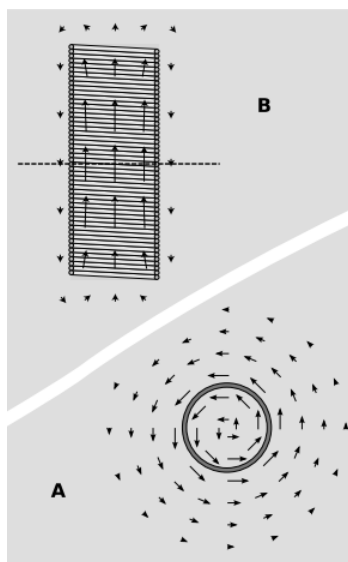
Figure 4.3.4 - The magnetic field (top) and vector potential (bottom) of a solenoid. The lower diagram is in the plane cutting through the waist of the solenoid, as indicated by the dashed line in the upper diagram. For an infinite solenoid, the magnetic field is uniform on the inside and zero on the outside, while the vector potential is proportional to r on the inside and to $\frac{1}{r}$ on the outside.

References

[10] *The Feynman Lectures on Physics*, Feynman, Leighton, and Sands, Addison Wesley Longman, 1970

4.4: The Tensor Transformation Laws

We may wish to represent a vector in more than one coordinate system, and to convert back and forth between the two representations. In general relativity, the transformation of the coordinates need not be linear, as in the Lorentz transformations; it can be any smooth, one-to-one function. For simplicity, however, we start by considering the one-dimensional case, and by assuming the coordinates are related in an affine manner,

$$x'^{\mu} = ax^{\mu} + b. \tag{4.4.1}$$

The addition of the constant b is merely a change in the choice of origin, so it has no effect on the components of the vector, but the dilation by the factor a gives a change in scale, which results in $v'^{\mu} = av^{\mu}$ for a contravariant vector. In the special case where v is an infinitesimal displacement, this is consistent with the result found by implicit differentiation of the coordinate transformation. For a contravariant vector, $v'_{\mu} = \frac{1}{a}v_{\mu}$. Generalizing to more than one dimension, and to a possibly nonlinear transformation, we have

$$v'^{\mu} = v^{\kappa}\frac{\partial x'^{\mu}}{\partial x^{\kappa}} \tag{4.4.2}$$

$$v'_{\mu} = v_{\kappa}\frac{\partial x'^{\kappa}}{\partial x^{\mu}} \tag{4.4.3}$$

Note the inversion of the partial derivative in one equation compared to the other. Because these equations describe a change from one coordinate system to another, they clearly depend on the coordinate system, so we use Greek indices rather than the Latin ones that would indicate a coordinate-independent equation. Note that the letter μ in these equations always appears as an index referring to the new coordinates, κ to the old ones. For this reason, we can get away with dropping the primes and writing, e.g., $v^{\mu} = \frac{v^{\kappa}\partial x'^{\mu}}{\partial x^{\kappa}}$ rather than v' counting on context to show that v^{μ} is the vector expressed in the new coordinates, v^{κ} in the old ones. This becomes especially natural if we start working in a specific coordinate system where the coordinates have names. For example, if we transform from coordinates (t, x, y, z) to (a, b, c, d), then it is clear that v^{t} is expressed in one system and v^{c} in the other.

> **Exercise 4.4.1**
>
> Recall that the gauge transformations allowed in general relativity are not just any coordinate transformations; they must be (1) smooth and (2) one-to-one. Relate both of these requirements to the features of the vector transformation laws above.

In Equation 4.4.3, μ appears as a subscript on the left side of the equation, but as a superscript on the right. This would appear to violate our rules of notation, but the interpretation here is that in expressions of the form $\frac{\partial}{\partial x^{i}}$ and $\frac{\partial}{\partial x_{i}}$, the superscripts and subscripts should be understood as being turned upside-down. Similarly, Equation 4.4.2 appears to have the implied sum over κ written ungrammatically, with both κ's appearing as superscripts. Normally we only have implied sums in which the index appears once as a superscript and once as a subscript. With our new rule for interpreting indices on the bottom of derivatives, the implied sum is seen to be written correctly. This rule is similar to the one for analyzing the units of derivatives written in Leibniz notation, with, e.g., $\frac{d^{2}x}{dt^{2}}$ having units of meters per second squared. That is, the flipping of the indices like this is required for consistency so that everything will work out properly when we change our units of measurement, causing all our vector components to be rescaled.

A quantity v that transforms according to Equations 4.4.2 or 4.4.3 is referred to as a rank-1 tensor, which is the same thing as a vector.

> **Example 17: The identity transformation**
>
> In the case of the identity transformation $x'^{\mu} = x^{\mu}$, Equation 4.4.2 clearly gives v' = v, since all the mixed partial derivatives $\frac{\partial x'^{\mu}}{\partial x^{\kappa}}$ with $\mu \neq \kappa$ are zero, and all the derivatives for $\kappa = \mu$ equal 1.
>
> In Equation 4.4.3, it is tempting to write
>
> $$\frac{\partial x^{\kappa}}{\partial x'^{\mu}} = \frac{1}{\frac{\partial x'^{\mu}}{\partial x^{\kappa}}} \quad (wrong!), \tag{4.4.4}$$

but this would give infinite results for the mixed terms! Only in the case of functions of a single variable is it possible to flip derivatives in this way; it doesn't work for partial derivatives. To evaluate these partial derivatives, we have to invert the transformation (which in this example is trivial to accomplish) and then take the partial derivatives.

The metric is a rank-2 tensor, and transforms analogously:

$$g_{\mu\nu} = g_{\kappa\lambda} \frac{\partial x^\kappa}{\partial x'^\mu} \frac{\partial x^\lambda}{\partial x'^\nu} \tag{4.4.5}$$

writing g rather than g' on the left, because context makes the distinction clear).

Exercise 4.4.2

Self-check: Write the similar expressions for $g^{\mu\nu}, g^\mu_\nu$, and g^ν_μ, which are entirely determined by the grammatical rules for writing superscripts and subscripts. Interpret the case of a rank-0 tensor.

Example 18: An accelerated coordinate system?

Let's see the effect on Lorentzian metric g of the transformation

$$t' = t \qquad x' = x + \frac{1}{2}at^2. \tag{4.4.6}$$

The inverse transformation is

$$t = t' \qquad x = x' - \frac{1}{2}at'^2. \tag{4.4.7}$$

The tensor transformation law gives

$$g'_{t't'} = 1 - (at')^2 \tag{4.4.8}$$

$$g'_{x'x'} = -1 \tag{4.4.9}$$

$$g'_{x't'} = -at'. \tag{4.4.10}$$

Clearly something bad happens at $at' = \pm 1$, when the relative velocity surpasses the speed of light: the t' component of the metric vanishes and then reverses its sign. This would be physically unreasonable if we viewed this as a transformation from observer A's Lorentzian frame into the accelerating reference frame of observer B aboard a spaceship who feels a constant acceleration. Several things prevent such an interpretation:

1. B cannot exceed the speed of light.
2. Even before B gets to the speed of light, the coordinate t' cannot correspond to B's proper time, which is dilated.
3. Due to time dilation, A and B do not agree on the rate at which B is accelerating. If B measures her own acceleration to be a', A will judge it to be a < a', with a → 0 as B approaches the speed of light.

There is nothing invalid about the coordinate system (t', x'), but neither does it have any physically interesting interpretation.

Example 19: Physically meaningful constant acceleration

To make a more physically meaningful version of example 18, we need to use the result of example 4. The somewhat messy derivation of the coordinate transformation is given by Semay.[11] The result is

$$t' = \left(x + \frac{1}{a}\right) \sinh at \tag{4.4.11}$$

$$x' = \left(x + \frac{1}{a}\right) \cosh at \tag{4.4.12}$$

Applying the tensor transformation law gives (problem 7):

$$g'_{t't'} = (1 + ax')^2 \tag{4.4.13}$$

$$g'_{x'x'} = -1 \tag{4.4.14}$$

Unlike the result of example 18, this one never misbehaves. The closely related topic of a uniform gravitational field in general relativity is considered in problem 7.

---

[11] arxiv.org/abs/physics/0601179

Example 20: Accurate timing signals

The relation between the potential **A** and the fields **E** and **B** given in section 4.2 can be written in manifestly covariant form as

$$F_{ij} = \partial_{[i} A_{j]} \tag{4.4.15}$$

where F, called the electromagnetic tensor, is an antisymmetric rank-two tensor whose six independent components correspond in a certain way with the components of the **E** and **B** three-vectors. If F vanishes completely at a certain point in spacetime, then the linear form of the tensor transformation laws guarantees that it will vanish in all coordinate systems, not just one. The GPS system takes advantage of this fact in the transmission of timing signals from the satellites to the users. The electromagnetic wave is modulated so that the bits it transmits are represented by phase reversals of the wave. At these phase reversals, F vanishes, and this vanishing holds true regardless of the motion of the user's unit or its position in the earth's gravitational field. Cf. problem 17.

Example 21: Momentum wants a lower index

In example 5, we saw that once we arbitrarily chose to write ruler measurements in Euclidean three-space as Δx^a rather than Δx_a, it became natural to think of the Newtonian force threevector as "wanting" to be notated with a lower index. We can do something similar with the momentum 3- or 4-vector. The Lagrangian is a relativistic scalar, and in Lagrangian mechanics momentum is defined by $p_a = \frac{\partial L}{\partial v^a}$. The upper index in the denominator on the right becomes a lower index on the left by the same reasoning as was employed in the notation of the tensor transformation laws. Newton's second law shows that this is consistent with the result of example 5.

This page titled 4.4: The Tensor Transformation Laws is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Benjamin Crowell via source content that was edited to the style and standards of the LibreTexts platform.

4.5: Experimental Tests

Universality of Tensor Behavior

The techniques developed in this chapter allow us to make a variety of new predictions that can be tested by experiment. In general, the mathematical treatment of all observables in relativity as tensors means that all observables must obey the same transformation laws. This is an extremely strict statement, because it requires that a wide variety of physical systems show identical behavior. For example, we already mentioned in section 2.5 the 2007 Gravity Probe B experiment (discussed in detail on in sections 5.5 and 6.2), in which four gyroscopes aboard a satellite were observed to precess due to special- and general-relativistic effects. The gyroscopes were complicated electromechanical systems, but the predicted precession was entirely independent of these complications. We argued that if two different types of gyroscopes displayed different behaviors, then the resulting discrepancy would allow us to map out some mysterious vector field. This field would be a built-in characteristic of spacetime (not produced by any physical objects nearby), and since all observables in general relativity are supposed to be tensors, the field would have to transform as a tensor. Let's say that this tensor was of rank 1. Since the tensor transformation law is linear, a nonzero tensor can never be transformed into a vanishing tensor in another coordinate system. But by the equivalence principle, any special, local property of spacetime can be made to vanish by transforming into a free-falling frame of reference, in which the spacetime is has a generic Lorentzian geometry. The mysterious new field should therefore vanish in such a frame. This is a contradiction, so we conclude that different types of gyroscopes cannot differ in their behavior.

This is an example of a new way of stating the equivalence principle: there is no way to associate a preferred tensor field with spacetime.[12]

> **Note**
>
> This statement of the equivalence principle, along with the others we have encountered, is summarized in Appendix C.

Speed of Light Differing from c

In a Lorentz invariant theory, we interpret c as a property of the underlying spacetime, not of the particles that inhabit it. One way in which Lorentz invariance could be violated would be if different types of particles had different maximum velocities. In 1997, Coleman and Glashow suggested a sensitive test for such an effect.[13]

Assuming Lorentz invariance, a photon cannot decay into an electron and a positron (example 12):

$$\gamma \to e^+ + e^- \tag{4.5.1}$$

Suppose, however, that material particles have a maximum speed $c_m = 1$, while photons have a maximum speed $c_p > 1$. Then the photon's momentum four-vector, $(E, \frac{E}{c_p})$ is timelike, so a frame does exist in which its three-momentum is zero. The detection of cosmicray gammas from distant sources with energies on the order of 10 TeV puts an upper limit on the decay rate, implying

$$c_p - 1 \lesssim 10^{-15}. \tag{4.5.2}$$

An even more stringent limit can be put on the possibility of $c_p < 1$. When a charged particle moves through a medium at a speed higher than the speed of light in the medium, *Cerenkov radiation* results. If c_p is less than 1, then Cerenkov radiation could be emitted by high-energy charged particles in a vacuum, and the particles would rapidly lose energy. The observation of cosmic-ray protons with energies $\sim 10^8$ TeV requires

$$c_p - 1 \lesssim -10^{-23}. \tag{4.5.3}$$

Degenerate Matter

The straightforward properties of the momentum four-vector have surprisingly far-reaching implications for matter subject to extreme pressure, as in a star that uses up all its fuel for nuclear fusion and collapses. These implications were initially considered too exotic to be taken seriously by astronomers. For historical perspective, consider that in 1916, when Einstein published the theory of general relativity, the Milky Way was believed to constitute the entire universe; the "spiral nebulae" were believed to be inside it, rather than being similar objects exterior to it. The only types of stars whose structure was understood even vaguely were those that were roughly analogous to our own sun. (It was not known that nuclear fusion was their source of energy.) The term "white dwarf" had not been invented, and neutron stars were unknown.

An ordinary, smallish star such as our own sun has enough hydrogen to sustain fusion reactions for billions of years, maintaining an equilibrium between its gravity and the pressure of its gases. When the hydrogen is used up, it has to begin fusing heavier elements. This leads to a period of relatively rapid fluctuations in structure. Nuclear fusion proceeds up until the formation of elements as heavy as oxygen ($Z = 8$), but the temperatures are not high enough to overcome the strong electrical repulsion of these nuclei to create even heavier ones. Some matter is blown off, but finally nuclear reactions cease and the star collapses under the pull of its own gravity.

To understand what happens in such a collapse, we have to understand the behavior of gases under very high pressures. In general, a surface area A within a gas is subject to collisions in a time t from the n particles occupying the volume $V = Avt$, where v is the typical velocity of the particles. The resulting pressure is given by $P \sim \frac{npv}{V}$, where p is the typical momentum.

> **Definition**
>
> - **Nondegenerate gas**: In an ordinary gas such as air, the particles are nonrelativistic, so $v = \frac{p}{m}$, and the thermal energy per particle is $\frac{p^2}{2m} \sim kT$, so the pressure is $P \sim \frac{nkT}{V}$.
> - **Nonrelativistic, degenerate gas**: When a fermionic gas is subject to extreme pressure, the dominant effects creating pressure are quantum-mechanical. Because of the Pauli exclusion principle, the volume available to each particle is $\sim \frac{V}{n}$, so its wavelength is no more than $\sim (\frac{V}{n})^{1/3}$, leading to $p = \frac{h}{\lambda} \sim h(\frac{n}{V})^{1/3}$. If the speeds of the particles are still nonrelativistic, then $v = \frac{p}{m}$ still holds, so the pressure becomes $P \sim (\frac{h^2}{m})(\frac{n}{V})^{5/3}$.
> - **Relativistic, degenerate gas**: If the compression is strong enough to cause highly relativistic motion for the particles, then v ≈ c, and the result is $P \sim hc(\frac{n}{V})^{4/3}$.

As a star with the mass of our sun collapses, it reaches a point at which the electrons begin to behave as a degenerate gas, and the collapse stops. The resulting object is called a white dwarf. A white dwarf should be an extremely compact body, about the size of the Earth. Because of its small surface area, it should emit very little light. In 1910, before the theoretical predictions had been made, Russell, Pickering, and Fleming discovered that 40 Eridani B had these characteristics. Russell recalled: "I knew enough about it, even in these paleozoic days, to realize at once that there was an extreme inconsistency between what we would then have called 'possible' values of the surface brightness and density. I must have shown that I was not only puzzled but crestfallen, at this exception to what looked like a very pretty rule of stellar characteristics; but Pickering smiled upon me, and said: 'It is just these exceptions that lead to an advance in our knowledge,' and so the white dwarfs entered the realm of study!"

S. Chandrasekhar showed in that 1930's that there was an upper limit to the mass of a white dwarf. We will recapitulate his calculation briefly in condensed order-of-magnitude form. The pressure at the core of the star is P ∼ ρgr ∼ GM2/r4 , where M is the total mass of the star. The star contains roughly equal numbers of neutrons, protons, and electrons, so M = Knm, where m is the mass of the electron, n is the number of electrons, and K ≈ 4000. For stars near the limit, the electrons are relativistic. Setting the pressure at the core equal to the degeneracy pressure of a relativistic gas, we find that the Chandrasekhar limit is $\sim (\frac{hc}{G})^{3/2}(Km)^{-2} = 6M_{\odot}$. A less sloppy calculation gives something more like 1.4M. The selfconsistency of this solution is investigated in homework problem 15.
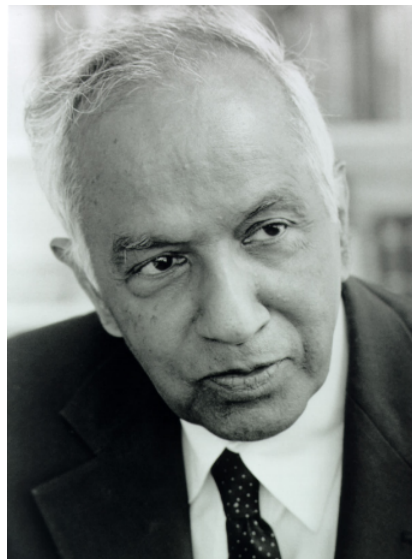
Figure 4.5.1 - Subrahmanyan Chandrasekhar (1910-1995)

What happens to a star whose mass is above the Chandrasekhar limit? As nuclear fusion reactions flicker out, the core of the star becomes a white dwarf, but once fusion ceases completely this cannot be an equilibrium state. Now consider the nuclear reactions

$$n \to p + e^- + \bar{\nu} \qquad (4.5.4)$$

$$p + e^- \to n + \nu, \qquad (4.5.5)$$

which happen due to the weak nuclear force. The first of these releases 0.8 MeV, and has a half-life of 14 minutes. This explains why free neutrons are not observed in significant numbers in our universe, e.g., in cosmic rays. The second reaction requires an *input* of 0.8 MeV of energy, so a free hydrogen atom is stable. The white dwarf contains fairly heavy nuclei, not individual protons, but similar considerations would seem to apply. A nucleus can absorb an electron and convert a proton into a neutron, and in this context the process is called electron capture. Ordinarily this process will only occur if the nucleus is neutron-deficient; once it reaches a neutronto-proton ratio that optimizes its binding energy, neutron capture cannot proceed without a source of energy to make the reaction go. In the environment of a white dwarf, however, there is such a source. The annihilation of an electron opens up a hole in the "Fermi sea." There is now an state into which another electron is allowed to drop without violating the exclusion principle, and the effect cascades upward. In a star with a mass above the Chandrasekhar limit, this process runs to completion, with every proton being converted into a neutron. The result is a *neutron star*, which is essentially an atomic nucleus (with Z = 0) with the mass of a star!

Observational evidence for the existence of neutron stars came in 1967 with the detection by Bell and Hewish at Cambridge of a mysterious radio signal with a period of 1.3373011 seconds. The signal's observability was synchronized with the rotation of the earth relative to the stars, rather than with legal clock time or the earth's rotation relative to the sun. This led to the conclusion that its origin was in space rather than on earth, and Bell and Hewish originally dubbed it LGM-1 for "little green men." The discovery of a second signal, from a different direction in the sky, convinced them that it was not actually an artificial signal being generated by aliens. Bell published the observation as an appendix to her PhD thesis, and it was soon interpreted as a signal from a neutron star. Neutron stars can be highly magnetized, and because of this magnetization they may emit a directional beam of electromagnetic radiation that sweeps across the sky once per rotational period — the "lighthouse effect." If the earth lies in the plane of the beam, a periodic signal can be detected, and the star is referred to as a pulsar. It is fairly easy to see that the short period of rotation makes it difficult to explain a pulsar as any kind of less exotic rotating object. In the approximation of Newtonian mechanics, a spherical body of density ρ, rotating with a period $T = \sqrt{\frac{3\pi}{G\rho}}$, has zero apparent gravity at its equator, since gravity is just strong enough to accelerate an object so that it follows a circular trajectory above a fixed point on the surface (problem 14). In reality, astronomical bodies of planetary size and greater are held together by their own gravity, so we have $T \gtrsim \frac{1}{\sqrt{G\rho}}$ for any body that does not fly apart spontaneously due to its own rotation. In the case of the Bell-Hewish pulsar, this implies $\rho \gtrsim 10^{10}$ kg/m^3, which is far larger than the density of normal matter, and also 10-100 times greater than the typical density of a white dwarf near the Chandrasekhar limit.

An upper limit on the mass of a neutron star can be found in a manner entirely analogous to the calculation of the Chandrasekhar limit. The only difference is that the mass of a neutron is much greater than the mass of an electron, and the neutrons are the only particles present, so there is no factor of K. Assuming the more precise result of $1.4M_\odot$ for the Chandrasekhar limit rather than our sloppy one, and ignoring the interaction of the neutrons via the strong nuclear force, we can infer an upper limit on the mass of a neutron star:

$$1.4M_\odot \left(\frac{Km_e}{m_n} \right)^2 \approx 5M_\odot \qquad (4.5.6)$$

The theoretical uncertainties in such an estimate are fairly large. Tolman, Oppenheimer, and Volkoff originally estimated it in 1939 as $0.7M_\odot$, whereas modern estimates are more in the range of 1.5 to $3M_\odot$. These are significantly lower than our crude estimate of $5M_\odot$, mainly because the attractive nature of the strong nuclear force tends to pull the star toward collapse. Unambiguous results are presently impossible because of uncertainties in extrapolating the behavior of the strong force from the regime of ordinary nuclei, where it has been relatively well parametrized, into the exotic environment of a neutron star, where the density is significantly different and no protons are present. There are a variety of effects that may be difficult to anticipate or to calculate. For example, Brown and Bethe found in 1994[14] that it might be possible for the mass limit to be drastically revised because of the process $e^- \to K^- + \nu_e$, which is impossible in free space due to conservation of energy, but might be possible in a neutron star. Observationally, nearly all neutron stars seem to lie in a surprisingly small range of mass, between 1.3 and $1.45M_\odot$, but in 2010 a neutron star with a mass of $1.97 \pm .04$ $M_\odot$ was discovered, ruling out most neutron-star models that included exotic matter.[15]

For stars with masses above the Tolman-Oppenheimer-Volkoff limit, theoretical predictions become even more speculative. A variety of bizarre objects has been proposed, including black stars, gravastars, quark stars, boson stars, Q-balls, and electroweak stars.

It seems likely, however, both on theoretical and observational grounds, that objects with masses of about 3 to 20 solar masses end up as black holes; see section 6.3.

References

[13] arxiv.org/abs/hep-ph/9703240

[14] H.A. Bethe and G.E. Brown, "Observational constraints on the maximum neutron star mass," Astrophys. J. 445 (1995) L129. G.E. Brown and H.A. Bethe, "A Scenario for a Large Number of Low-Mass Black Holes in the Galaxy," Astrophys. J. 423 (1994) 659. Both papers are available at adsabs.harvard.edu.

[15] Demorest et al., arxiv.org/abs/1010.5788v1.

4.6: Conservation Laws

No General Conservation Laws

Some of the first tensors we discussed were mass and charge, both rank-0 tensors, and the rank-1 momentum tensor, which contains both the classical energy and the classical momentum. Physicists originally decided that mass, charge, energy, and momentum were interesting because these things were found to be conserved. This makes it natural to ask how conservation laws can be formulated in relativity. We're used to stating conservation laws casually in terms of the amount of something in the whole universe, e.g., that classically the total amount of mass in the universe stays constant. Relativity does allow us to make physical models of the universe as a whole, so it seems as though we ought to be able to talk about conservation laws in relativity.

We can't.

First, how do we define "stays constant?" Simultaneity isn't well-defined, so we can't just take two snapshots, call them initial and final, and compare the total amount of, say, electric charge in each snapshot. This difficulty isn't insurmountable. As in figure a, we can arbitrarily pick out three-dimensional spacelike surfaces — one initial and one final — and integrate the charge over each one. A law of conservation of charge would say that no matter what spacelike surface we picked, the total charge on each would be the same.
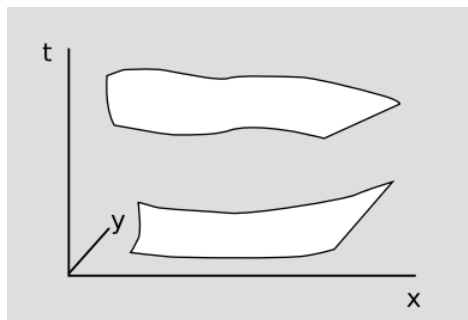
Figure 4.6.1 - Two spacelike surfaces.

Next there's the issue that the integral might diverge, especially if the universe was spatially infinite. For now, let's assume a spatially finite universe. For simplicity, let's assume that it has the topology of a three-sphere (see section 8.2 for reassurance that this isn't physically unreasonable), and we can visualize it as a two-sphere.
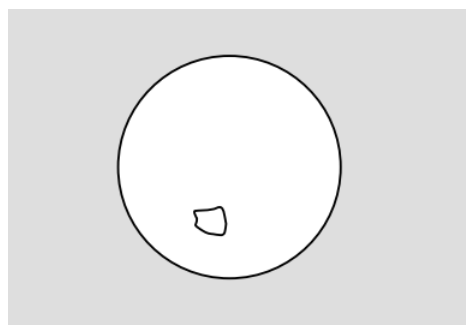
Figure 4.6.2 - We define a boundary around a region whose charge we want to measure.

In the case of the momentum four-vector, what coordinate system would we express it in? In general, we do not even expect to be able to define a smooth, well-behaved coordinate system that covers the entire universe, and even if we did, it would not make sense to add a vector expressed in that coordinate system at point A to another vector from point B; the best we could do would be to parallel-transport the vectors to one point and then add them, but parallel transport is path-dependent. (Similar issues occur with angular momentum.) For this reason, let's restrict ourselves to the easier case of a scalar, such as electric charge.
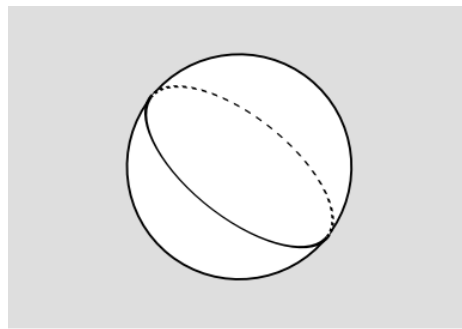
Figure 4.6.3 - This boundary cuts the sphere into equal parts.

But now we're in real trouble. How would we go about actually measuring the total electric charge of the universe? The only way to do it is to measure electric fields, and then apply Gauss's law. This requires us to single out some surface that we can integrate the flux over, as in Figure 4.5.2. This would really be a two-dimensional surface on the three-sphere, but we can visualize it as a one-dimensional surface — a closed curve — on the two-sphere. But now suppose this curve is a great circle, Figure 4.5.3. If we measure a nonvanishing total flux across it, how do we know where the charge is? It could be on either side.

The conclusion is that conservation laws only make sense in relativity under very special circumstances.[16] We do not have anything like over-arching, global principles of conservation. As an example of the appropriate special circumstances, section 6.2 shows how to define conserved quantities, which behave like energy and momentum, for the motion of a test particle in a particular metric that has a certain symmetry. This is generalized in section 7.1 to a general, global conservation law corresponding to every continuous symmetry of a spacetime.

> **Note**
>
> For another argument leading to the same conclusion, see section 7.5.

Conservation of Angular Momentum and Frame Dragging

Another special case where conservation laws work is that if the spacetime we're studying gets very flat at large distances from a small system we're studying, then we can define a far-away boundary that surrounds the system, measure the flux through that boundary, and find the system's charge. For such *asymptotic flatness* spacetimes, we can also get around the problems that crop up with conserved vectors, such as momentum. (Asymptotic flatness is discussed in more detail in section 7.4.) If the spacetime far away is nearly flat, then parallel transport loses its path-dependence, so we can unambiguously define a notion of parallel-transporting all the contributions to the flux to one arbitrarily chosen point P and then adding them. Asymptotic flatness also allows us to define an approximate notion of a global Lorentz frame, so that the choice of P doesn't matter.

As an example, Figure 4.5.4 shows a jet of matter being ejected from the galaxy M87 at ultrarelativistic fields. The blue color of the jet in the visible-light image comes from *synchrotron radiation*, which is the electromagnetic radiation emitted by relativistic charged particles accelerated by a magnetic field. The jet is believed to be coming from a supermassive black hole at the center of M87. The emission of the jet in a particular direction suggests that the black hole is not spherically symmetric. It seems to have a particular axis associated with it. How can this be? Our sun's spherical symmetry is broken by the existence of externally observable features such as sunspots and the equatorial bulge, but the only information we can get about a black hole comes from its external gravitational (and possibly electromagnetic) fields. It appears that something about the spacetime metric surrounding this black hole breaks spherical symmetry, but preserves symmetry about some preferred axis. What aspect of the initial conditions in the formation of the hole could have determined such an axis? The most likely candidate is the angular momentum. We are thus led to suspect that black holes can possess angular momentum, that angular momentum preserves information about their formation, and that angular momentum is externally detectable via its effect on the spacetime metric.
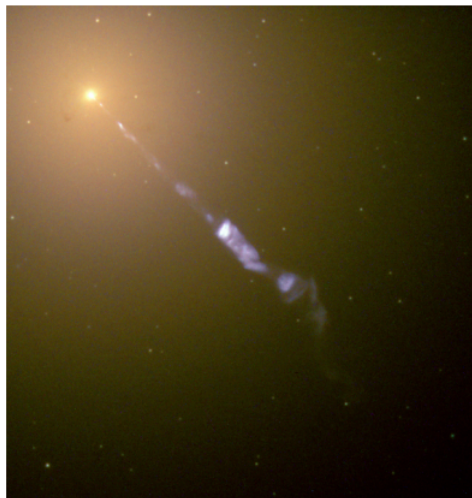
Figure 4.6.4 - A relativistic jet.

What would the form of such a metric be? Spherical coordinates in flat spacetime give a metric like this:

$$ds^2 = dt^2 - dr^2 - r^2 d\theta^2 - r^2 \sin^2 \theta d\phi^2. \qquad (4.6.1)$$

We'll see in chapter 6 that for a non-rotating black hole, the metric is of the form

$$ds^2 = (\ldots)dt^2 - (\ldots)dr^2 - r^2 d\theta^2 - r^2 \sin^2 \theta d\phi^2, \qquad (4.6.2)$$

where (. . .) represents functions of r. In fact, there is nothing special about the metric of a black hole, at least far away; the same external metric applies to *any* spherically symmetric, non-rotating body, such as the moon. Now what about the metric of a rotating body? We expect it to have the following properties:

1. It has terms that are odd under time-reversal, corresponding to reversal of the body's angular momentum.
2. Similarly, it has terms that are odd under reversal of the differential $d\phi$ of the azimuthal coordinate.
3. The metric should have axial symmetry, i.e., it should be independent of ϕ.

Restricting our attention to the equatorial plane $\theta = \frac{\pi}{2}$, the simplest modification that has these three properties is to add a term of the form

$$f(\ldots)Ld\phi dt, \qquad (4.6.3)$$

where (. . .) again gives the r-dependence and L is a constant, interpreted as the angular momentum. A detailed treatment is beyond the scope of this book, but solutions of this form to the relativistic field equations were found by New Zealand-born physicist Roy Kerr in 1963 at the University of Texas at Austin.

The astrophysical modeling of observations like figure d is complicated, but we can see in a simplified thought experiment that if we want to determine the angular momentum of a rotating body via its gravitational field, it will be difficult unless we use a measuring process that takes advantage of the asymptotic flatness of the space. For example, suppose we send two beams of light past the earth, in its equatorial plane, one on each side, and measure their deflections, e. The deflections will be different, because the sign of $d\phi$ dt will be opposite for the two beams. But the entire notion of a "deflection" only makes sense if we have an asymptotically flat background, as indicated by the dashed tangent lines. Also, if spacetime were not asymptotically flat in this example, then there might be no unambiguous way to determine whether the asymmetry was due to the earth's rotation, to some external factor, or to some kind of interaction between the earth and other bodies nearby.
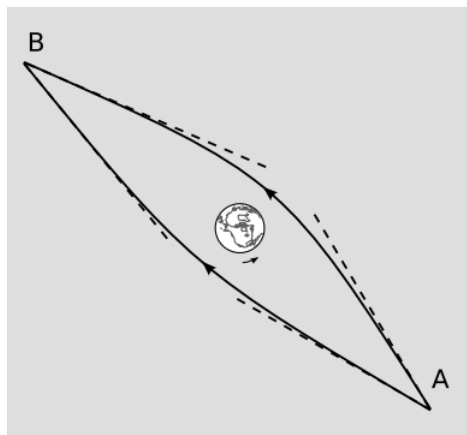
Figure 4.6.5 - Two light rays travel in the earth's equatorial plane from A to B. Due to frame-dragging, the ray moving with the earth's rotation is deflected by a greater amount than the one moving contrary to it. As a result, the figure has an asymmetric banana shape. Both the deflection and its asymmetry are greatly exaggerated.

It also turns out that a gyroscope in such a gravitational field precesses. This effect, called frame dragging, was predicted by Lense and Thirring in 1918, and was finally verified experimentally in 2008 by analysis of data from the Gravity Probe B experiment, to a precision of about 15%. The experiment was arranged so that the relatively strong geodetic effect (6.6 arc-seconds per year) and the much weaker Lense-Thirring effect (.041 arc-sec/yr) produced precessions in perpendicular directions. Again, the presence of an asymptotically flat background was involved, because the probe measured the orientations of its gyroscopes relative to the guide star IM Pegasi.
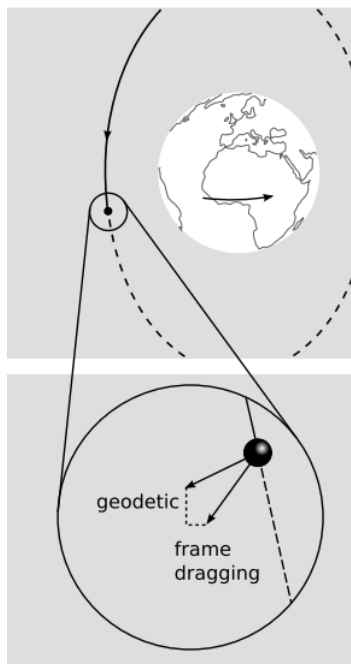
Figure 4.6.6 - Gravity Probe B verified the existence of frame-dragging. The rotational axis of the gyroscope precesses in two perpendicular planes due to the two separate effects: geodetic and frame-dragging.

This page titled 4.6: Conservation Laws is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Benjamin Crowell via source content that was edited to the style and standards of the LibreTexts platform.

4.7: Things that Aren't Quite Tensors

This section can be skipped on a first reading.

Area, Volume, and Tensor Densities

We've embarked on a program of redefining every possible physical quantity as a tensor, but so far we haven't tackled area and volume. Is there, for example, an area tensor in a locally Euclidean plane? We are encouraged to hope that there is such a thing, because in Example 3 we saw that we could cook up a measure of area with no other ingredients than the axioms of affine geometry. What kind of tensor would it be? The notions of vector and scalar from freshman mechanics are distinguished from one another by the fact that one has a direction in space and the other does not. Therefore we expect that area would be a scalar, i.e., a rank-0 tensor. But this can't be right, for the following reason. Under a rescaling of Cartesian coordinates by a factor k, area should change by a factor of k^2. But by the tensor transformation laws, a rank-0 tensor is supposed to be invariant under a change of coordinates. We therefore conclude that quantities like area and volume are not tensors.

In the language of ordinary vectors and scalars in Euclidean three-space, one way to express area and volume is by using dot and cross products. The area of the parallelogram spanned by **u** and **v** is measured by the area vector **u** × **v**, and similarly the volume of the parallelepiped formed by **u**, **v**, and **w** can be computed as the scalar triple product **u** · (**v** × **w**). Both of these quantities are defined such that interchanging two of the inputs negates the output. In differential geometry, we do have a scalar product, which is defined by contracting the indices of two vectors, as in $u^a v_a$. If we also had a a tensorial cross product, we would be able to define area and volume tensors, so we conclude that there is no tensorial cross product, i.e., an operation that would multiply two rank-1 tensors to produce a rank-1 tensor. Since one of the most important physical applications of the cross product is to calculate the angular momentum **L** = **r** × **p**, we find that angular momentum in relativity is either not a tensor or not a rank-1 tensor.

When someone tells you that it's impossible to do a seemingly straightforward thing, the typical response is to look for a way to get around the supposed limitation. In the case of a locally Euclidean plane, what is to stop us from making a small, standard square, and then sliding the square around to any desired location? If we have some figure whose area we wish to measure, we can then dissect it into squares of that size and count the number of squares.

There are two problems with this plan, neither of which is completely insurmountable. First, the area vector **u** × **v** is a vector, with its orientation specified by the direction of the normal to the surface. We need this orientation, for example, when we calculate the electric flux as $\int$ **E** · d**A**. Figure 4.6.1 shows that we cannot always define such an orientation in a consistent way. When the x − y coordinate system is slid around the Möbius strip, it ends up with the opposite orientation. In general relativity, there is not any guarantee of orientability in space — or even in time! But the vast majority of spacetimes of physical interest are in fact orientable in every desired way, and even for those that aren't, orientability still holds in any sufficiently small neighborhood.
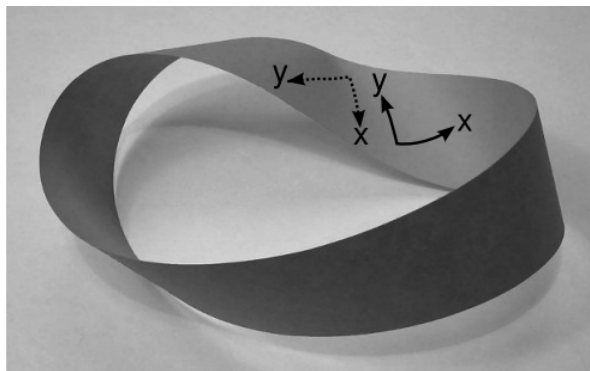
Figure 4.7.1 - A Möbius strip is not an orientable surface.

The other problem is that area has the wrong scaling properties to be a rank-0 tensor. We can get around this problem by being willing to discuss quantities that don't transform exactly like tensors. Often we only care about transformations, such as rotations and translations, that don't involve any scaling. We saw in section 2.2 that Lorentz boosts also have the special property of preserving area in a space-time plane containing the boost. We therefore define a *tensor density* as a quantity that transforms like a tensor under rotations, translations, and boosts, but that rescales and possibly flips its sign under other types of coordinate transformations. In general, the additional factor comes from the determinant d of the matrix consisting of the partial derivatives $\frac{\partial x'^\mu}{\partial x^\nu}$ (called the Jacobian matrix). This determinant is raised to a power W, known as the weight of the tensor density. Weight zero

corresponds to the case of a real tensor. The definition of the sign of W is not standardized in the literature. The convention in this book is the one used by Carroll and Weinberg, but the opposite sign is used, for example, by Misner, Thorne, and Wheeler, and in the Wikipedia article "Tensor density."

> **Example 22: Area as a tensor density**
>
> In a Euclidean plane, making our rulers shorter by a factor of k causes the area measured in the new coordinates to increase by a factor of k^2. The rescaling is represented by a matrix of partial derivatives that is simply kI, where I is the identity matrix. The determinant is k^2. Therefore area is a tensor density of weight +1.

> **Example 23: Mass density**
>
> A piece of aluminum foil as a certain number of milligrams per square centimeter. Shrinking rulers by $\frac{1}{k}$ causes this number to decrease by k^{-2}, so this mass density has W = −1.

In Weyl's apt characterization,[17] tensors represent intensities, while tensor densities measure quantity.

The Levi-Civita Symbol

Although there is no tensorial vector cross product, we can define a similar operation whose output is a tensor density. This is most easily expressed in terms of the Levi-Civita symbol ϵ. (See section 3.3 for biographical information about Levi-Civita.)

In n dimensions, the Levi-Civita symbol has n indices. It is defined so as to be totally asymmetric, in the sense that if any two of the indices are interchanged, its sign flips. This is sufficient to define the symbol completely except for an over-all scaling, which is fixed by arbitrarily taking one of the nonvanishing elements and setting it to +1. To see that this is enough to define ϵ completely, first note that it must vanish when any index is repeated. For example, in three dimensions labeled by $\kappa, \lambda,$ and $\mu, \epsilon_{\kappa\lambda\lambda}$ is unchanged under an interchange of the second and third indices, but it must also flip its sign under this operation, which means that it must be zero. If we arbitrarily fix $\epsilon_{\kappa\lambda\mu} = +1$, then interchange of the second and third indices gives $\epsilon_{\kappa\mu\lambda} = -1$, and a further interchange of the first and second yields $\epsilon_{\mu\kappa\lambda} = +1$. Any permutation of the three distinct indices can be reached from any other by a series of such pairwise swaps, and the number of swaps is uniquely odd or even.[18] In Cartesian coordinates in three dimensions, it is conventional to choose $\epsilon_{xyz} = +1$ when x, y, and z form a right-handed spatial coordinate system. In four dimensions, we take ϵ_{txyz} = +1 when t is future-timelike and (x, y, z) are right-handed.

In Euclidean three-space, in coordinates such that g = diag(1, 1, 1), the vector cross product **A** = **u** × **v**, where we have in mind the interpretation of A as area, can be expressed as $A_\mu = \epsilon_{\mu\kappa\lambda} u^\kappa v^\lambda$.

> **Note**
>
> For a proof, see the Wikipedia article "Parity of a permutation."

> **Exercise 4.7.1**
>
> Self-check: Check that this matches up with the more familiar definition of the vector cross product.

Now suppose that we want to generalize to curved spaces, where g cannot be constant. There are two ways to proceed.

Tensorial ϵ

One is to let ϵ have the values 0 and ±1 at some arbitrarily chosen point, in some arbitrarily chosen coordinate system, but to let it transform like a tensor. Then $A_\mu = \epsilon_{\mu\kappa\lambda} u^\kappa v^\lambda$ needs to be modified, since the right-hand side is a tensor, and that would make A a tensor, but if A is an area we don't want it to transform like a 1-tensor. We therefore need to revise the definition of area to be $A_\mu = g^{-1/2}\epsilon_{\mu\kappa\lambda} u^\kappa v^\lambda$, where g is the determinant of the lower-index form of the metric. The following two examples justify this procedure in a locally Euclidean three-space.

> ### Example 24: Scaling coordinates with tensorial ϵ
>
> Then scaling of coordinates by k scales all the elements of the metric by k^{-2}, g by k^{-6}, $g^{-1/2}$ by k^3, $\epsilon_{\mu\kappa\lambda}$ by k^{-3}, and $u^\kappa v^\lambda$ by k^2. The result is to scale A_μ by $k^{+3-3+2} = k^2$, which makes sense if A is an area.

> ### Example 25: oblique coordinates with tensorial ϵ
>
> In oblique coordinates (example 9), the two basis vectors have unit length but are at an angle $\phi \neq \frac{\pi}{2}$ to one another. The determinant of the metric is g = $\sin^2 \phi$, so $\sqrt{g} = \sin\phi$, which is exactly the correction factor needed in order to get the right area when u and v are the two basis vectors.

This procedure works more generally, the sole modification being that in a space such as a locally Lorentzian one where g < 0 we need to use $\sqrt{-g}$ as the correction factor rather than $\sqrt{g}$.

Tensor-density ϵ

The other option is to let ϵ have the same 0 and ±1 values at all points. Then is clearly not a tensor, because it doesn't scale by a factor of k^n when the coordinates are scaled by k; ϵ is a tensor density with weight −1 for the upper-index version and +1 for the lower-index one. The relation $A_\mu = \epsilon_{\mu\kappa\lambda} u^\kappa v^\lambda$ gives an area that is a tensor density, not a tensor, because A is not written in terms of purely tensorial quantities. Scaling the coordinates by k leaves $\epsilon_{\mu\kappa\lambda}$ unchanged, scales up $u^\kappa v^\lambda$ by k^2, and scales up the area by k^2, as expected.

Unfortunately, there is no consistency in the literature as to whether ϵ should be a tensor or a tensor density. Some authors define both a tensor and a nontensor version, with notations like ϵ and $\tilde{\epsilon}$, or[19] ϵ_{0123} and [0123]. Others avoid writing the letter ϵ completely.[20] The tensor-density version is convenient because we always know that its value is 0 or ±1. The tensor version has the advantage that it transforms as a tensor.

Spacetime Volume

We saw in section 2.2 that area in the 1 + 1-dimensional plane of flat spacetime is preserved by a Lorentz boost. This makes sense because when we express the area spanned by a parallelogram with edges **p** and **q** as $\epsilon^{ab} p_a s_b$, all the indices have been contracted, leaving a rank-0 tensor density. In 3 + 1 dimensions, we have the spacetime volume V = $\epsilon^{abcd} p_a q_b r_c s_d$ spanned by the paralellepiped with edges **p**, **q**, **r**, and **s**. A typical situation in which this volume is nonzero would be that in which one of the vectors is timelike and the other three spacelike. Let the timelike one be **p**. Assume |p| = 1, since an example with |p| ≠ 1 can be reduced to this by scaling. Then **p** can be interpreted as the velocity vector of some observer, and V as the spatial volume that the observer says is spanned by the 3-paralellepiped with edges **q**, **r**, and **s**.

Angular Momentum

As discussed above, angular momentum cannot be a rank-1 tensor. One approach is to define a rank-2 angular momentum tensor $L^{ab} = r^a p^b - r^b p^a$.

In a frame whose origin is instantaneously moving along with a certain system's center of mass at a certain time, the time-space components of L vanish, and the components L^{yz}, L^{zx}, and L^{xy} coincide in the nonrelativistic limit with the x, y, and z components of the Newtonian angular momentum vector. We can also define a three-dimensional object $L^a = \epsilon_{abc} L^{bc}$ (with three-dimensional tensor-density ϵ in the spatial dimensions) that doesn't transform like a tensor.

References

[17] Hermann Weyl, "Space-Time-Matter," 1922, p. 109, available online at archive.org/details/spacetimematter00weyluoft.

[19] Misner, Thorne, and Wheeler

[20] Hawking and Ellis

4.E: Tensors (Exercises)

1. Describe the four-velocity of a photon.
2. The Large Hadron Collider is designed to accelerate protons to energies of 7 TeV. Find $1 - v$ for such a proton.
3. Prove that an electron in a vacuum cannot absorb a photon. (This is the reason that the ability of materials to absorb gammarays is strongly dependent on atomic number Z. The case of Z = 0 corresponds to the vacuum.)
4. (a) For an object moving in a circle at constant speed, the dot product of the classical three-vectors **v** and **a** is zero. Give an interpretation in terms of the work-kinetic energy theorem. (b) In the case of relativistic four-vectors, $v^i a_i = 0$ for *any* world-line. Give a similar interpretation. *Hint*: find the rate of change of the four-velocity's squared magnitude.
5. Starting from coordinates (t, x) having a Lorentzian metric g, transform the metric tensor into reflected coordinates (t', x') = (t, −x), and verify that g' is the same as g.
6. Starting from coordinates (t, x) having a Lorentzian metric g, transform the metric tensor into Lorentz-boosted coordinates (t', x'), and verify that g' is the same as g.
7. Verify the transformation of the metric given in example 19.
8. A skeptic claims that the Hafele-Keating experiment can only be explained correctly by relativity in a frame in which the earth's axis is at rest. Prove mathematically that this is incorrect. Does it matter whether the frame is inertial?
9. Assume the metric g = diag(+1, +1, +1). Which of the following correctly expresses the noncommutative property of ordinary matrix multiplication?
10. Example 10 introduced the Dirac sea, whose existence is implied by the two roots of the relativistic relation $E = \pm\sqrt{p^2 + m^2}$. Prove that a Lorentz boost will never transform a positive-energy state into a negative-energy state.
11. In section 4.2, we found the relativistic Doppler shift in 1+1 dimensions. Extend this to 3+1 dimensions, and check your result against the one given by Einstein in Appendix A.
12. Estimate the energy contained in the electric field of an electron, if the electron's radius is r. Classically (i.e., assuming relativity but no quantum mechanics), this energy contributes to the electron's rest mass, so it must be less than the rest mass. Estimate the resulting lower limit on r, which is known as the classical electron radius.
13. For gamma-rays in the MeV range, the most frequent mode of interaction with matter is Compton scattering, in which the photon is scattered by an electron without being absorbed. Only part of the gamma's energy is deposited, and the amount is related to the angle of scattering. Use conservation of four-momentum to show that in the case of scattering at 180 degrees, the scattered photon has energy E' = E/(1+2E/m), where m is the mass of the electron.
14. Derive the equation $T = \sqrt{\frac{3\pi}{G\rho}}$ given in section 4.4 for the period of a rotating, spherical object that results in zero apparent gravity at its surface.
15. Section 4.4 presented an estimate of the upper limit on the mass of a white dwarf. Check the self-consistency of the solution in the following respects: (1) Why is it valid to ignore the contribution of the nuclei to the degeneracy pressure? (2) Although the electrons are ultrarelativistic, spacetime is approximated as being flat. As suggested in example 14, a reasonable order-ofmagnitude check on this result is that we should have $\frac{M}{r} << \frac{c^2}{G}$.
16. The laws of physics in our universe imply that for bodies with a certain range of masses, a neutron star is the unique equilibrium state. Suppose we knew of the existence of neutron stars, but didn't know the mass of the neutron. Infer upper and lower bounds on the mass of the neutron.
17. Example 20 briefly introduced the electromagnetic potential four-vector F_{ij}, and this implicitly defines the transformation properties of the electric and magnetic fields under a Lorentz boost **v**. To lowest order in **v**, this transformation is given by

$$\mathbf{E}' \approx \mathbf{E} + \mathbf{v} \times \mathbf{B} \qquad and$$
$$\mathbf{B}' \approx \mathbf{B} - \mathbf{v} \times \mathbf{E}.$$

I'm not a historian of science, but apparently ca. 1905 people like Hertz believed that these were the *exact* transformations of the field.[21] Show that this can't be the case, because performing two such transformations in a row does not in general result in a transformation of the same form.
18. We know of massive particles, whose velocity vectors always lie inside the future light cone, and massless particles, whose velocities lie on it. In principle, we could have a third class of particles, called tachyons, with spacelike velocity vectors. Tachyons would have m^2 < 0, i.e., their masses would have to be imaginary. Show that it is possible to pick momentum four-vectors $\mathbf{p}_1$ and $\mathbf{p}_2$ for a pair of tachyons such that $\mathbf{p}_1 + \mathbf{p}_2 = 0$. This implies that the vacuum would be unstable with respect to spontaneous creation of tachyon-antitachyon pairs.

References

21 Montigny and Rousseaux, arxiv.org/abs/physics/0512200.

5: Curvature

General relativity describes gravitation as a curvature of spacetime, with matter acting as the source of the curvature in the same way that electric charge acts as the source of electric fields. Our goal is to arrive at Einstein's field equations, which relate the local intrinsic curvature to the locally ambient matter in the same way that Gauss's law relates the local divergence of the electric field to the charge density. The locality of the equations is necessary because relativity has no action at a distance; cause and effect propagate at a maximum velocity of c.

Contributors and Attributions

- Benjamin Crowell (Fullerton College). General Relativity is copyrighted with a CC-BY-SA license.
- Thumbnail: The overall geometry of the universe is determined by whether the Omega cosmological parameter is less than, equal to or greater than 1. Shown from top to bottom are a closed universe with positive curvature, a hyperbolic universe with negative curvature and a flat universe with zero curvature. (Public Domain; NASA).

5.1: Introduction to Curvature

General relativity describes gravitation as a curvature of spacetime, with matter acting as the source of the curvature in the same way that electric charge acts as the source of electric fields. Our goal is to arrive at Einstein's field equations, which relate the local intrinsic curvature to the locally ambient matter in the same way that Gauss's law relates the local divergence of the electric field to the charge density. The locality of the equations is necessary because relativity has no action at a distance; cause and effect propagate at a maximum velocity of $c(=1)$.

The hard part is arriving at the right way of defining curvature. We've already seen that it can be tricky to distinguish *intrinsic curvature*, which is real, from *extrinsic curvature*, which can never produce observable effects. E.g., Example 5 showed that spheres have intrinsic curvature, while cylinders do not. The manifestly intrinsic tensor notation protects us from being misled in this respect. If we can formulate a definition of curvature expressed using only tensors that are expressed without reference to any preordained coordinate system, then we know it is physically observable, and not just a superficial feature of a particular model.
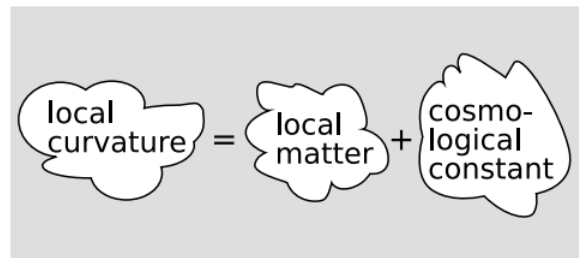
Figure 5.1.1 - The expected structure of the field equations in general relativity.

As an example, drop two rocks side by side, Figure 5.0.2. Their trajectories are vertical, but on a (t, x) coordinate plot rendered in the Earth's frame of reference, they appear as parallel parabolas. The curvature of these parabolas is extrinsic. The Earth-fixed frame of reference is defined by an observer who is subject to non-gravitational forces, and is therefore not a valid Lorentz frame. In a free-falling Lorentz frame (t', x'), the two rocks are either motionless or moving at constant velocity in straight lines. We can therefore see that the curvature of world-lines in a particular coordinate system is not an intrinsic measure of curvature; it can arise simply from the choice of the coordinate system. What would indicate intrinsic curvature would be, for example, if geodesics that were initially parallel were to converge or diverge.
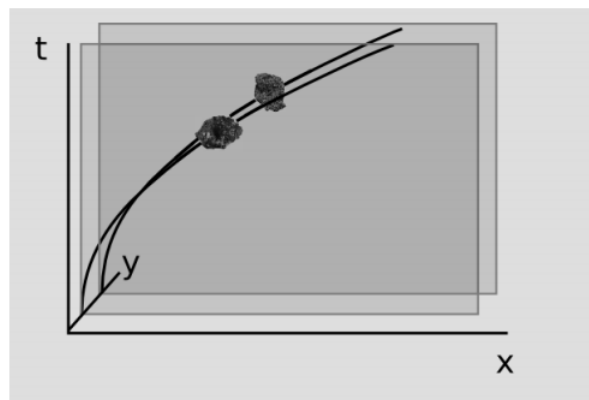
Figure 5.1.2 - Two rocks are dropped side by side. The curvatures of their world-lines are not intrinsic. In a free-falling frame, both would appear straight. If initially parallel world-lines became non-parallel, that would be evidence of intrinsic curvature.

Nor is the metric a measure of intrinsic curvature. In example 19, we found the metric for an accelerated observer to be

$$g'_{t't'} = (1 + ax')^2 \qquad g_{x'x'} = -1, \tag{5.1.1}$$

where the primes indicate the accelerated observer's frame. The fact that the timelike element is not equal to −1 is not an indication of intrinsic curvature. It arises only from the choice of the coordinates (t', x') defined by a frame tied to the accelerating rocket ship.

The fact that the above metric has nonvanishing derivatives, unlike a constant Lorentz metric, does indicate the presence of a gravitational field. However, a gravitational field is not the same thing as intrinsic curvature. The gravitational field seen by an observer aboard the ship is, by the equivalence principle, indistinguishable from an acceleration, and indeed the Lorentzian

observer in the earth's frame does describe it as arising from the ship's acceleration, not from a gravitational field permeating all of space. Both observers must agree that "I got plenty of nothin' " — that the region of the universe to which they have access lacks any stars, neutrinos, or clouds of dust. The observer aboard the ship must describe the gravitational field he detects as arising from some source very far away, perhaps a hypothetical vast sheet of lead lying billions of light-years aft of the ship's deckplates. Such a hypothesis is fine, but it is unrelated to the structure of our hoped-for field equation, which is to be *local* in nature.

Not only does the metric tensor not represent the gravitational field, but no tensor can represent it. By the equivalence principle, any gravitational field seen by observer A can be eliminated by switching to the frame of a free-falling observer B who is instantaneously at rest with respect to A at a certain time. The structure of the tensor transformation law guarantees that A and B will agree on whether a given tensor is zero at the point in spacetime where they pass by one another. Since they agree on all tensors, and disagree on the gravitational field, the gravitational field cannot be a tensor.

We therefore conclude that a nonzero intrinsic curvature of the type that is to be included in the Einstein field equations is not encoded in any simple way in the metric or its first derivatives. Since neither the metric nor its first derivatives indicate curvature, we can reasonably conjecture that the curvature might be encoded in its second derivatives.

5.2: Tidal Curvature Versus Curvature Caused by Local Sources

A further complication is the need to distinguish tidal curvature from curvature caused by local sources. Figure $5.2.1$ shows Comet Shoemaker-Levy, broken up into a string of fragments by Jupiter's tidal forces shortly before its spectacular impact with the planet in 1994. Immediately after each fracture, the newly separated chunks had almost zero velocity relative to one another, so once the comet finished breaking up, the fragments' world-lines were a sheaf of nearly parallel lines separated by spatial distances of only 1 km. These initially parallel geodesics then diverged, eventually fanning out to span millions of kilometers.

Figure $5.2.1$: Tidal forces disrupt comet Shoemaker-Levy.

If initially parallel lines lose their parallelism, that is clearly an indication of intrinsic curvature. We call it a measure of *sectional curvature*, because the loss of parallelism occurs within a particular plane, in this case the (t, x) plane represented by Figure $5.2.2$.
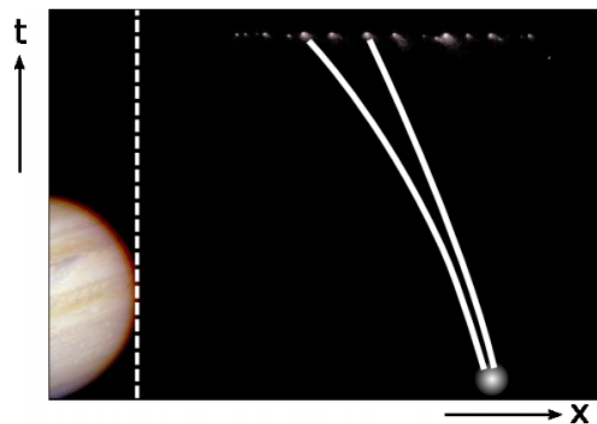
Figure $5.2.2$: Tidal forces cause the initially parallel world-lines of the fragments to diverge. The spacetime occupied by the comet has intrinsic curvature, but it is not caused by any local mass; it is caused by the distant mass of Jupiter.

But this curvature was not caused by a local source lurking in among the fragments. It was caused by a distant source: Jupiter. We therefore see that the mere presence of sectional curvature is not enough to demonstrate the existence of local sources. Even the sign of the sectional curvature is not a reliable indication. Although this example showed a divergence of initially parallel geodesics, referred to as a negative curvature, it is also possible for tidal forces exerted by distant masses to create positive curvature. For example, the ocean tides on earth oscillate both above and below mean sea level, Figure Figure $5.2.3$.
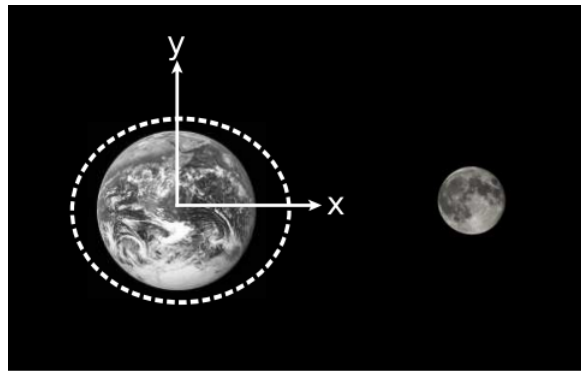
Figure $5.2.3$: The moon's gravitational field causes the Earth's oceans to be distorted into an ellipsoid. The sign of the sectional curvature is negative in the x − t plane, but positive in the y − t plane.

As an example that really would indicate the presence of a local source, we could release a cloud of test masses at rest in a spherical shell around the earth, and allow them to drop, Figure $5.2.4$. We would then have positive and equal sectional curvature in the $t - x$, $t - y$, and $t - z$ planes. Such an observation cannot be due to a distant mass. It demonstrates an over-all contraction of the volume of an initially parallel sheaf of geodesics, which can never be induced by tidal forces. The earth's oceans, for example, do not change their total volume due to the tides, and this would be true even if the oceans were a gas rather than an incompressible fluid. It is a unique property of $\frac{1}{r^2}$ forces such as gravity that they conserve volume in this way; this is essentially a restatement of Gauss's law in a vacuum.
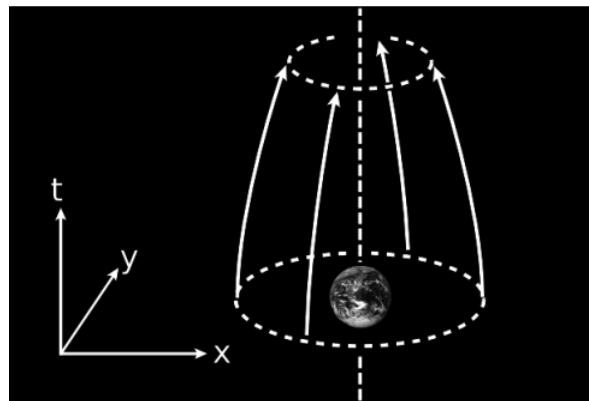
Figure $5.2.4$: A cloud of test masses is released at rest in a spherical shell around the earth, shown here as a circle because the z axis is omitted. The volume of the shell contracts over time, which demonstrates that the local curvature of spacetime is generated by a local source — the earth — rather than some distant one.

5.3: The Stress-energy Tensor

In general, the curvature of spacetime will contain contributions from both tidal forces and local sources, superimposed on one another. To develop the right formulation for the Einstein field equations, we need to eliminate the tidal part. Roughly speaking, we will do this by averaging the sectional curvature over all three of the planes $t - x$, $t - y$, and $t - z$, giving a measure of curvature called the *Ricci curvature*. The "roughly speaking" is because such a prescription would treat the time and space coordinates in an extremely asymmetric manner, which would violate local Lorentz invariance.

To get an idea of how this would work, let's compare with the Newtonian case, where there really is an asymmetry between the treatment of time and space. In the Cartan curved-spacetime theory of Newtonian gravity (Chapter 2), the field equation has a kind of scalar Ricci curvature on one side, and on the other side is the density of mass, which is also a scalar. In relativity, however, the source term in the equation clearly cannot be the scalar mass density. We know that mass and energy are equivalent in relativity, so for example the curvature of spacetime around the earth depends not just on the mass of its atoms but also on all the other forms of energy it contains, such as thermal energy and electromagnetic and nuclear binding energy. Can the source term in the Einstein field equations therefore be the mass-energy E? No, because E is merely the timelike component of a particle's momentum four-vector. To single it out would violate Lorentz invariance just as much as an asymmetric treatment of time and space in constructing a Ricci measure of curvature. To get a properly Lorentz invariant theory, we need to find a way to formulate everything in terms of tensor equations that make no explicit reference to coordinates. The proper generalization of the Newtonian mass density in relativity is the stress-energy tensor T^{ij}, whose 16 elements measure the local density of mass-energy and momentum, and also the rate of transport of these quantities in various directions. If we happen to be able to find a frame of reference in which the local matter is all at rest, then T^{tt} represents the mass density. The reason for the word "stress" in the name is that, for example, the flux of x-momentum in the x direction is a measure of pressure.

For the purposes of the present discussion, it's not necessary to introduce the explicit definition of T; the point is merely that we should expect the Einstein field equations to be tensor equations, which tells us that the definition of curvature we're seeking clearly has to be a rank-2 tensor, not a scalar. The implications in four-dimensional spacetime are fairly complex. We'll end up with a rank-4 tensor that measures the sectional curvature, and a rank-2 Ricci tensor derived from it that averages away the tidal effects. The Einstein field equations then relate the Ricci tensor to the energy-momentum tensor in a certain way. The stress-energy tensor is discussed further in Section 8.1.

5.4: Curvature in Two Spacelike Dimensions

Since the curvature tensors in 3+1 dimensions are complicated, let's start by considering lower dimensions. In one dimension, Figure 5.3.1, there is no such thing as intrinsic curvature. This is because curvature describes the failure of parallelism to behave as in E5, but there is no notion of parallelism in one dimension.

Figure 5.4.1 - This curve has no intrinsic curvature.

The lowest interesting dimension is therefore two, and this case was studied by Carl Friedrich Gauss in the early nineteenth century. Gauss ran a geodesic survey of the state of Hanover, inventing an optical surveying instrument called a heliotrope that in effect was used to cover the Earth's surface with a triangular mesh of light rays. If one of the mesh points lies, for example, at the peak of a mountain, then the sum $\Sigma\theta$ of the angles of the vertices meeting at that point will be less than 2π, in contradiction to Euclid. Although the light rays do travel through the air above the dirt, we can think of them as approximations to geodesics painted directly on the dirt, which would be intrinsic rather than extrinsic. The angular defect around a vertex now vanishes, because the space is locally Euclidean, but we now pick up a different kind of *angular defect*, which is that the interior angles of a triangle no longer add up to the Euclidean value of π.

Figure 5.4.2 - A surveyor on a mountaintop uses a heliotrope.
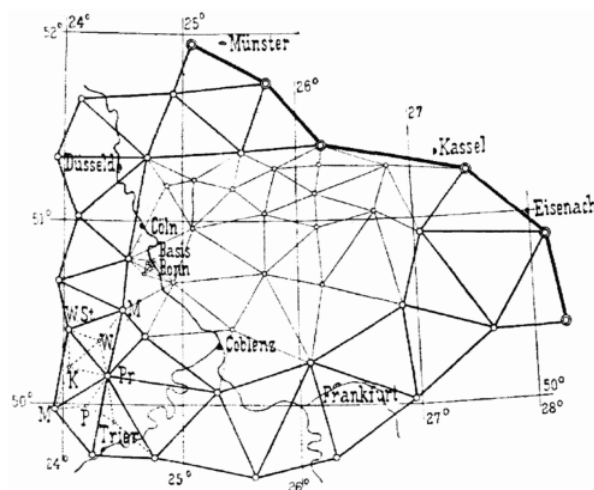
Figure 5.4.3 - A map of a triangulation survey such as the one Gauss carried out. By measuring the interior angles of the triangles, one can determine not just the two-dimensional projection of the grid but its complete three-dimensional form, including both the curvature of the earth (note the curvature of the lines of latitude) and the height of features above and below sea level.

Example 1: A polygonal survey of a soccer ball

Figure 5.3.3 applies similar ideas to a soccer ball, the only difference being the use of pentagons and hexagons rather than triangles.
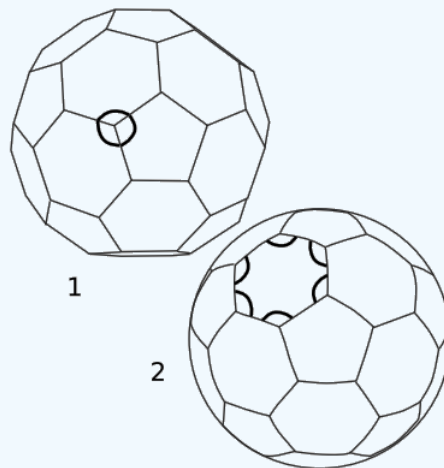
Figure 5.4.4

In Figure 5.3.4 (1), the survey is extrinsic, because the lines pass below the surface of the sphere. The curvature is detectable because the angles at each vertex add up to 120 + 120 + 110 = 350 degrees, giving an angular defect of 10 degrees.

In Figure 5.3.4 (2), the lines have been projected to form arcs of great circles on the surface of the sphere. Because the space is locally Euclidean, the sum of the angles at a vertex has its Euclidean value of 360 degrees. The curvature can be detected, however, because the sum of the internal angles of a polygon is greater than the Euclidean value. For example, each spherical hexagon gives a sum of 6 × 124.31 degrees, rather than the Euclidean 6 × 120. The angular defect of 6 × 4.31 degrees is an intrinsic measure of curvature.

Example 2: Angular defect on the earth's surface

Divide the Earth's northern hemisphere into four octants, with their boundaries running through the north pole. These octants have sides that are geodesics, so they are equilateral triangles. Assuming Euclidean geometry, the interior angles of an equilateral triangle are each equal to 60 degrees, and, as with any triangle, they add up to 180 degrees. The octant-triangle in Figure 5.3.5 has angles that are each 90 degrees, and the sum is 270. This shows that the Earth's surface has intrinsic curvature.

Figure 5.4.5

This example suggests another way of measuring intrinsic curvature, in terms of the ratio $\frac{C}{r}$ of the circumference of a circle to its radius. In Euclidean geometry, this ratio equals 2π. Let ρ be the radius of the Earth, and consider the equator to be a circle centered on the north pole, so that its radius is the length of one of the sides of the triangle in Figure 5.3.5, $r = \left(\frac{\pi}{2}\right)\rho$. (Don't

confuse r, which is intrinsic, with ρ, the radius of the sphere, which is extrinsic and not equal to r.) Then the ratio $\frac{C}{r}$ is equal to 4, which is smaller than the Euclidean value of 2π.

Let

$$\epsilon = \sum \theta - \pi \qquad (5.4.1)$$

be the *angular defect of a triangle*, and for concreteness let the triangle be in a space with an elliptic geometry, so that it has constant curvature and can be modeled as a sphere of radius ρ, with antipodal points identified.

> **Exercise 5.4.1**
>
> In elliptic geometry, what is the minimum possible value of the quantity $\frac{C}{r}$ discussed in example 2? How does this differ from the case of spherical geometry?

We want a measure of curvature that is local, but if our space is locally flat, we must have $\epsilon \to 0$ as the size of the triangles approaches zero. This is why Euclidean geometry is a good approximation for small-scale maps of the earth. The discrete nature of the triangular mesh is just an artifact of the definition, so we want a measure of curvature that, unlike ϵ, approaches some finite limit as the scale of the triangles approaches zero. Should we expect this scaling to go as $\epsilon \propto \rho$? ρ^2? Let's determine the scaling. First we prove a classic lemma by Gauss, concerning a slightly different version of the angular defect, for a single triangle.

> **Theorem**
>
> In elliptic geometry, the angular defect $\epsilon = \alpha + \beta + \gamma - \pi$ of a triangle is proportional to its area A.

> **proof**
>
> By axiom E2, extend each side of the triangle to form a line, Figure 5.3.6 (1). Each pair of lines crosses at only one point (E1) and divides the plane into two lunes with their four vertices touching at this point, Figure 5.3.6 (2). Of the six lunes, we focus on the three shaded ones, which overlap the triangle. In each of these, the two interior angles at the vertex are the same (Euclid I.15). The area of a lune is proportional to its interior angle, as follows from dissection into narrower lunes; since a lune with an interior angle of π covers the entire area P of the plane, the constant of proportionality is $\frac{P}{\pi}$. The sum of the areas of the three lunes is $\left(\frac{P}{\pi}\right)(\alpha + \beta + \gamma)$, but these three areas also cover the entire plane, overlapping three times on the given triangle, and therefore their sum also equals $P + 2A$. Equating the two expressions leads to the desired result.
>
>
>
> Figure 5.4.6 - Proof that the angular defect of a triangle in elliptic geometry is proportional to its area. Each white circle represents the entire elliptic plane. The dashed line at the edge is not really a boundary; lines that go off the edge simply wrap back around. In the spherical model, the white circle corresponds to one hemisphere, which is identified with the opposite hemisphere.
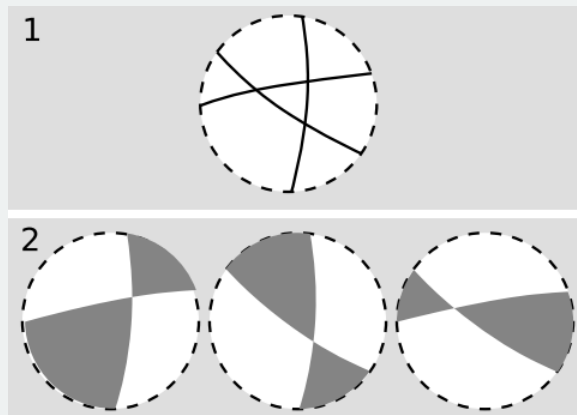
This calculation was purely intrinsic, because it made no use of any model or coordinates. We can therefore construct a measure of curvature that we can be assured is intrinsic, $K = \frac{\epsilon}{A}$. This is called the *Gaussian curvature*, and in elliptic geometry it is constant rather than varying from point to point. In the model on a sphere of radius ρ, we have

$$K = \frac{1}{\rho^2}. \tag{5.4.2}$$

Exercise $5.4.2$

Self-check: Verify the equation $K = \frac{1}{\rho^2}$ by considering a triangle covering one octant of the sphere, as in example 2.

Gaussian Normal Coordinates

It is useful to introduce *normal* or *Gaussian normal coordinates*, defined as follows. Through point O, construct perpendicular geodesics, and define affine coordinates x and y along these. For any point P off the axis, define coordinates by constructing the lines through P that cross the axes perpendicularly. For P in a sufficiently small neighborhood of O, these lines exist and are uniquely determined. Gaussian polar coordinates can be defined in a similar way.
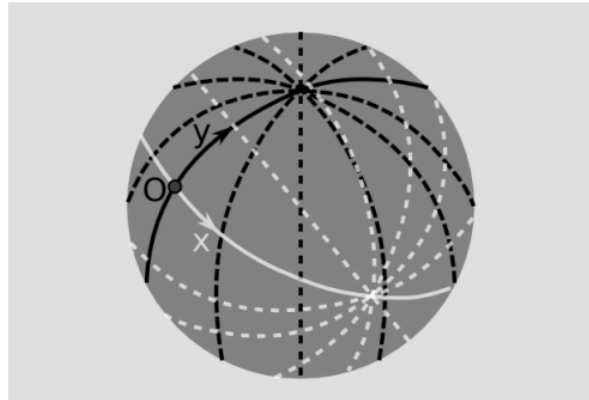
Figure 5.4.7 - Gaussian normal coordinates on a sphere.

Here are two useful interpretations of K.

1. The Gaussian curvature measures the failure of parallelism in the following sense. Let line ℓ be constructed so that it crosses the normal y axis at (0, dy) at an angle that differs from perpendicular by the infinitesimal amount dα (Figure 5.3.8). Construct the line x' = dx, and let dα' be the angle its perpendicular forms with ℓ. Then[1] the Gaussian curvature at O is

$$K = \frac{d^2\alpha}{dxdy}, \tag{5.4.3}$$

where $d^2\alpha = d\alpha' - d\alpha$.
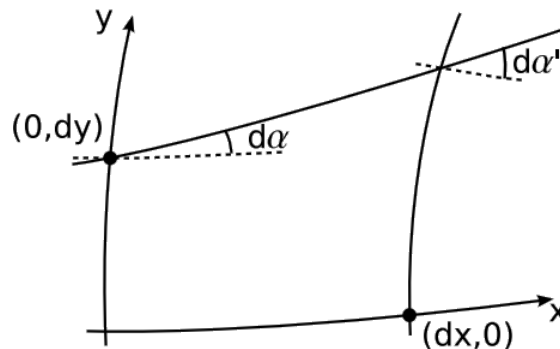
Figure 5.4.8 - 1. Gaussian curvature can be interpreted as the failure of parallelism represented by $\frac{d^2\alpha}{dxdy}$.

Proof

Since any two lines cross in elliptic geometry, ℓ crosses the x axis. The corollary then follows by application of the definition of the Gaussian curvature to the right triangles formed by ℓ, the x axis, and the lines at x = 0 and x = dx, so that

$K = \frac{d\epsilon}{dA} = \frac{d^2\alpha}{dxdy}$, where third powers of infinitesimals have been discarded.

2. From a point P, emit a fan of rays at angles filling a certain range θ of angles in Gaussian polar coordinates (Figure 5.3.9). Let the arc length of this fan at r be L, which may not be equal to its Euclidean value $L_E = r\theta$. Then[2]

$$K = -3\frac{d^2}{dr^2}\left(\frac{L}{L_E}\right). \tag{5.4.4}$$
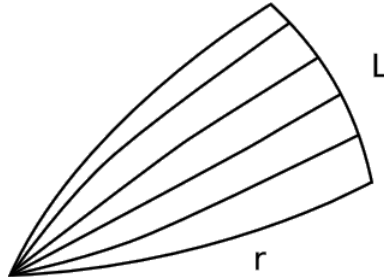
Figure 5.4.9 - *2. Gaussian curvature as L ≠ rθ.*

> **Note**
>
> In the spherical model, L = $\rho\theta \sin u$, where u is the angle subtended at the center of the sphere by an arc of length r. We then have $\frac{L}{L_E} = \frac{\sin u}{u}$, whose second derivative with respect to u is $-\frac{1}{3}$. Since r = ρu, the second derivative of the same quantity with respect to r equals $-\frac{1}{3\rho^2} = -\frac{K}{3}$.

Let's now generalize beyond elliptic geometry. Consider a space modeled by a surface embedded in three dimensions, with geodesics defined as curves of extremal length, i.e., the curves made by a piece of string stretched taut across the surface. At a particular point P, we can always pick a coordinate system (x, y, z) such that the surface $z = \frac{1}{2}k_1 x^2 + \frac{1}{2}k_2 y^2$ locally approximates the surface to the level of precision needed in order to discuss curvature. The surface is either paraboloidal or hyperboloidal (a saddle), depending on the signs of k_1 and k_2. We might naively think that k_1 and k_2 could be independently determined by intrinsic measurements, but as we've seen in example 5, a cylinder is locally indistinguishable from a Euclidean plane, so if one k is zero, the other k clearly cannot be determined. In fact all that can be measured is the Gaussian curvature, which equals the product $k_1 k_2$. To see why this should be true, first consider that any measure of curvature has units of inverse distance squared, and the k's have units of inverse distance. The only possible intrinsic measures of curvature based on the k's are therefore $k^2_1 + k^2_2$ and $k_1 k_2$. (We can't have, for example, just k^2_1, because that would change under an extrinsic rotation about the z axis.) Only $k_1 k_2$ vanishes on a cylinder, so it is the only possible intrinsic curvature.

Example 3: Eating pizza

When people eat pizza by folding the slice lengthwise, they are taking advantage of the intrinsic nature of the Gaussian curvature. Once k_1 is fixed to a nonzero value, k_2 can't change without varying K, so the slice can't droop.

Example 4: Elliptic and hyperbolic geometry

We've seen that figures behaving according to the axioms of elliptic geometry can be modeled on part of a sphere, which is a surface of constant K > 0. The model can be made into global one satisfying all the axioms if the appropriate topological properties are ensured by identifying antipodal points. A paraboloidal surface $z = k_1 x^2 + k_2 y^2$ can be a good local approximation to a sphere, but for points far from its apex, K varies significantly. Elliptic geometry has no parallels; all lines meet if extended far enough.

A space of constant negative curvature has a geometry called hyperbolic, and is of some interest because it appears to be the one that describes the spatial dimensions of our universe on a cosmological scale. A hyperboloidal surface works locally as a model, but its curvature is only approximately constant; the surface of constant curvature is a horn-shaped one created by revolving a mountain-shaped curve called a tractrix about its axis. The tractrix of revolution is not as satisfactory a model as the sphere is for elliptic geometry, because lines are cut off at the cusp of the horn. Hyperbolic geometry is richer in parallels

than Euclidean geometry; given a line ℓ and a point P not on ℓ, there are infinitely many lines through P that do not pass through ℓ.
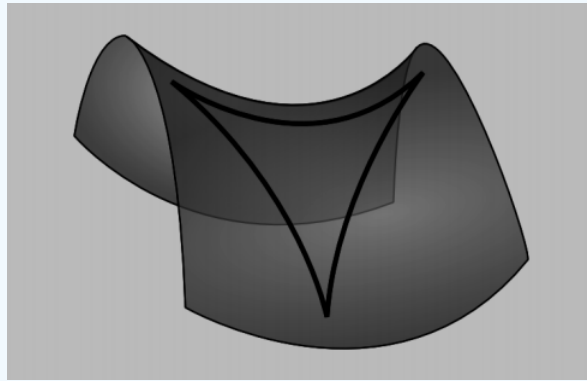
Figure 5.4.10 - A triangle in a space with negative curvature has angles that add to less than π.

Example 5: A flea on a football

We might imagine that a flea on the surface of an American football could determine by intrinsic, local measurements which direction to go in order to get to the nearest tip. This is impossible, because the flea would have to determine a vector, and curvature cannot be a vector, since $z = \frac{1}{2}k_1 x^2 + \frac{1}{2}k_2 y^2$ is invariant under the parity inversion x → −x, y → −y. For similar reasons, a measure of curvature can never have odd rank.

Without violating reflection symmetry, it is still conceivable that the flea could determine the orientation of the tip-to-tip line running through his position. Surprisingly, even this is impossible. The flea can only measure the single number K, which carries no information about directions in space.
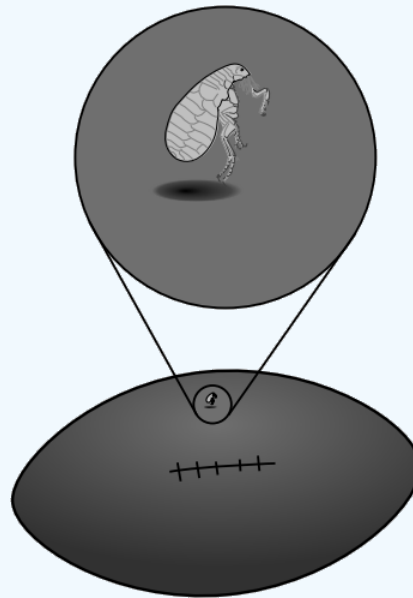
Figure 5.4.11 - A flea on the football cannot orient himself by intrinsic, local measurements.

Example 6: The lightning rod

Suppose you have a pear-shaped conductor like the one in Figure 5.3.12 (1). Since the pear is a conductor, there are free charges everywhere inside it. Panels 1 and 2 of the figure show a computer simulation with 100 identical electric charges. In 1, the charges are released at random positions inside the pear. Repulsion causes them all to fly outward onto the surface and then settle down into an orderly but nonuniform pattern.
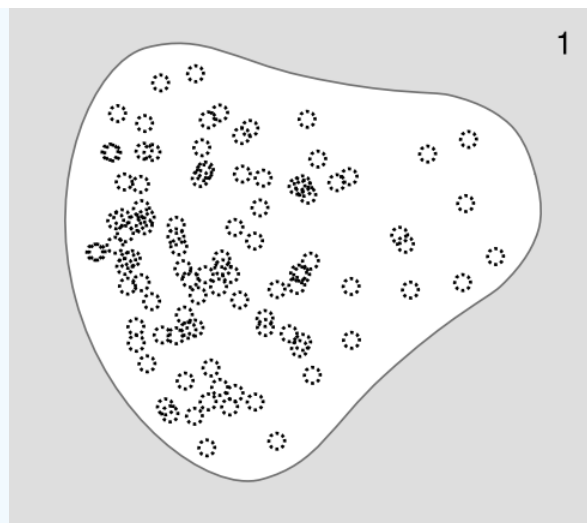
Figure_5.3.12b.png" />
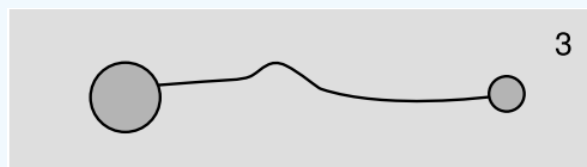
Figure 5.4.12 - In 1 and 2, charges that are visible on the front surface of the conductor are shown as solid dots; the others would have to be seen through the conductor, which we imagine is semi-transparent.

We might not have been able to guess the pattern in advance, but we can verify that some of its features make sense. For example, charge A has more neighbors on the right than on the left, which would tend to make it accelerate off to the left. But when we look at the picture as a whole, it appears reasonable that this is prevented by the larger number of more distant charges on its left than on its right.

There also seems to be a pattern to the nonuniformity: the charges collect more densely in areas like B, where the Gaussian curvature is large, and less densely in areas like C, where K is nearly zero (slightly negative).

To understand the reason for this pattern, consider Figure 5.3.12 (3). It's straightforward to show that the density of charge σ on each sphere is inversely proportional to its radius, or proportional to $K^{1/2}$. Lord Kelvin proved that on a conducting ellipsoid, the density of charge is proportional to the distance from the center to the tangent plane, which is equivalent[3] to $\sigma \propto K^{1/4}$; this result looks similar except for the different exponent. McAllister showed in 1990[4] that this $K^{1/4}$ behavior applies to a certain class of examples, but it clearly can't apply in all cases, since, for example, K could be negative, or we could have a deep concavity, which would form a Faraday cage. Problem 1 discusses the case of a knife-edge.

Similar reasoning shows why Benjamin Franklin used a sharp tip when he invented the lightning rod. The charged stormclouds induce positive and negative charges to move to opposite ends of the rod. At the pointed upper end of the rod, the charge tends to concentrate at the point, and this charge attracts the lightning. The same effect can sometimes be seen when a scrap of aluminum foil is inadvertently put in a microwave oven. Modern experiments[5] show that although a sharp tip is best at starting a spark, a more moderate curve, like the right-hand tip of the pear in this example, is better at successfully sustaining the spark for long enough to connect a discharge to the clouds.

---

[3] http://math.stackexchange.com/questi...f-the-distance

4 I W McAllister 1990 J. Phys. D: Appl. Phys. 23 359

[5] Moore et al., Journal of Applied Meteorology 39 (1999) 593

5.5: Curvature Tensors

The example of the flea suggests that if we want to express curvature as a tensor, it should have even rank. Also, in a coordinate system in which the coordinates have units of distance (they are not angles, for instance, as in spherical coordinates), we expect that the units of curvature will always be inverse distance squared. Another way of putting this is that if we start with normal coordinates and then rescale all the coordinates by a factor of μ, a curvature tensor should scale down by μ^{-2}. (See section 5.11, for more on this topic.)

Combining these two facts, we find that a curvature tensor should have one of the forms R_{ab} $R^a{}_{bcd}$, . . . , i.e., the number of lower indices should be two greater than the number of upper indices. The following definition has this property, and is equivalent to the earlier definitions of the Gaussian curvature that were not written in tensor notation.

Definition of the Riemann curvature tensor: Let dp^c and dq^d be two infinitesimal vectors, and use them to form a quadrilateral that is a good approximation to a parallelogram.[6] Parallel-transport vector v^b all the way around the parallelogram. When it comes back to its starting place, it has a new value $v^b \rightarrow v^b + dv^b$. Then the *Riemann curvature tensor* is defined as the tensor that computes dv a according to $dv^a = R^a{}_{bcd}v^b dp^c dq^d$. (There is no standardization in the literature of the order of the indices.)

> **Note**
>
> Section 5.8 discusses the sense in which this approximation is good enough.

> **Example 7: A symmetry of the Riemann tensor**
>
> If vectors dp^c and dq^d lie along the same line, then dv^a must vanish, and interchanging dp^c and dq^d simply reverses the direction of the circuit around the quadrilateral, giving $dv^a \rightarrow -dv^a$. This shows that $R^a{}_{bcd}$ must be *antisymmetric* under interchange of the indices c and d,
>
> $$R^a_{bcd} = -R^a_{bdc}. \tag{5.5.1}$$
>
>
>
> Figure 5.5.1 - The definition of the Riemann tensor. The vector v^b changes by dv^b when parallel-transported around the approximate parallelogram. (v^b is drawn on a scale that makes its length comparable to the infinitesimals dp^c, dq^d, and dv^b; in reality, its size would be greater than theirs by an infinite factor.)
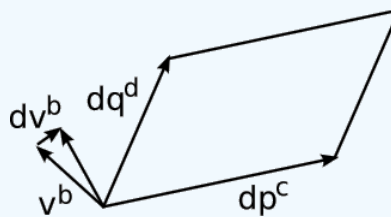
In local normal coordinates, the interpretation of the Riemann tensor becomes particularly transparent. The constant-coordinate lines are geodesics, so when the vector v^b is transported along them, it maintains a constant angle with respect to them. Any rotation of the vector after it is brought around the perimeter of the quadrilateral can therefore be attributed to something that happens at the vertices. In other words, it is simply a measure of the angular defect. We can therefore see that the Riemann tensor is really just a tensorial way of writing the Gaussian curvature

$$K = \frac{d\epsilon}{dA}. \tag{5.5.2}$$

In normal coordinates, the local geometry is nearly Cartesian, and when we take the product of two vectors in an antisymmetric manner, we are essentially measuring the area of the parallelogram they span, as in the three-dimensional vector cross product. We can therefore see that the Riemann tensor tells us something about the amount of curvature contained within the infinitesimal area spanned by dp^c and dq^d. A finite two-dimensional region can be broken down into infinitesimal elements of area, and the Riemann tensor integrated over them. The result is equal to the finite change Δv^b in a vector transported around the whole boundary of the region.

Example 8: Curvature tensors on a sphere

Let's find the curvature tensors on a sphere of radius ρ.

Construct normal coordinates (x, y) with origin O, and let vectors dp^c and dq^d represent infinitesimal displacements along x and y, forming a quadrilateral as described above. Then $R^x{}_{yxy}$ represents the change in the x direction that occurs in a vector that is initially in the y direction. If the vector has unit magnitude, then $R^x{}_{yxy}$ equals the angular deficit of the quadrilateral. Comparing with the definition of the Gaussian curvature, we find $R^x{}_{yxy} = K = \frac{1}{\rho^2}$. Interchanging x and y, we find the same result for R y xyx . Thus although the Riemann tensor in two dimensions has sixteen components, only these two are nonzero, and they are equal to each other.

This result represents the defect in parallel transport around a closed loop per unit area. Suppose we parallel-transport a vector around an octant, as shown in Figure 5.4.2. The area of the octant is $(\frac{\pi}{2})\rho^2$, and multiplying it by the Riemann tensor, we find that the defect in parallel transport is $\frac{\pi}{2}$, i.e., a right angle, as is also evident from the figure.

The above treatment may be somewhat misleading in that it may lead you to believe that there is a single coordinate system in which the Riemann tensor is always constant. This is not the case, since the calculation of the Riemann tensor was only valid near the origin O of the normal coordinates. The character of these coordinates becomes quite complicated far from O; we end up with all our constant-x lines converging at north and south poles of the sphere, and all the constant-y lines at east and west poles.

Angular coordinates (ϕ, θ) are more suitable as a large-scale description of the sphere. We can use the tensor transformation law to find the Riemann tensor in these coordinates. If O, the origin of the (x, y) coordinates, is at coordinates (ϕ, θ), then $\frac{dx}{d\phi} = \rho \sin\theta$ and $\frac{dy}{d\theta} = \rho$. The result is $R^\phi_{\theta\phi\theta} = R^x_{yxy}(\frac{dy}{d\theta})^2 = 1$ and $R^\theta_{\phi\theta\phi} = R^y_{xyx}(\frac{dx}{d\phi})^2 = \sin^2\theta$. The variation in $R^\theta_{\phi\theta\phi}$ is not due to any variation in the sphere's intrinsic curvature; it represents the behavior of the coordinate system.

The Riemann tensor only measures curvature within a particular plane, the one defined by dp^c and dq^d, so it is a kind of sectional curvature. Since we're currently working in two dimensions, however, there is only one plane, and no real distinction between sectional curvature and Ricci curvature, which is the average of the sectional curvature over all planes that include dq^d: $R_{cd} = R^a{}_{cad}$. The Ricci curvature in two spacelike dimensions, expressed in normal coordinates, is simply the diagonal matrix diag(K, K).

5.6: Some Order-of-magnitude Estimates

As a general proposition, calculating an order-of-magnitude estimate of a physical effect requires an understanding of 50% of the physics, while an exact calculation requires about 75%.[7] We've reached the point where it's reasonable to attempt a variety of order-of-magnitude estimates.

> **Note**
>
> This statement is itself only a rough estimate. Anyone who has taught physics knows that students will often calculate an effect exactly while not understanding the underlying physics at all.

The Geodetic Effect

How could we confirm experimentally that parallel transport around a closed path can cause a vector to rotate? The rotation is related to the amount of spacetime curvature contained within the path, so it would make sense to choose a loop going around a gravitating body. The rotation is a purely relativistic effect, so we expect it to be small. To make it easier to detect, we should go around the loop many times, causing the effect to accumulate. This is essentially a description of a body orbiting another body. A gyroscope aboard the orbiting body is expected to precess. This is known as the geodetic effect. In 1916, shortly after Einstein published the general theory of relativity, Willem de Sitter calculated the effect on the earth-moon system. The effect was not directly verified until the 1980's, and the first high-precision measurement was in 2007, from analysis of the results collected by the Gravity Probe B satellite experiment. The probe carried four gyroscopes made of quartz, which were the most perfect spheres ever manufactured, varying from sphericity by no more than about 40 atoms.
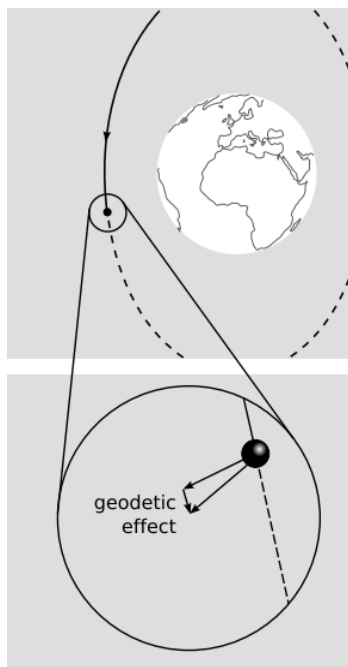
Figure 5.6.1 - The geodetic effect as measured by Gravity Probe B.

Let's estimate the size of the effect. The first derivative of the metric is, roughly, the gravitational field, whereas the second derivative has to do with curvature. The curvature of spacetime around the earth should therefore vary as GMr^{-3}, where M is the earth's mass and G is the gravitational constant. The area enclosed by a circular orbit is proportional to r^2, so we expect the geodetic effect to vary as $\frac{nGM}{r}$, where n is the number of orbits. The angle of precession is unitless, and the only way to make this result unitless is to put in a factor of $\frac{1}{c^2}$. In units with c = 1, this factor is unnecessary. In ordinary metric units, the $\frac{1}{c^2}$ makes sense, because it causes the purely relativistic effect to come out to be small. The result, up to unitless factors that we didn't pretend to find, is

$$\Delta\theta \sim \frac{nGM}{c^2 r}.$$ (5.6.1)

We might also expect a Thomas precession. Like the spacetime curvature effect, it would be proportional to $\frac{nGM}{c^2r}$. Since we're not worrying about unitless factors, we can just lump the Thomas precession together with the effect already calculated.

The data for Gravity Probe B are r = r_e+(650 km) and n ≈ 5000 (orbiting once every 90 minutes for the 353-day duration of the experiment), giving $\Delta\theta \sim 3 \times 10^{-6}$ radians. Figure 5.5.2 shows the actual results[8] the four gyroscopes aboard the probe. The precession was about 6 arc-seconds, or 3×10^{-5} radians. Our crude estimate was on the right order of magnitude. The missing unitless factor on the right-hand side of the equation above is 3π, which brings the two results into fairly close quantitative agreement. The full derivation, including the factor of 3π, is given in section 6.2.
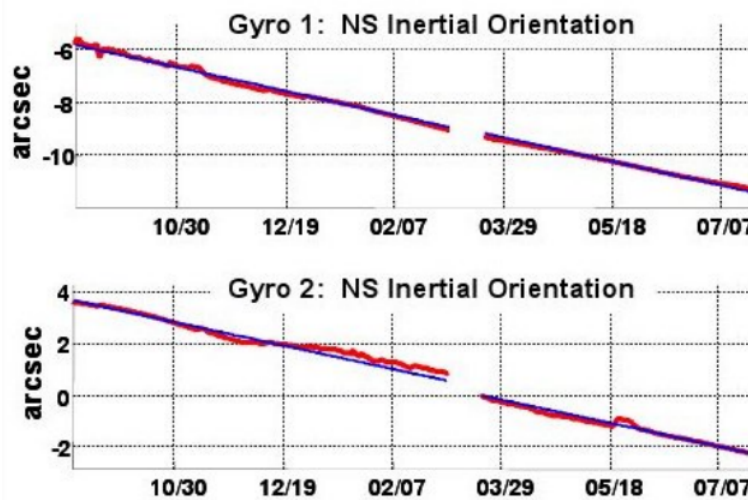
Figure_5.5.2b.png" />

Figure 5.6.2 - *Precession angle as a function of time as measured by the four gyroscopes aboard Gravity Probe B.*

Deflection of Light Rays

In the discussion of the momentum four vector in section 4.2, we saw that due to the equivalence principle, light must be affected by gravity. There are two ways in which such an effect could occur. Light can gain and lose momentum as it travels up and down in a gravitational field, or its momentum vector can be deflected by a transverse gravitational field. As an example of the latter, a ray of starlight can be deflected by the sun's gravity, causing the star's apparent position in the sky to be shifted. The detection of this effect was one of the first experimental tests of general relativity. Ordinarily the bright light from the sun would make it impossible to accurately measure a star's location on the celestial sphere, but this problem was sidestepped by Arthur Eddington during an eclipse of the sun in 1919.

Let's estimate the size of this effect. We've already seen that the Riemann tensor is essentially just a tensorial way of writing the Gaussian curvature $K = \frac{d\epsilon}{dA}$. Suppose, for the sake of this rough estimate, that the sun, earth, and star form a non-Euclidean triangle with a right angle at the sun. Then the angular deflection is the same as the angular defect ϵ of this triangle, and equals the integral of the curvature over the interior of the triangle. Ignoring unitless constants, this ends up being exactly the same calculation as earlier, and the result is $\epsilon \sim \frac{GM}{c^2r}$, where r is the light ray's distance of closest approach to the sun. The value of r can't be less than the radius of the sun, so the maximum size of the effect is on the order of $\frac{GM}{c^2r}$, where M is the sun's mass, and r is its radius. We find $\epsilon \sim 10^{-5}$ radians, or about a second of arc. To measure a star's position to within an arc second was well within the state of the art in 1919, under good conditions in a comfortable observatory. This observation, however, required that Eddington's team travel to the island of Principe, off the coast of West Africa. The weather was cloudy, and only during the last 10 seconds of the seven-minute eclipse did the sky clear enough to allow photographic plates to be taken of the Hyades star cluster against the background of the eclipse-darkened sky. The observed deflection was 1.6 seconds of arc, in agreement with the relativistic prediction. The relativistic prediction is derived in section 6.2.
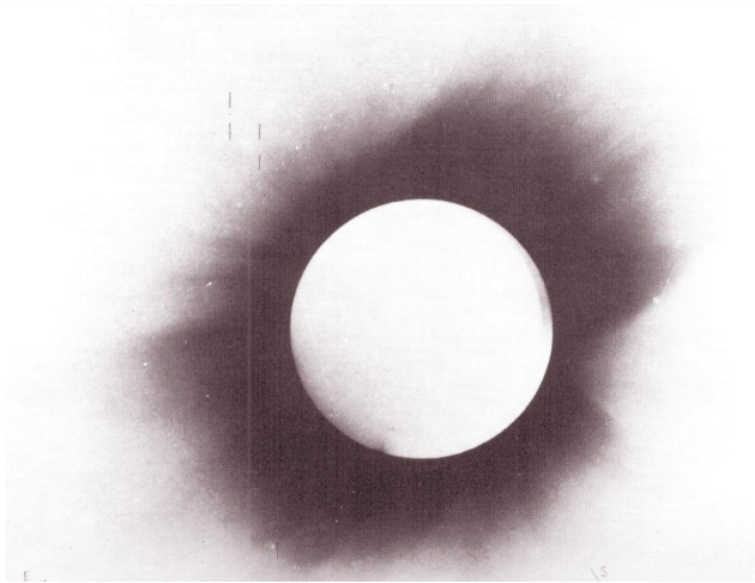
Figure 5.6.3 - One of the photos from Eddington's observations of the 1919 eclipse. This is a photographic negative, so the circle that appears bright is actually the dark face of the moon, and the dark area is really the bright corona of the sun. The stars, marked by lines above and below them, appeared at positions slightly different than their normal ones, indicating that their light had been bent by the sun's gravity on its way to our planet.

References

[8] arxiv.org/abs/1105.3456

5.7: The Covariant Derivative

In the preceding section we were able to estimate a nontrivial general relativistic effect, the geodetic precession of the gyroscopes aboard Gravity Probe B, up to a unitless constant 3π. Let's think about what additional machinery would be needed in order to carry out the calculation in detail, including the 3π.

First we would need to know the Einstein field equation, but in a vacuum this is fairly straightforward:

$$R_{ab} = 0. \tag{5.7.1}$$

Einstein posited this equation based essentially on the considerations laid out in Section 5.1.

But just knowing that a certain tensor vanishes identically in the space surrounding the earth clearly doesn't tell us anything explicit about the structure of the spacetime in that region. We want to know the metric. As suggested at the beginning of the chapter, we expect that the first derivatives of the metric will give a quantity analogous to the gravitational field of Newtonian mechanics, but this quantity will not be directly observable, and will not be a tensor. The second derivatives of the metric are the ones that we expect to relate to the Ricci tensor R_{ab}.

The Covariant Derivative in Electromagnetism

We're talking blithely about derivatives, but it's not obvious how to define a derivative in the context of general relativity in such a way that taking a derivative results in well-behaved tensor.

To see how this issue arises, let's retreat to the more familiar terrain of electromagnetism. In quantum mechanics, the phase of a charged particle's wavefunction is unobservable, so that for example the transformation $\Psi \to -\Psi$ does not change the results of experiments. As a less trivial example, we can redefine the ground of our electrical potential, $\Phi \to \Phi + \delta\Phi$, and this will add a constant onto the energy of every electron in the universe, causing their phases to oscillate at a greater rate due to the quantum-mechanical relation

$$E = hf. \tag{5.7.2}$$

There are no observable consequences, however, because what is observable is the phase of one electron relative to another, as in a double-slit interference experiment. Since every electron has been made to oscillate faster, the effect is simply like letting the conductor of an orchestra wave her baton more quickly; every musician is still in step with every other musician. The rate of change of the wavefunction, i.e., its derivative, has some built-in ambiguity.
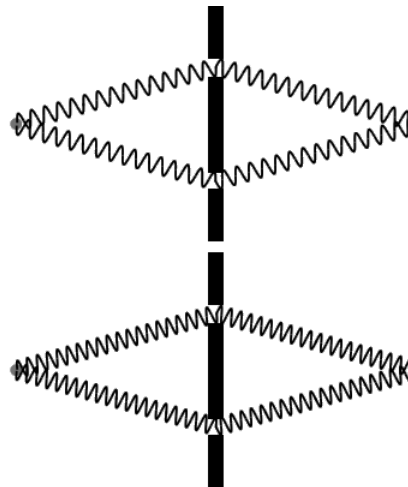
Figure 5.7.1 - A double-slit experiment with electrons. If we add an arbitrary constant to the potential, no observable changes result. The wavelength is shortened, but the relative phase of the two parts of the waves stays the same.

For simplicity, let's now restrict ourselves to spin-zero particles, since details of electrons' polarization clearly won't tell us anything useful when we make the analogy with relativity. For a spin-zero particle, the wavefunction is simply a complex number, and there are no observable consequences arising from the transformation

$$\Psi \to \Psi' = e^{i\alpha}\Psi \tag{5.7.3}$$

where α is a constant. The transformation $\Phi \to \Phi + \delta\Phi$ is also allowed, and it gives $\alpha(t) = (\frac{q\delta\Phi}{\hbar})t$, so that the phase factor $e^{i\alpha(t)}$ is a function of time t. Now from the point of view of electromagnetism in the age of Maxwell, with the electric and magnetic fields imagined as playing their roles against a background of Euclidean space and absolute time, the form of this time-dependent phase factor is very special and symmetrical; it depends only on the absolute time variable. But to a relativist, there is nothing very nice about this function at all, because there is nothing special about a time coordinate. If we're going to allow a function of this form, then based on the coordinate-invariance of relativity, it seems that we should probably allow α to be any function at all of the spacetime coordinates. The proper generalization of $\Phi \to \Phi - \delta\Phi$ is now $A_b \to A_b - \partial_b\alpha$, where A_b is the electromagnetic potential four-vector (section 4.2).
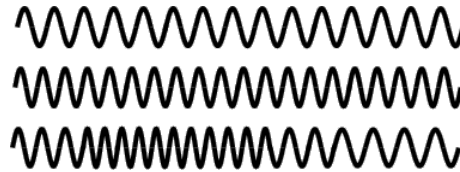
Figure 5.7.2 - Two wavefunctions with constant wavelengths, and a third with a varying wavelength. None of these are physically distinguishable, provided that the same variation in wavelength is applied to all electrons in the universe at any given point in spacetime. There is not even any unambiguous way to pick out the third one as the one with a varying wavelength. We could choose a different gauge in which the third wave was the only one with a constant *wavelength*.

> ### Exercise 5.7.1
>
> Self-check: Suppose we said we would allow α to be a function of t, but forbid it to depend on the spatial coordinates. Prove that this would violate Lorentz invariance.

The transformation has no effect on the electromagnetic fields, which are the direct observables. We can also verify that the change of gauge will have no effect on observable behavior of charged particles. This is because the phase of a wavefunction can only be determined relative to the phase of another particle's wavefunction, when they occupy the same point in space and, for example, interfere. Since the phase shift depends only on the location in spacetime, there is no change in the relative phase.

But bad things will happen if we don't make a corresponding adjustment to the derivatives appearing in the Schrödinger equation. These derivatives are essentially the momentum operators, and they give different results when applied to Ψ' than when applied to Ψ:

$$\begin{aligned}
\partial_b\Psi &\to \partial_b(e^{i\alpha}\Psi) \\
&= e^{i\alpha}\partial_b\Psi + i\partial_b\alpha(e^{i\alpha}\Psi) \\
&= (\partial_b + A'_b - A_b)\Psi'
\end{aligned}$$

To avoid getting incorrect results, we have to do the substitution $\partial_b \to \partial_b + ieA_b$, where the correction term compensates for the change of gauge. We call the operator ∇ defined as

$$\nabla_b = \partial_b + ieA_b \tag{5.7.4}$$

the *covariant derivative*. It gives the right answer regardless of a change of gauge.

The Covariant Derivative in General Relativity

Now consider how all of this plays out in the context of general relativity. The gauge transformations of general relativity are arbitrary smooth changes of coordinates. One of the most basic properties we could require of a derivative operator is that it must give zero on a constant function. A constant scalar function remains constant when expressed in a new coordinate system, but the same is not true for a constant vector function, or for any tensor of higher rank. This is because the change of coordinates changes the units in which the vector is measured, and if the change of coordinates is nonlinear, the units vary from point to point.
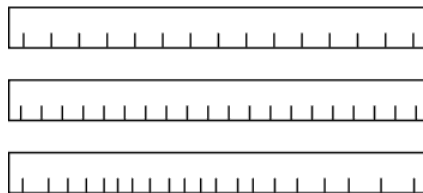
Figure 5.7.2 switching from one set of coordinates to another has no effect on any experimental observables. It is merely a choice of gauge.

Consider the one-dimensional case, in which a vector v^a has only one component, and the metric is also a single number, so that we can omit the indices and simply write v and g. (We just have to remember that v is really a covariant vector, even though we're leaving out the upper index.) If v is constant, its derivative $\frac{dv}{dx}$, computed in the ordinary way without any correction term, is zero. If we further assume that the coordinate x is a normal coordinate, so that the metric is simply the constant g = 1, then zero is not just the answer but the right answer. (The existence of a preferred, global set of normal coordinates is a special feature of a one-dimensional space, because there is no curvature in one dimension. In more than one dimension, there will typically be no possible set of coordinates in which the metric is constant, and normal coordinates only give a metric that is approximately constant in the neighborhood around a certain point. See Figure 5.3.7 for an example of normal coordinates on a sphere, which do not have a constant metric.)

Now suppose we transform into a new coordinate system X, which is not normal. The metric G, expressed in this coordinate system, is not constant. Applying the tensor transformation law, we have $V = v \frac{dX}{dx}$, and differentiation with respect to X will not give zero, because the factor $\frac{dX}{dx}$ isn't constant. This is the wrong answer: V isn't really varying, it just appears to vary because G does.

We want to add a correction term onto the derivative operator $\frac{d}{dX}$, forming a covariant derivative operator ∇_X that gives the right answer. This correction term is easy to find if we consider what the result ought to be when differentiating the metric itself. In general, if a tensor appears to vary, it could vary either because it really does vary or because the metric varies. If the metric *itself* varies, it could be either because the metric really does vary or . . . because the metric varies. In other words, there is no sensible way to assign a nonzero covariant derivative to the metric itself, so we must have $\nabla_X G = 0$. The required correction therefore consists of replacing $\frac{d}{dX}$ with

$$\nabla_X = \frac{d}{dX} - G^{-1} \frac{dG}{dX}. \qquad (5.7.5)$$

Applying this to G gives zero. G is a second-rank contravariant tensor. If we apply the same correction to the derivatives of other second-rank contravariant tensors, we will get nonzero results, and they will be the right nonzero results. For example, the covariant derivative of the stress-energy tensor T (assuming such a thing could have some physical significance in one dimension!) will be $\nabla_X T = \frac{dT}{dX} - G^{-1} \left(\frac{dG}{dX} \right) T$.

Physically, the correction term is a derivative of the metric, and we've already seen that the derivatives of the metric (1) are the closest thing we get in general relativity to the gravitational field, and (2) are not tensors. In 1+1 dimensions, suppose we observe that a free-falling rock has $\frac{dV}{dT}$ = 9.8 m/s². This acceleration cannot be a tensor, because we could make it vanish by changing from Earthfixed coordinates X to free-falling (normal, locally Lorentzian) coordinates x, and a tensor cannot be made to vanish by a change of coordinates. According to a free-falling observer, the vector v isn't changing at all; it is only the variation in the Earth-fixed observer's metric G that makes it appear to change.

Mathematically, the form of the derivative is $\left(\frac{1}{y} \right) \frac{dy}{dx}$, which is known as a logarithmic derivative, since it equals $\frac{d(\ln y)}{dx}$. It measures the *multiplicative* rate of change of y. For example, if y scales up by a factor of k when x increases by 1 unit, then the logarithmic derivative of y is ln k. The logarithmic derivative of e^{cx} is c. The logarithmic nature of the correction term to ∇_X is a good thing, because it lets us take changes of scale, which are multiplicative changes, and convert them to additive corrections to the derivative operator. The additivity of the corrections is necessary if the result of a covariant derivative is to be a tensor, since tensors are additive creatures.

What about quantities that are not second-rank covariant tensors? Under a rescaling of contravariant coordinates by a factor of k, covariant vectors scale by k^{-1}, and second-rank covariant tensors by k^{-2}. The correction term should therefore be half as much for covariant vectors,

$$\nabla_X = \frac{d}{dX} - \frac{1}{2}G^{-1}\frac{dG}{dX}. \tag{5.7.6}$$

and should have an opposite sign for contravariant vectors.

Generalizing the correction term to derivatives of vectors in more than one dimension, we should have something of this form:

$$\nabla_a v^b = \partial_a v^b + \Gamma^b_{ac} v^c$$
$$\nabla_a v_b = \partial_a v_b - \Gamma^c_{ba} v_c,$$

where Γ^b_{ac}, called the Christoffel symbol, does not transform like a tensor, and involves derivatives of the metric. ("Christoffel" is pronounced "Krist-AWful," with the accent on the middle syllable.) The explicit computation of the Christoffel symbols from the metric is deferred until section 5.9, but the intervening sections 5.7 and 5.8 can be omitted on a first reading without loss of continuity.

An important gotcha is that when we evaluate a particular component of a covariant derivative such as $\nabla_2 v^3$, it is possible for the result to be nonzero even if the component v^3 vanishes identically. This can be seen in example 5 and example 21.

Example 9: Christoffel symbols on the globe

As a qualitative example, consider the geodesic airplane trajectory shown in Figure 5.6.4, from London to Mexico City. In physics it is customary to work with the *colatitude*, θ, measured down from the north pole, rather then the latitude, measured from the equator. At P, over the North Atlantic, the plane's colatitude has a minimum. (We can see, without having to take it on faith from the figure, that such a minimum must occur. The easiest way to convince oneself of this is to consider a path that goes directly over the pole, at $\theta = 0$.)
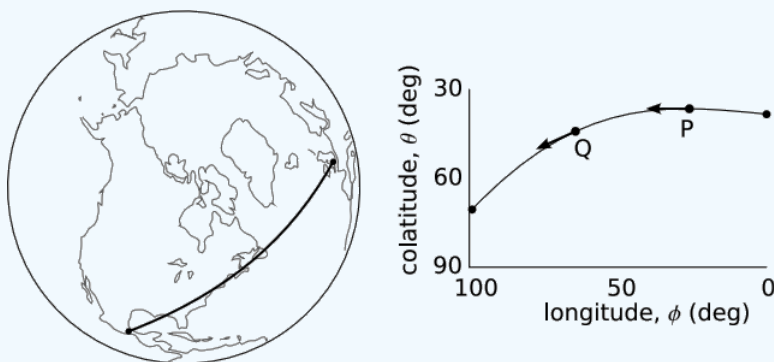
Figure 5.7.4

At P, the plane's velocity vector points directly west. At Q, over New England, its velocity has a large component to the south. Since the path is a geodesic and the plane has constant speed, the velocity vector is simply being parallel-transported; the vector's covariant derivative is zero. Since we have $v_\theta = 0$ at P, the only way to explain the nonzero and positive value of $\partial_\phi v^\theta$ is that we have a nonzero and negative value of $\Gamma^\theta_{\phi\phi}$.

By symmetry, we can infer that $\Gamma^\theta_{\phi\phi}$ must have a positive value in the southern hemisphere, and must vanish at the equator.

$\Gamma^\theta_{\phi\phi}$ is computed in example 10.

Symmetry also requires that this Christoffel symbol be independent of ϕ, and it must also be independent of the radius of the sphere.

Example 9 is in two spatial dimensions. In spacetime, Γ is essentially the gravitational field (see problem 7), and early papers in relativity essentially refer to it that way.[9] This may feel like a joyous reunion with our old friend from freshman mechanics, g = 9.8 m/s. But our old friend has changed. In Newtonian mechanics, accelerations like g are frame-invariant (considering only inertial frames, which are the only legitimate ones in that theory). In general relativity they are frame-dependent, and as we saw earlier, the acceleration of gravity can be made to equal anything we like, based on our choice of a frame of reference.

To compute the covariant derivative of a higher-rank tensor, we just add more correction terms, e.g.,

$$\nabla_a U_{bc} = \partial_a U_{bc} - \Gamma^d_{ba} U_{dc} - \Gamma^d_{ca} U_{bd} \tag{5.7.7}$$

or

$$\nabla_a U_b^c = \partial_a U_b^c - \Gamma_{ba}^d U_d^c + \Gamma_{ad}^c U_b^d. \tag{5.7.8}$$

With the partial derivative ∂_μ, it does not make sense to use the metric to raise the index and form ∂^μ. It *does* make sense to do so with covariant derivatives, so $\nabla^a = g^{ab}\nabla_b$ is a correct identity.

Comma, Semicolon, and Birdtracks Notation

Some authors use superscripts with commas and semicolons to indicate partial and covariant derivatives. The following equations give equivalent notations for the same derivatives:

$$\partial_\mu X_\nu = X_{\nu,\,\mu} \tag{5.7.9}$$

$$\nabla_a X_b = X_{b;a} \tag{5.7.10}$$

$$\nabla^a X_b = X_b^{;a} \tag{5.7.11}$$

Figure 5.6.5 shows two examples of the corresponding birdtracks notation. Because birdtracks are meant to be manifestly coordinate-independent, they do not have a way of expressing non-covariant derivatives. We no longer want to use the circle as a notation for a non-covariant gradient as we did when we first introduced it in section 2.1.

References

[9] "On the gravitational field of a point mass according to Einstein's theory," Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften 1 (1916) 189, translated in arxiv.org/abs/physics/9905030v1.

5.8: The Geodesic Equation

In this section, which can be skipped at a first reading, we show how the Christoffel symbols can be used to find differential equations that describe geodesics.

Characterization of the Geodesic

A geodesic can be defined as a world-line that preserves tangency under parallel transport, Figure 5.8.1. This is essentially a mathematical way of expressing the notion that we have previously expressed more informally in terms of "staying on course" or moving "inertially."
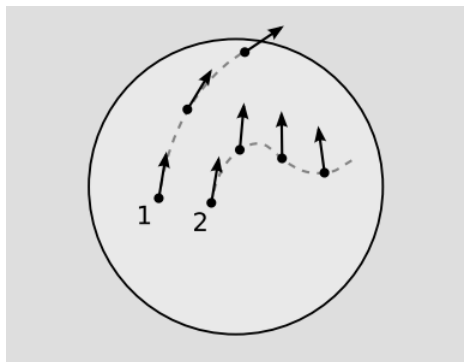
Figure 5.8.1: The geodesic, 1, preserves tangency under parallel transport. The non-geodesic curve, 2, doesn't have this property; a vector initially tangent to the curve is no longer tangent to it when parallel-transported along it.

A curve can be specified by giving functions $x^\mu(\lambda)$ for its coordinates, where λ is a real parameter. A vector lying tangent to the curve can then be calculated using partial derivatives, $T^\mu = \frac{\partial x^\mu}{\partial \lambda}$. There are three ways in which a vector function of λ could change: (1) it could change for the trivial reason that the metric is changing, so that its components changed when expressed in the new metric; (2) it could change its components perpendicular to the curve; or (3) it could change its component parallel to the curve. Possibility 1 should not really be considered a change at all, and the definition of the covariant derivative is specifically designed to be insensitive to this kind of thing. 2 cannot apply to T^μ, which is tangent by construction. It would therefore be convenient if T^μ happened to be always the same length. If so, then 3 would not happen either, and we could reexpress the definition of a geodesic by saying that the covariant derivative of T^μ was zero. For this reason, we will assume for the remainder of this section that the parametrization of the curve has this property. In a Newtonian context, we could imagine the x^μ to be purely spatial coordinates, and λ to be a universal time coordinate. We would then interpret T^μ as the velocity, and the restriction would be to a parametrization describing motion with constant speed. In relativity, the restriction is that λ must be an affine parameter. For example, it could be the proper time of a particle, if the curve in question is timelike.

Covariant Derivative with Respect to a Parameter

The notation of section 5.6 is not quite adapted to our present purposes, since it allows us to express a covariant derivative with respect to one of the coordinates, but not with respect to a parameter such as λ. We would like to notate the covariant derivative of T^μ with respect to λ as $\nabla_\lambda T^\mu$, even though λ isn't a coordinate. To connect the two types of derivatives, we can use a total derivative. To make the idea clear, here is how we calculate a total derivative for a scalar function f(x, y), without tensor notation:

$$\frac{df}{d\lambda} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial \lambda} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial \lambda} \tag{5.8.1}$$

This is just the generalization of the chain rule to a function of two variables. For example, if λ represents time and f temperature, then this would tell us the rate of change of the temperature as a thermometer was carried through space. Applying this to the present problem, we express the total covariant derivative as

$$\nabla_\lambda T^\mu = (\nabla_\kappa T^\mu)\frac{dx^\kappa}{d\lambda}$$
$$= (\partial_\kappa T^\mu + \Gamma^\mu_{\kappa\nu} T^\nu)\frac{dx^\kappa}{d\lambda}.$$

The Geodesic Equation

Recognizing $\frac{\partial_\kappa T^\mu dx^\kappa}{d\lambda}$ as a total non-covariant derivative, we find

$$\nabla_\lambda T^\mu = \frac{dT^\mu}{d\lambda} + \Gamma^\mu_{\kappa\nu} T^\nu \frac{dx^\kappa}{d\lambda}. \tag{5.8.2}$$

This is known as the geodesic equation. There is a factor of two that is a common gotcha when applying this equation. The symmetry of the Christoffel symbols $\Gamma^\mu_{\kappa\nu} = \Gamma^\mu_{\nu\kappa}$ implies that when κ and ν are distinct, the same term will appear twice in the summation.

If this differential equation is satisfied for one affine parameter λ, then it is also satisfied for any other affine parameter $\lambda' = a\lambda + b$, where a and b are constants (problem 5). Recall that affine parameters are only defined along geodesics, not along arbitrary curves.

We can't start by defining an affine parameter and then use it to find geodesics using this equation, because we can't define an affine parameter without *first* specifying a geodesic. Likewise, we can't do the geodesic first and then the affine parameter, because if we already had a geodesic in hand, we wouldn't need the differential equation in order to find a geodesic. The solution to this chicken-and-egg conundrum is to write down the differential equations and try to find a solution, without trying to specify either the affine parameter or the geodesic in advance. We will seldom have occasion to resort to this technique, an exception being example 19.

Uniqueness

The geodesic equation is useful in establishing one of the necessary theoretical foundations of relativity, which is the uniqueness of geodesics for a given set of initial conditions. This is related to axiom O1 of ordered geometry, that two points determine a line, and is necessary physically for the reasons discussed in section 1.5; briefly, if the geodesic were not uniquely determined, then particles would have no way of deciding how to move. The form of the geodesic equation guarantees uniqueness. To see this, consider the following algorithm for determining a numerical approximation to a geodesic:

1. Initialize λ, the x^μ and their derivatives $\frac{dx^\mu}{d\lambda}$. Also, set a small step-size $\Delta\lambda$ by which to increment λ at each step below.
2. For each i, calculate $\frac{d^2 x^\mu}{d\lambda^2}$ using the geodesic equation.
3. Add $\left(\frac{d^2 x^\mu}{d\lambda^2}\right)\Delta\lambda$ to the currently stored value of $\frac{dx^\mu}{d\lambda}$.
4. Add $\left(\frac{dx^\mu}{d\lambda}\right)\Delta\lambda$ to x^μ.
5. Add $\Delta\lambda$ to λ.
6. Repeat steps 2-5 until the geodesic has been extended to the desired affine distance.

Since the result of the calculation depends only on the inputs at step 1, we find that the geodesic is uniquely determined.

To see that this is really a valid way of proving uniqueness, it may be helpful to consider how the proof could have failed. Omitting some of the details of the tensors and the multidimensionality of the space, the form of the geodesic equation is essentially $\ddot{x} + f\dot{x}^2 = 0$, where dots indicate derivatives with respect to λ. Suppose that it had instead had the form $\ddot{x}^2 + f\dot{x} = 0$. Then at step 2 we would have had to pick either a positive or a negative square root for $\ddot{x}$. Although continuity would usually suffice to maintain a consistent sign from one iteration to the next, that would not work if we ever came to a point where $\ddot{x}$ vanished momentarily. An equation of this form therefore would not have a unique solution for a given set of initial conditions.

The practical use of this algorithm to compute geodesics numerically is demonstrated in section 5.9.

5.9: Torsion

This section describes the concept of gravitational torsion. It can be skipped without loss of continuity, provided that you accept the symmetry property $\Gamma^a_{[bc]} = 0$ without worrying about what it means physically or what empirical evidence supports it.

> **Exercise 5.9.1**
>
> Self-check: Interpret the mathematical meaning of the equation $\Gamma^a_{[bc]} = 0$, which is expressed in the notation introduced earlier.

Are Scalars Path-dependent?

It seems clear that something like the covariant derivative is needed for vectors, since they have a direction in spacetime, and thus their measures vary when the measure of spacetime itself varies. Since scalars don't have a direction in spacetime, the same reasoning doesn't apply to them, and this is reflected in our rules for covariant derivatives. The covariant derivative has one Γ term for every index of the tensor being differentiated, so for a scalar there should be no Γ terms at all, i.e., ∇_a is the same as ∂_a.

But just because derivatives of scalars don't require special treatment *for this particular reason*, that doesn't mean they are guaranteed to behave as we intuitively expect, in the strange world of coordinate-invariant relativity.

One possible way for scalars to behave counterintuitively would be by analogy with parallel transport of vectors. If we stick a vector in a box (as with, e.g., the gyroscopes aboard Gravity Probe B) and carry it around a closed loop, it changes. Could the same happen with a scalar? This is extremely counterintuitive, since there is no reason to imagine such an effect in any of the models we've constructed of curved spaces. In fact, it is not just counterintuitive but mathematically impossible, according to the following argument. The only reason we can interpret the vector-in-a-box effect as arising from the geometry of spacetime is that it applies equally to all vectors. If, for example, it only applied to the magnetic polarization vectors of ferromagnetic substances, then we would interpret it as a magnetic field living in spacetime, not a property of spacetime itself. If the value of a scalar-in-a-box was path-dependent, and this path-dependence was a geometric property of spacetime, then it would have to apply to all scalars, including, say, masses and charges of particles. Thus if an electron's mass increased by 1% when transported in a box along a certain path, its charge would have to increase by 1% as well. But then its charge-to-mass ratio would remain invariant, and this is a contradiction, since the charge-to-mass ratio is also a scalar, and should have felt the same 1% effect. Since the varying scalar-in-a-box idea leads to a contradiction, it wasn't a coincidence that we couldn't find a model that produced such an effect; a theory that lacks self-consistency doesn't have any models.

> **Exercise 5.9.2**
>
> Self-check: Explain why parallel transporting a vector can only rotate it, not change its magnitude.

There is, however, a different way in which scalars could behave counterintuitively, and this one is mathematically self-consistent. Suppose that Helen lives in two spatial dimensions and owns a thermometer. She wants to measure the spatial variation of temperature, in particular its mixed second derivative $\frac{\partial^2 T}{\partial x \partial y}$. At home in the morning at point A, she prepares by calibrating her gyrocompass to point north and measuring the temperature. Then she travels $\ell = 1$ km east along a geodesic to B, consults her gyrocompass, and turns north. She continues one kilometer north to C, samples the change in temperature ΔT_1 relative to her home, and then retraces her steps to come home for lunch. In the afternoon, she checks her work by carrying out the same process, but this time she interchanges the roles of north and east, traveling along ADE. If she were living in a flat space, this would form the other two sides of a square, and her afternoon temperature sample ΔT_2 would be at the same point in space C as her morning sample. She actually doesn't recognize the landscape, so the sample points C and E are different, but this just confirms what she already knew: the space isn't flat.[10]
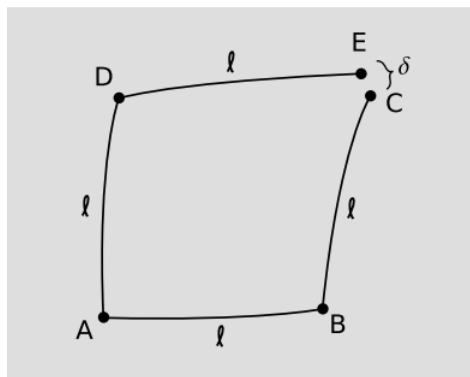
Figure 5.9.1 - Measuring $\frac{\partial^2 T}{\partial x \partial y}$ for a scalar T.
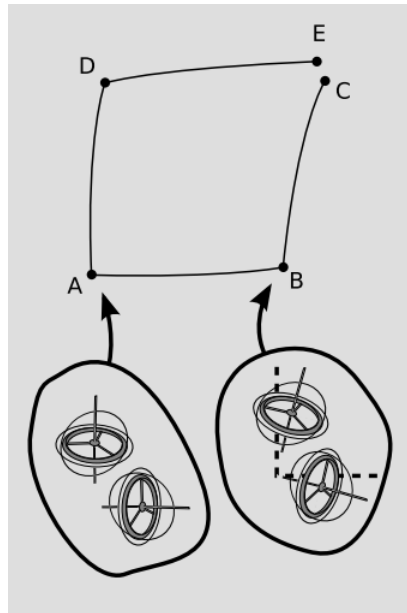
Figure 5.9.2 - The gyroscopes both rotate when transported from A to B, causing Helen to navigate along BC, which does not form a right angle with AB. The angle between the two gyroscopes' axes is always the same, so the rotation is not locally observable, but it does produce an observable gap between C and E.

> **Note**
>
> This point was mentioned in section 5.4, in connection with the definition of the Riemann tensor.

None of this seems surprising yet, but there are now two qualitatively different ways that her analysis of her data could turn out, indicating qualitatively different things about the laws of physics in her universe. The definition of the derivative as a limit requires that she repeat the experiment at smaller scales. As $\ell \to 0$, the result for $\frac{\partial^2 T}{\partial x \partial y}$ should approach a definite limit, and the error should diminish in proportion to ℓ. In particular the difference between the results inferred from ΔT_1 and ΔT_2 indicate an error, and the discrepancy between the second derivatives inferred from them should shrink appropriately as ℓ shrinks. Suppose this doesn't happen. Since partial derivatives commute, we conclude that her measuring procedure is not the same as a partial derivative. Let's call her measuring procedure ∇, so that she is observing a discrepancy between $\nabla_x \nabla_y$ and $\nabla_y \nabla_x$. The fact that the commutator $\nabla_x \nabla_y - \nabla_y \nabla_x$ doesn't vanish cannot be explained by the Christoffel symbols, because what she's differentiating is a scalar. Since the discrepancy arises entirely from the failure of $\Delta T_1 - \Delta T_2$ to scale down appropriately, the conclusion is that the distance δ between the two sampling points is not scaling down as quickly as we expect. In our familiar models of two-dimensional spaces as surfaces embedded in three-space, we always have $\delta \sim \ell^3$ for small ℓ, but she has found that it only shrinks as quickly as ℓ^2.

For a clue as to what is going on, note that the commutator $\nabla_x \nabla_y - \nabla_y \nabla_x$ has a particular handedness to it. For example, it flips its sign under a reflection across the line y = x. When we "parallel"-transport vectors, they aren't actually staying parallel. In this hypothetical universe, a vector in a box transported by a small distance ℓ rotates by an angle proportional to ℓ. This effect is called

torsion. Although no torsion effect shows up in our familiar models, that is not because torsion lacks self-consistency. Models of spaces with torsion do exist. In particular, we can see that torsion doesn't lead to the same kind of logical contradiction as the varying-scalar-in-a-box idea. Since all vectors twist by the same amount when transported, inner products are preserved, so it is not possible to put two vectors in one box and get the scalar-in-a-box paradox by watching their inner product change when the box is transported.

Note that the elbows ABC and ADE are not right angles. If Helen had brought a pair of gyrocompasses with her, one for x and one for y, she would have found that the right angle between the gyrocompasses was preserved under parallel transport, but that a gyrocompass initially tangent to a geodesic did not remain so. There are in fact two inequivalent definitions of a geodesic in a space with torsion. The shortest path between two points is not necessarily the same as the straightest possible path, i.e., the one that parallel-transports its own tangent vector.

The Torsion Tensor

Since torsion is odd under parity, it must be represented by an odd-rank tensor, which we call τ_{ab}^{c} and define according to

$$(\nabla_a \nabla_b - \nabla_b \nabla_a)f = -\tau_{ab}^{c}\nabla_c f, \tag{5.9.1}$$

where f is any scalar field, such as the temperature in the preceding section. There are two different ways in which a space can be non-Euclidean: it can have curvature, or it can have torsion. For a full discussion of how to handle the mathematics of a spacetime with both curvature and torsion, see the article by Steuard Jensen at http://www.slimy.com/~steuard/teaching/tutorials/GRtorsion.pdf. For our present purposes, the main mathematical fact worth noting is that vanishing torsion is equivalent to the symmetry $\Gamma_{bc}^{a} = \Gamma_{cb}^{a}$ of the Christoffel symbols. Using the notation introduced earlier, $\Gamma_{[bc]}^{a} = 0$ if $\tau = 0$.

> ### Exercise 5.9.3
>
> Self-check: Use an argument similar to the one in example 5 to prove that no model of a two-space embedded in a three-space can have torsion.

Generalizing to more dimensions, the torsion tensor is odd under the full spacetime reflection $x_a \rightarrow -x_a$, i.e., a parity inversion plus a time-reversal, PT.

In the story above, we had a torsion that didn't preserve tangent vectors. In three or more dimensions, however, it is possible to have torsion that does preserve tangent vectors. For example, transporting a vector along the x axis could cause only a rotation in the y-z plane. This relates to the symmetries of the torsion tensor, which for convenience we'll write in an x-y-z coordinate system and in the fully covariant form $\tau_{\lambda\mu\nu}$. The definition of the torsion tensor implies $\tau_{\lambda(\mu\nu)} = 0$, i.e., that the torsion tensor is antisymmetric in its two final indices. Torsion that does not preserve tangent vectors will have nonvanishing elements such as τ_{xxy}, meaning that parallel-transporting a vector along the x axis can change its x component. Torsion that preserves tangent vectors will have vanishing $\tau_{\lambda\mu\nu}$ unless $\lambda, \mu,$ and ν are all distinct. This is an example of the type of antisymmetry that is familiar from the vector cross product, in which the cross products of the basis vectors behave as $\mathbf{x} \times \mathbf{y} = \mathbf{z}, \mathbf{y} \times \mathbf{z} = \mathbf{x}, \mathbf{y} \times \mathbf{z} = \mathbf{x}$. Generalizing the notation for symmetrization and antisymmetrization of tensors from earlier, we have

$$T_{(abc)} = \frac{1}{3!}\Sigma T_{abc} \tag{5.9.2}$$

$$T_{[abc]} = \frac{1}{3!}\Sigma \epsilon^{abc} T_{abc}, \tag{5.9.3}$$

where the sums are over all permutations of the indices, and in the second line we have used the Levi-Civita symbol. In this notation, a totally antisymmetric torsion tensor is one with $\tau_{\lambda\mu\nu} = \tau_{[\lambda\mu\nu]}$, and torsion of this type preserves tangent vectors under translation.
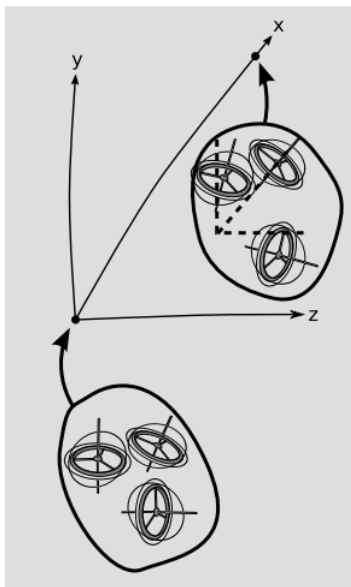
Figure 5.9.3 - Three gyroscopes are initially aligned with the x, y, and z axes. After parallel transport along the geodesic x axis, the x gyro is still aligned with the x axis, but the y and z gyros have rotated.

In two dimensions, there are no totally antisymmetric objects with three indices, because we can't write three indices without repeating one. In three dimensions, an antisymmetric object with three indices is simply a multiple of the Levi-Civita tensor, so a totally antisymmetric torsion, if it exists, is represented by a single number; under translation, vectors rotate like either right-handed or left-handed screws, and this number tells us the rate of rotation. In four dimensions, we have four independently variable quantities, $\tau_{xyz}, \tau_{tyz}, \tau_{txz}$, and τ_{txy}. In other words, an antisymmetric torsion of 3+1 spacetime can be represented by a four-vector, $\tau^a = \epsilon^{abcd}\tau_{bcd}$.

Experimental Searches for Torsion

One way of stating the equivalence principle (see section 4.4) is that it forbids spacetime from coming equipped with a vector field that could be measured by free-falling observers, i.e., observers in local Lorentz frames. A variety of high-precision tests of the equivalence principle have been carried out. From the point of view of an experimenter doing this kind of test, it is important to distinguish between fields that are "built in" to spacetime and those that live in spacetime. For example, the existence of the earth's magnetic field does not violate the equivalence principle, but if an experiment was sensitive to the earth's field, and the experimenter didn't know about it, there would appear to be a violation. Antisymmetric torsion in four dimensions acts like a vector. If it constitutes a universal background effect built into spacetime, then it violates the equivalence principle. If it instead arises from specific material sources, then it may still show up as a measurable effect in experimental tests designed to detect Lorentz-invariance. Let's consider the latter possibility.

Since curvature in general relativity comes from mass and energy, as represented by the stress-energy tensor T_{ab}, we could ask what would be the sources of torsion, if it exists in our universe. The source can't be the rank-2 stress-energy tensor. It would have to be an odd-rank tensor, i.e., a quantity that is odd under PT, and in theories that include torsion it is commonly assumed that the source is the quantum-mechanical angular momentum of subatomic particles. If this is the case, then torsion effects are expected to be proportional to $\hbar G$, the product of Planck's constant and the gravitational constant, and they should therefore be extremely small and hard to measure. String theory, for example, includes torsion, but nobody has found a way to test string theory empirically because it essentially makes predictions about phenomena at the Planck scale, $\sqrt{\frac{\hbar G}{c^3}} \sim 10^{-35}$ m, where both gravity and quantum mechanics are strong effects.

There are, however, some high-precision experiments that have a reasonable chance of detecting whether our universe has torsion. Torsion violates the equivalence principle, and by the turn of the century tests of the equivalence principle had reached a level of precision sufficient to rule out some models that include torsion. Figure 5.8.4 shows a torsion pendulum used in an experiment by the Eöt-Wash group at the University of Washington.[11] If torsion exists, then the intrinsic spin σ of an electron should have an energy $\sigma \cdot \tau$, where τ is the spacelike part of the torsion vector. The torsion could be generated by the earth, the sun, or some other object at a greater distance. The interaction $\sigma \cdot \tau$ will modify the behavior of a torsion pendulum if the spins of the electrons in the pendulum are polarized nonrandomly, as in a magnetic material. The pendulum will tend to precess around the axis defined by τ.
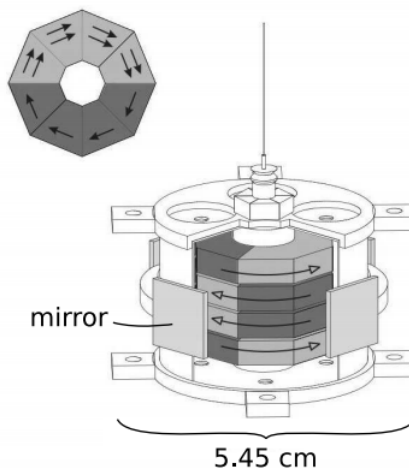
Figure 5.9.4 - The University of Washington torsion pendulum used to search for torsion. The light gray wedges are Alnico, the darker ones SmCo5. The arrows with the filled heads represent the directions of the electron spins, with denser arrows indicating higher polarization. The arrows with the open heads show the direction of the **B** field.

This type of experiment is extremely difficult, because the pendulum tends to act as an ultra-sensitive magnetic compass, resulting in a measurement of the ambient magnetic field rather than the hypothetical torsion field $\boldsymbol{\tau}$. To eliminate this source of systematic error, the UW group first eliminated the ambient magnetic field as well as possible, using mu-metal shielding and Helmholtz coils. They also constructed the pendulum out of a combination of two magnetic materials, Alnico 5 and SmCo$_5$, in such a way that the magnetic dipole moment vanished, but the spin dipole moment did not; Alnico 5's magnetic field is due almost entirely to electron spin, whereas the magnetic field of SmCo$_5$ contains significant contributions from orbital motion. The result was a nonmagnetic object whose spins were polarized. After four years of data collection, they found $|\boldsymbol{\tau}| \lesssim 10^{-21}$ eV. Models that include torsion typically predict such effects to be of the order of $\frac{m_e^2}{m_P} \sim 10^{-17}$ eV, where m_e is the mass of the electron and $m_P = \sqrt{\frac{\hbar c}{G}} \approx 10^{19}$ GeV ≈ 20 μg is the Planck mass. A wide class of these models is therefore ruled out by these experiments.

Since there appears to be no experimental evidence for the existence of gravitational torsion in our universe, we will assume from now on that it vanishes identically. Einstein made the same assumption when he originally created general relativity, although he and Cartan later tinkered with non-torsion-free theories in a failed attempt to unify gravity with electromagnetism. Some models that include torsion remain viable. For example, it has been argued that the torsion tensor should fall off quickly with distance from the source.[12]

References

[11] http://arxiv.org/abs/hep-ph/0606218

[12] Carroll and Field, http://arxiv.org/abs/gr-qc/9403058

5.10: From Metric to Curvature

Finding the Christoffel Symbol from the Metric

We've already found the Christoffel symbol in terms of the metric in one dimension. Expressing it in tensor notation, we have

$$\Gamma^d_{ba} = \frac{1}{2}g^{cd}(\partial_? g_{??}),\tag{5.10.1}$$

where inversion of the one-component matrix G has been replaced by matrix inversion, and, more importantly, the question marks indicate that there would be more than one way to place the subscripts so that the result would be a grammatical tensor equation. The most general form for the Christoffel symbol would be

$$\Gamma^b_{ac} = \frac{1}{2}g^{db}(L\partial_c g_{ab} + M\partial_a g_{cb} + N\partial_b g_{ca}),\tag{5.10.2}$$

where L, M, and N are constants. Consistency with the onedimensional expression requires

$$L + M + N = 1\tag{5.10.3}$$

and vanishing torsion gives $L = M$. The L and M terms have a different physical significance than the N term.

Suppose an observer uses coordinates such that all objects are described as lengthening over time, and the change of scale accumulated over one day is a factor of k > 1. This is described by the derivative $\partial_t g_{xx} < 1$, which affects the M term. Since the metric is used to calculate squared distances, the g_{xx} matrix element scales down by $\frac{1}{\sqrt{k}}$. To compensate for $\partial_t v^x < 0$, so we need to add a positive correction term, $M > 0$, to the covariant derivative. When the same observer measures the rate of change of a vector v^t with *respect* to space, the rate of change comes out to be too *small*, because the variable she differentiates with respect to is too big. This requires $N < 0$, and the correction is of the same size as th$M$ correction, so $|M| = |N|$. We find $L = M = -N = 1$.

> **Exercise 5.10.1**
>
> Does the above argument depend on the use of space for one coordinate and time for the other?

The resulting general expression for the Christoffel symbol in terms of the metric is

$$\Gamma^c_{ab} = \frac{1}{2}g^{cd}(\partial_a g_{bd} + \partial_b g_{ad} - \partial_d g_{ab}).\tag{5.10.4}$$

One can readily go back and check that this gives

$$\nabla_c g_{ab} = 0.\tag{5.10.5}$$

> Confirming Equation 5.10.5 is a bit tedious. For that matter, tensor calculations in general can be infamously time-consuming and error-prone. Any reasonable person living in the 21st century will therefore resort to a computer algebra system. The most widely used computer algebra system is Mathematica, but it's expensive and proprietary, and it doesn't have extensive built-in facilities for handling tensors. It turns out that there is quite a bit of free and open-source tensor software, and it falls into two classes: coordinate-based and coordinate-independent. The best open-source coordinate-independent facility available appears to be Cadabra, and in fact the verification of $\nabla_c g_{ab} = 0$ is the first example given in the Leo Brewin's handy guide to applications of Cadabra to general relativity.[13]

> **Exercise 5.10.2**
>
> In the case of 1 dimension, show that this reduces to the earlier result of $-(\frac{1}{2})\frac{dG}{dX}$.

Since Γ is not a tensor, it is not obvious that the covariant derivative, which is constructed from it, is a tensor. But if it isn't obvious, neither is it surprising – the goal of the above derivation was to get results that would be coordinate-independent.

Example 10: Christoffel symbols on the globe, quantitatively

In example 9, we inferred the following properties for the Christoffel symbol $\Gamma^\theta_{\phi\phi}$ on a sphere of radius R: $\Gamma^\theta_{\phi\phi}$ is independent of ϕ and R, $\Gamma^\theta_{\phi\phi} < 0$ in the northern hemisphere (colatitude θ less than $\pi/2$), $\Gamma^\theta_{\phi\phi} = 0$ on the equator, and $\Gamma^\theta_{\phi\phi} > 0$ in the southern hemisphere.

The metric on a sphere is

$$ds^2 = R^2 d\theta^2 + R^2 \sin^2\theta d\phi^2. \tag{5.10.6}$$

The only nonvanishing term in the expression for $\Gamma^\theta_{\phi\phi}$ is the one involving $\partial_\theta g_{\phi\phi} = 2R^2 \sin\theta\cos\theta$. The result is

$$\Gamma^\theta_{\phi\phi} = -\sin\theta\cos\theta \tag{5.10.7}$$

which can be verified to have the properties claimed above.

Numerical Solution of the Geodesic Equation

In Section 5.7, I gave an algorithm that demonstrated the uniqueness of the solutions to the geodesic equation. This algorithm can also be used to find geodesics in cases where the metric is known. The following program, written in the computer language Python, carries out a very simple calculation of this kind, in a case where we know what the answer should be; even without any previous familiarity with Python, it shouldn't be difficult to see the correspondence between the abstract algorithm presented in Section 5.7 and its concrete realization below. For polar coordinates in a Euclidean plane, one can compute $\Gamma^r_{\phi\phi} = -r$ and $\Gamma^\phi_{r\phi} = \frac{1}{r}$ (problem 2). Here we compute the geodesic that starts out tangent to the unit circle at $\phi = 0$.

```
1   import math
2
3   l = 0     # affine parameter lambda
4   dl = .001  # change in l with each iteration
5   l_max = 100.
6
7   # initial position:
8   r=1
9   phi=0
10  # initial derivatives of coordinates w.r.t. lambda
11  vr = 0
12  vphi = 1
13
14  k = 0 # keep track of how often to print out updates
15  while l<l_max:
16    l = l+dl
17    # Christoffel symbols:
18    Grphiphi = -r
19    Gphirphi = 1/r
20    # second derivatives:
21    ar   = -Grphiphi*vphi*vphi
22    aphi = -2.*Gphirphi*vr*vphi
23        # ... factor of 2 because G^a_{bc}=G^a_{cb} and b
24        #      is not the same as c
25    # update velocity:
26    vr = vr + dl*ar
27    vphi = vphi + dl*aphi
28    # update position:
29    r = r + vr*dl
30    phi = phi + vphi*dl
31    if k%10000==0: # k is divisible by 10000
32      phi_deg = phi*180./math.pi
33      print "lambda=%6.2f   r=%6.2f   phi=%6.2f deg." % (l,r,phi_deg)
34    k = k+1
```

It is not necessary to worry about all the technical details of the language (e.g., line 1, which makes available such conveniences as math.pi for π). Comments are set off by pound signs. Lines 16-34 are indented because they are all to be executed repeatedly, until it is no longer true that $\lambda < \lambda_{max}$ (line 15).

Exercise 5.10.3

By inspecting lines 18-22, find the signs of $\ddot{r}$ and $\ddot{\phi}$ at $\lambda = 0$. Convince yourself that these signs are what we expect geometrically.

The output is as follows:

```
 1   lambda=  0.00   r=  1.00   phi=  0.06 deg.
 2   lambda= 10.00   r= 10.06   phi= 84.23 deg.
 3   lambda= 20.00   r= 20.04   phi= 87.07 deg.
 4   lambda= 30.00   r= 30.04   phi= 88.02 deg.
 5   lambda= 40.00   r= 40.04   phi= 88.50 deg.
 6   lambda= 50.00   r= 50.04   phi= 88.78 deg.
 7   lambda= 60.00   r= 60.05   phi= 88.98 deg.
 8   lambda= 70.00   r= 70.05   phi= 89.11 deg.
 9   lambda= 80.00   r= 80.06   phi= 89.21 deg.
10   lambda= 90.00   r= 90.06   phi= 89.29 deg.
```

We can see that $\phi \to 90$ deg. as $\lambda \to \infty$, which makes sense, because the geodesic is a straight line parallel to the y axis. A less trivial use of the technique is demonstrated in Section 6.2, where we calculate the deflection of light rays in a gravitational field, one of the classic observational tests of general relativity.

The Riemann Tensor in Terms of the Christoffel Symbols

The covariant derivative of a vector can be interpreted as the rate of change of a vector in a certain direction, relative to the result of parallel-transporting the original vector in the same direction. We can therefore see that the definition of the Riemann curvature tensor in Section 5.4 is a measure of the failure of covariant derivatives to commute:

$$(\nabla_a \nabla_b - \nabla_b \nabla_a)A^c = A^d R^c_{dab} \tag{5.10.8}$$

A tedious calculation now gives R in terms of the Γs:

$$R^a_{bcd} = \partial_c \Gamma^a_{db} - \partial_d \Gamma^a_{cb} + \Gamma^a_{ce}\Gamma^e_{db} - \Gamma^a_{de}\Gamma^e_{cb} \tag{5.10.9}$$

This is given as another example later in Brewin's manual for applying Cadabra to general relativity.[14] (Brewin writes the upper index in the second slot of R.)

Some General Ideas about Gauge

Let's step back now for a moment and try to gain some physical insight by looking at the features that the electromagnetic and relativistic gauge transformations have in common. We have the following analogies:

	electromagnetism	*differential geometry*
global symmetry	A constant phase shift α has no observable effects.	Adding a constant onto a coordinate has no observable effects.
local symmetry	A phase shift α that varies from point to point has no observable effects.	An arbitrary coordinate transformation has no observable effects.
The gauge is described by . . .	α	$g_{\mu\nu}$
. . . and differentiation of this gives the gauge field. . .	A_b	Γ^c_{ab}
A second differentiation gives the directly observable field(s) . . .	**E** and **B**	R^c_{dab}

The interesting thing here is that the directly observable fields do not carry all of the necessary information, but the gauge fields are not directly observable. In electromagnetism, we can see this from the *Aharonov-Bohm effect*, shown in Figure 5.10.1[15] The solenoid has **B** = 0 externally, and the electron beams only ever move through the external region, so they never experience any magnetic field. Experiments show, however, that turning the solenoid on and off does change the interference between the two beams. This is because the vector potential does not vanish outside the solenoid, and as we've seen in section 4.2, the phase of the beams varies according to the path integral of the A_b. We are therefore left with an uncomfortable, but unavoidable, situation. The concept of a field is supposed to eliminate the need for instantaneous action at a distance, which is forbidden by relativity; that is, (1) we want our fields to have only local effects. On the other hand, (2) we would like our fields to be directly observable quantities. We cannot have both 1 and 2. The gauge field satisfies 1 but not 2, and the electromagnetic fields give 2 but not 1.
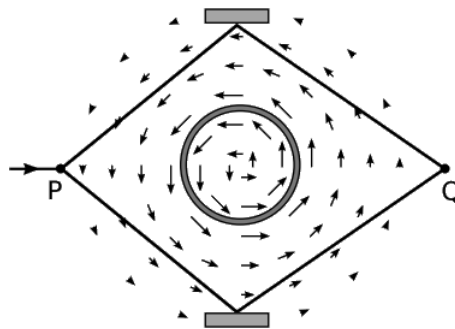
Figure 5.10.1: The Aharonov-Bohm effect. An electron enters a beam splitter at P, and is sent out in two different directions. The two parts of the wave are reflected so that they reunite at Q. The arrows represent the vector potential **A**. The observable magnetic field **B** is zero everywhere outside the solenoid, and yet the interference observed at Q depends on whether the field is turned on. See section 4.2 for further discussion of the **A** and **B** fields of a solenoid.

> **Note**
>
> We describe the effect here in terms of an idealized, impractical experiment. For the actual empirical status of the Aharonov-Bohm effect, see Batelaan and Tonomura, Physics Today 62 (2009) 38.

Figure 5.9.2 shows an analog of the Aharonov-Bohm experiment in differential geometry. Everywhere but at the tip, the cone has zero curvature, as we can see by cutting it and laying it out flat. But even an observer who never visits the tightly curved region at the tip can detect its existence, because parallel-transporting a vector around a closed loop can change the vector's direction, provided that the loop surrounds the tip.
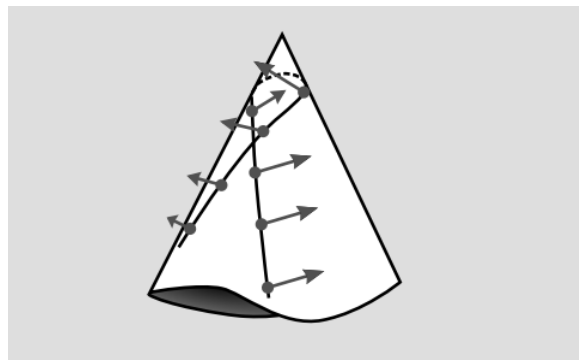
Figure 5.10.2 - The cone has zero intrinsic curvature everywhere except at its tip. An observer who never visits the tip can nevertheless detect its existence, because parallel transport around a path that encloses the tip causes a vector to change its direction.

In the electromagnetic example, integrating **A** around a closed loop reveals, via Stokes' theorem, the existence of a magnetic flux through the loop, even though the magnetic field is zero at every location where **A** has to be sampled. In the relativistic example, integrating Γ around a closed loop shows that there is curvature inside the loop, even though the curvature is zero at all the places where Γ has to be sampled.

The fact that Γ is a gauge field, and therefore not locally observable, is simply a fancy way of expressing the ideas introduced on in section 5.6, that due to the equivalence principle, the gravitational field in general relativity is not locally observable. This nonobservability is local because the equivalence principle is a statement about local Lorentz frames. The example in Figure 5.9.2 is non-local.

> **Example 11: Geodetic effect and structure of the source**
>
> - In Section 5.5, we estimated the geodetic effect on Gravity Probe B and found a result that was only off by a factor of 3π. The mathematically pure form of the 3π suggests that the geodetic effect is insensitive to the distribution of mass inside the earth. Why should this be so?
> - The change in a vector upon parallel transporting it around a closed loop can be expressed in terms of either (1) the area integral of the curvature within the loop or (2) the line integral of the Christoffel symbol (essentially the gravitational field) on the loop itself. Although I expressed the estimate as 1, it would have been equally valid to use 2. By Newton's shell

theorem, the gravitational field is not sensitive to anything about its mass distribution other than its near spherical symmetry. The earth spins, and this does affect the stress-energy tensor, but since the velocity with which it spins is everywhere much smaller than c, the resulting effect, called *frame dragging*, is much smaller.

References

[13] http://arxiv.org/abs/0903.2085

[14] http://arxiv.org/abs/0903.2085

This page titled 5.10: From Metric to Curvature is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Benjamin Crowell via source content that was edited to the style and standards of the LibreTexts platform.

5.11: Manifolds (Part 1)

This section can be omitted on a first reading.

Why We Need Manifolds

General relativity doesn't assume a predefined background metric, and this creates a chicken-and-egg problem. We want to define a metric on some space, but how do we even specify the set of points that make up that space? The usual way to define a set of points would be by their coordinates. For example, in two dimensions we could define the space as the set of all ordered pairs of real numbers (x, y). But this doesn't work in general relativity, because space is not guaranteed to have this structure. For example, in the classic 1979 computer game Asteroids, space "wraps around," so that if your spaceship flies off the right edge of the screen, it reappears on the left, and similarly at the top and bottom (Figure 5.11.1).
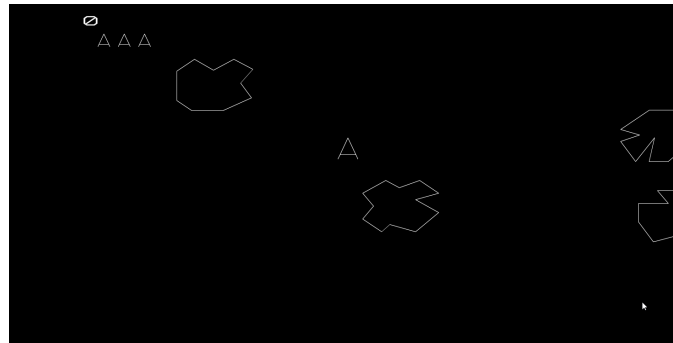
Figure 5.11.1: In Asteroids, space "wraps around."

Even before we impose a metric on this space, it has topological properties that differ from those of the Euclidean plane. By "topological" we mean properties that are preserved if the space is thought of as a sheet of rubber that can be stretched in any way, but not cut or glued back together. Topologically, the space in Asteroids is equivalent to a torus (surface of a doughnut), but not to the Euclidean plane.
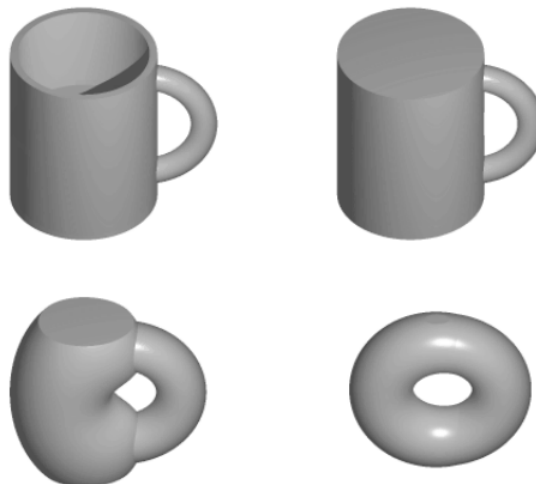
Figure 5.11.2: A coffee cup is topologically equivalent to a torus.

Another useful example is the surface of a sphere. In [example 10](#), we calculated $\Gamma^{\theta}_{\phi\phi}$. A similar calculation gives $\Gamma^{\theta}_{\phi\phi} = \cot \frac{\theta}{R}$. Now consider what happens as we drive our dogsled north along the line of longitude $\phi = 0$, cross the north pole at $\theta = 0$, and continue along the same geodesic. As we cross the pole, our longitude changes discontinuously from 0 to π. Consulting the geodesic equation, we see that this happens because $\Gamma^{\theta}_{\phi\phi}$ blows up at $\theta = 0$. Of course nothing really special happens at the pole. The bad behavior isn't the fault of the sphere, it's the fault of the (θ, ϕ) coordinates we've chosen, that happen to misbehave at the pole. Unfortunately, it is impossible to define a pair of coordinates on a two-sphere without having them misbehave somewhere.

(This follows from Brouwer's famous 1912 "Hairy ball theorem," which states that it is impossible to comb the hair on a sphere without creating a cowlick somewhere.)

Topological Definition of a Manifold

This motivates us to try to define a "bare-bones" geometrical space in which there is no predefined metric or even any predefined set of coordinates.

There is a general notion of a topological space, which is too general for our purposes. In such a space, the only structure we are guaranteed is that certain sets are defined as "open," in the same sense that an interval like $0 < x < 1$ is called "open." A point in an open set can be moved in any direction without leaving the set. An open set is essentially a set without a boundary, for in a set like $0 \leq x \leq 1$, the boundary points 0 and 1 can only be moved in one direction without taking them outside.
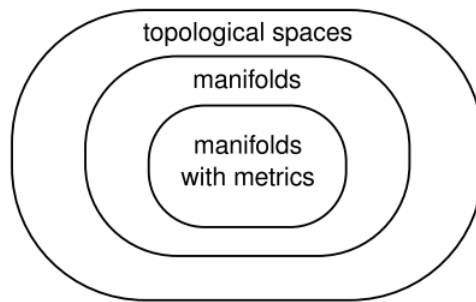
Figure 5.11.3: General relativity does not assume a predefined background metric. Therefore all we can really know before we calculate anything is that we're working on a manifold, without a metric imposed on it.

A topological space is too general for us because it can include spaces like fractals, infinite-dimensional spaces, and spaces that have different numbers of dimensions in different regions. It is nevertheless useful to recognize certain concepts that can be defined using only the generic apparatus of a topological space, so that we know they do not depend in any way on the presence of a metric. An open set surrounding a point is called a neighborhood of that point. In a topological space we have a notion of getting arbitrarily close to a certain point, which means to take smaller and smaller neighborhoods, each of which is a subset of the last. But since there is no metric, we do not have any concept of comparing distances of distant points, e.g., that P is closer to Q than R is to S. A continuous function is a purely topological idea; a continuous function is one such that for any open subset U of its range, the set V of points in its domain that are mapped to points in U is also open. Although some definitions of continuous functions talk about real numbers like ϵ and δ, the notion of continuity doesn't depend on the existence of any structure such as the real number system. A homeomorphism is a function that is invertible and continuous in both directions. Homeomorphisms formalize the informal notion of "rubber-sheet geometry without cutting or gluing." If a homeomorphism exists between two topological spaces, we say that they are homeomorphic; they have the same structure and are in some sense the same space.

The more specific type of topological space we want is called a manifold. Without attempting any high level of mathematical rigor, we define an n-dimensional manifold M according to the following informal principles:[16]

> **Note**
>
> For those with knowledge of topology, these can be formalized a little more: we want a completely normal, second-countable, locally connected topological space that has Lebesgue covering dimension n, is a homogeneous space under its own homeomorphism group, and is a complete uniform space. I don't know whether this is sufficient to characterize a manifold completely, but it suffices to rule out all the counterexamples of which I know.

- **M1** Dimension: M's dimension is n.
- **M2** Homogeneity: No point has any locally definable property that distinguishes it from any other point.
- **M3** Completeness: M is complete, in the sense that specifying an arbitrarily small neighborhood gives a unique definition of a point.

Example 12: Lines

The set of all real numbers is a 1-manifold. Similarly, any line with the properties specified in Euclid's *Elements* is a 1-manifold. All such lines are homeomorphic to one another, and we can therefore speak of "the line."

Example 13: a circle

A circle (not including its interior) is a 1-manifold, and it is not homeomorphic to the line. To see this, note that deleting a point from a circle leaves it in one connected piece, but deleting a point from a line makes two. Here we use the fact that a homeomorphism is guaranteed to preserve "rubber-sheet" properties like the number of pieces.

Example 14: No changes of dimension

A "lollipop" formed by gluing an open 2-circle (i.e., a circle not including its boundary) to an open line segment is not a manifold, because there is no n for which it satisfies M1.

It also violates M2, because points in this set fall into three distinct classes: classes that live in 2-dimensional neighborhoods, those that live in 1-dimensional neighborhoods, and the point where the line segment intersects the boundary of the circle.

Example 15: No manifolds made from the rational numbers

The rational numbers are not a manifold, because specifying an arbitrarily small neighborhood around $\sqrt{2}$ excludes every rational number, violating M3.

Similarly, the rational plane defined by rational-number coordinate pairs (x, y) is not a 2-manifold. It's good that we've excluded this space, because it has the unphysical property that curves can cross without having a point in common. For example, the curve $y = x^2$ crosses from one side of the line $y = 2$ to the other, but never intersects it. This is physically undesirable because it doesn't match up with what we have in mind when we talk about collisions between particles as intersections of their world-lines, or when we say that electric field lines aren't supposed to intersect.

Example 16: no boundary

The open half-plane $y > 0$ in the Cartesian plane is a 2-manifold. The closed half-plane $y \geq 0$ is not, because it violates M2; the boundary points have different properties than the ones on the interior.

Example 17: Disconnected manifolds

Two nonintersecting lines are a 1-manifold. Physically, disconnected manifolds of this type would represent a universe in which an observer in one region would never be able to find out about the existence of the other region.

Example 18: No bad glue jobs

Hold your hands like you're pretending you know karate, and then use one hand to karate-chop the other. Suppose we want to join two open half-planes in this way. As long as they're separate, then we have a perfectly legitimate disconnected manifold. But if we want to join them by adding the point P where their boundaries coincide, then we violate M2, because this point has special properties not possessed by any others. An example of such a property is that there exist points Q and R such that every continuous curve joining them passes through P. (Cf. problem 5)

Hausdorff Property

Pioneering topologist Felix Hausdorff defined the following property of a topological space.

Definition: Hausdorff property

Given any two points, it is possible to find disjoint neighborhoods of them.

A joke/mnemonic, which probably works best for people with a certain type of British accent, is that in a Hausdorff space, any two points can be "housed off" from one another inside their own nonintersecting open sets. The notion appeals strongly to our intuitive

ideas about how space and time behave, and the standard definition of a manifold implies that it is Hausdorff. When we model Minkowski space using real-number coordinates, it is Hausdorff. Since the equivalence principle says that spacetime is locally Minkowski, we could also say that it implies spacetime is Hausdorff. However, general relativity allows spacetime to behave badly in cases such as singularities, so it is imaginable that our universe contains points that violate the Hausdorff property. There are interesting and physically well-motivated spacetimes, such as some versions of the Taub-NUT space, that are non-Hausdorff. Since we have no empirical data on the behavior of spacetime under the most extreme conditions, we cannot say whether spacetime is really Hausdorff. One should maintain some skepticism about whether such an idealized category is even meaningful in science, since it refers to phenomena at arbitrarily small scales, whereas theories and measurements are limited in the scales they can deal with. A good discussion of the Hausdorff property as applied to relativity is given by Earman.[17]

Example 19: Branching universes

Figure $5.11.4$ shows spacetime diagrams of 1 + 1-dimensional universes that branch like a tree. These are meant to be pictures of classical general relativity, although some of the strongest motivation for considering such possibilities comes from attempts to construct a theory of quantum gravity. In such theories, it is commonly expected that spacetime will have a structure at the Planck scale that is a kind of "quantum foam."
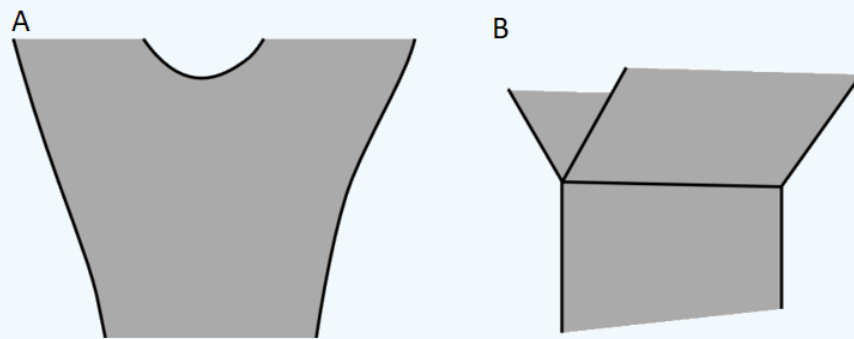
Figure 5.11.4: *Spacetime diagrams of 1 + 1-dimensional universes.*

The example in Figure $5.11.4A$ is a manifold, and is Hausdorff. This is an example of topology change, meaning that the spacelike section at one time has a different topology than the section at another.[18] Although such a branching can occur without the existence of any singularities, theorems by Tipler and Geroch show that other types of misbehavior must occur, including causality violations and the need for forms of matter that violate energy conditions.

Figure $5.11.4B$ is qualitatively different. Here we have formed a spacetime by gluing together three pieces. No curvature is implied; these could be three pieces of Minkowski space. The spacetime is not a manifold, since the points at the join have different local properties than points elsewhere. The machinery of general relativity breaks down in a case like this, but for example we could imagine that a geodesic in this spacetime could fork off into two different geodesics after the split.

Yet a third possibility is to reinterpret Figure $5.11.4B$ so that there are two different copies of the seam. For example, we could let the portion of the diagram extending into the past be represented by points with t < 0, while the two branches continuing into the future could each have t ≥ 0, so that for a given x we would have two different events with coordinates (t = 0, x). It would not be possible to put these two points into disjoint neighborhoods, so this version of the space is not Hausdorff.

[18] For a recent treatment, see Borde, 1994, "Topology change in classical general relativity," arxiv.org/abs/gr-qc/9406053.

References

[17] John Earman, 'Pruning some branches from "branching spacetimes",' pitt.edu/~jearman/Earman2008a.pdf

5.12: Manifolds (Part 2)

Local-coordinate Definition of a Manifold

An alternative way of characterizing an n-manifold is as an object that can locally be described by n real coordinates. That is, any sufficiently small neighborhood is homeomorphic to an open set in the space of real-valued n-tuples of the form $(x_1, x_2, \ldots, x_n)$. For example, a closed half-plane is not a 2-manifold because no neighborhood of a point on its edge is homeomorphic to any open set in the Cartesian plane.

Exercise 5.12.1

Self-check: Verify that this alternative definition of a manifold gives the same answers as M1-M3 in all the examples above.

Roughly speaking, the equivalence of the two definitions occurs because we're using n real numbers as coordinates for the dimensions specified by M1, and the real numbers are the unique number system that has the usual arithmetic operations, is ordered, and is complete in the sense of M3.

As usual when we say that something is "local," a question arises as to how local is local enough. The language in the definition above about "any sufficiently small neighborhood" is logically akin to the Weierstrass $\epsilon - \delta$ approach: if Alice gives Bob a manifold and a point on a manifold, Bob can always find some neighborhood around that point that is compatible with coordinates, but it may be an extremely small neighborhood.

Example 20: Coordinates on a circle

If we are to define coordinates on a circle, they should be continuous functions. The angle ϕ about the center therefore doesn't quite work as a global coordinate, because it has a discontinuity where $\phi = 0$ is identified with $\phi = 2\pi$. We can get around this by using different coordinates in different regions, as is guaranteed to be possible by the local-coordinate definition of a manifold. For example, we can cover the circle with two open sets, one on the left and one on the right. The left one, L, is defined by deleting only the $\phi = 0$ point from the circle. The right one, R, is defined by deleting only the one at $\phi = \pi$. On L, we use coordinates $0 < \phi_L < 2\pi$, which are always a continuous function from L to the real numbers. On R, we use $-\pi < \phi_R < \pi$.

In examples like this one, the sets like L and R are referred to as patches. We require that the coordinate maps on the different patches match up smoothly. In this example, we would like all four of the following functions, known as transition maps, to be continuous:

- ϕ_L as a function of ϕ_R on the domain $0 < \phi_R < \pi$
- ϕ_L as a function of ϕ_R on the domain $-\pi < \phi_R < 0$
- ϕ_R as a function of ϕ_L on the domain $0 < \phi_L < \pi$
- ϕ_R as a function of ϕ_LL on the domain $\pi < \phi_L < 2\pi$

The local-coordinate definition only states that a manifold can be coordinatized. That is, the functions that define the coordinate maps are not part of the definition of the manifold, so, for example, if two people define coordinates patches on the unit circle in different ways, they are still talking about exactly the same manifold.

Example 21: Open line segment homeomorphic to a line

Let L be an open line segment, such as the open interval (0, 1). L is homeomorphic to a line, because we can map (0, 1) to the real line through the function f(x) = $\tan(\pi x - \frac{\pi}{2})$.

Example 22: Closed line segment not homeomorphic to a line

A closed line segment (which is not a manifold) is not homeomorphic to a line. If we map it to a line, then the endpoints have to go to two special points A and B. There is then no way for the mapping to visit the points exterior to the interval [A, B] without visiting A and B more than once.

If the interior of a circle could be mapped by a homeomorphism f to an open line segment, then consider what would happen if we took a closed curve lying inside the circle and found its image. By the intermediate value theorem, f would not be one-to-one, but this is a contradiction since f was assumed to be a homeomorphism. This is an example of a more general fact that homeomorphism preserves the dimensionality of a manifold.
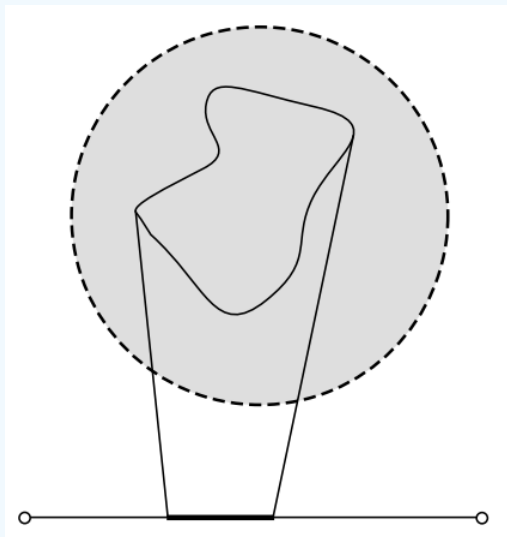
Figure 5.12.5

Differentiable Manifolds

A differentiable manifold means a manifold with enough extra structure so you can do calculus on it, but this extra structure doesn't necessarily include anything as fancy as a metric. As a concrete example, suppose that in a 1 + 1-dimensional Galilean universe, observer Alice constructs a global coordinate system (t, x). Her spacetime is clearly a manifold, based on the local-coordinate definition, and this is true even though Galilean spacetime doesn't have a metric. Meanwhile, observer Bob constructs his own coordinate system (t', x'). But something disturbing happens when Alice constructs the transition map from Bob's coordinate grid to hers. As shown in figure f, Bob's grid has a kink in it. "Bob," says Alice, "something is wrong with your coordinate system. I hypothesize that at a certain time, which we can call t = 0, an invisible giant struck your body with an invisible croquet mallet and suddenly changed your state of motion." "No way, Alice," Bob answers. "I didn't feel anything happen at t = 0. I think you're the one who got whacked."
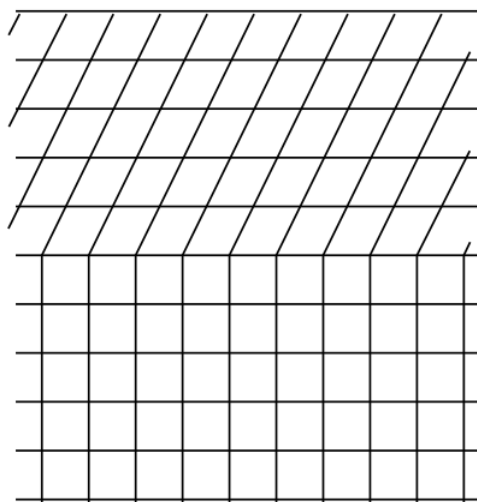
Figure 5.12.6 - *"Bob, your manifold isn't smooth!"*

By a differentiable manifold we mean one in which this sort of controversy never happens. The manifold comes with an a collection of local coordinate systems, called *charts,* and wherever these charts overlap, the transition map is differentiable. Every coordinate is a differentiable function of every other coordinate. In fact, we will assume for convenience that not just the first derivative but derivatives of all orders are defined. This makes our manifold not just a differentiable manifold but a *smooth manifold*. This definition sounds coordinate-dependent, but it isn't. Our collection of charts (called an *atlas*) can contain infinitely many possible coordinate systems; we can even specify that it contains *all* possible coordinate systems that could be obtained from one another by any diffeomorphism.

The Tangent Space

We now formalize the intuitive notion of a tangent vector (section 3.1), following Nowik and Katz.[19] Let M be an n-dimensional smooth manifold, so that locally it looks like Euclidean space, describable by real-number coordinates x, y, . . .We now enhance M to form a new topological space, in which the coordinates can include not only real numbers, but numbers that differ infinitesimally from reals, as outlined in example 3. From now on when we say things like "the manifold," we mean this enhanced version.[20] Fix some infinitesimal number ϵ for once and for all, and define the notation x = $O(\epsilon)$ to mean that $\frac{x}{\epsilon}$ is not infinite.[21]

> **Note**
>
> It is possible to define a different and larger enhancement, called $^*M$, that would include points with infinitely large coordinates. For example, suppose we have a coordinate patch with bounds on the coordinates that can be written down using inequalities, t > 0, $0 \le \theta \le \frac{\pi}{4}$, . . . Then $^*M$ would contain any finite, infinitesimal, and infinite values of (t, θ, . . .) satisfying these inequalities, and this would include infinite values of t. We will not do this here, because the inclusion of idealized points at infinity is more useful in relativity if we do it using a different approach, discussed in section 7.3.

> **Note**
>
> An equivalent and manifestly coordinate-independent definition is that for every smooth real function in a neighborhood of the points, the function differs at these points by an amount that is $O(\epsilon)$.

Points in the manifold are considered *close* if the Euclidean distance between them in coordinate space is $O(\epsilon)$. This definition sounds coordinate-dependent, but isn't, and sounds like it's assuming an actual Euclidean metric, but isn't.[22] Define a *prevector* at point P as a pair (P,Q) of points that are close, Figure 5.10.7 (1). Define prevectors to be equivalent if the difference between them is infinitesimal even compared to ϵ.
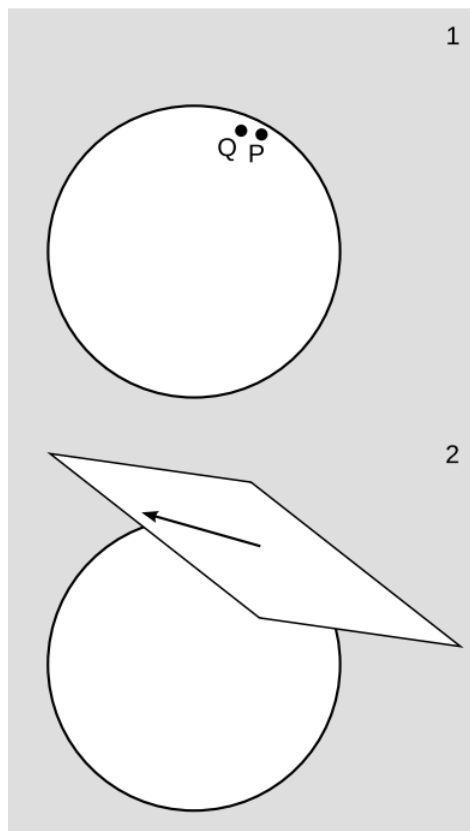
Figure 5.12.7 - *1. A tangent vector can be thought of as an infinitesimal displacement. 2. For a sphere embedded in three-space, the space of tangent vectors is visualized as a plane tangent to the sphere at a certain point.*

> **Definition**
>
> A tangent vector at point P is the set of all prevectors at P that are equivalent to a particular prevector at P.

The *tangent space* T_P is the set of all tangent vectors at P. The tangent space has the structure of a vector space over the reals simply by using the coordinate differences to define the vector-space operations, just as we would do if (P, Q) meant an arrow extending from P to Q, as in freshman physics.

In practice, we don't really care about the details of the construction of the tangent space, and different people don't even have to use the same construction. All we care about is that the tangent space has a certain structure. In particular, it has n dimensions, as we would expect intuitively. Since we're going to forget the details of the construction, it doesn't matter that we've made all tangent vectors infinitesimal by definition. The vector space's internal structure only has to do with how big the vectors are *compared to each other*. (If we wanted to, we could scale up all the tangent vectors by a factor of $\frac{1}{\epsilon}$.) This justifies the visualization in Figure 5.10.7 (2).

Actually it's not quite true that we only care about the tangent space's internal structure, because then we could have avoided the fancy definition and simply used the ordinary vector space consisting of n-tuples of real numbers. The fancy definition is needed because it ties the tangent space in a natural way to the structure of the manifold at a particular point. Therefore it will allow us (1) to define parallel transport, which brings a vector from one tangent space to another, and (2) to define components of vectors in a particular coordinate system.

For an alternative definition of the tangent space, see ch. 2 of Carroll.[23] Briefly, this involves taking a tangent vector to be something that behaves like a directional derivative. In particular, a partial derivative with respect to a coordinate such as $\frac{\partial}{\partial x}$ qualifies as a tangent vector, which we think of as pointing in the x direction. The set of such coordinate derivatives forms a basis for the tangent space and gives a convenient way of notating tangent vectors. We will find this notation convenient in section 7.1.

References

[19] "Differential geometry via infinitesimal displacements," arxiv.org/abs/1405.0984

[23] Lecture Notes on General Relativity, http://ned.ipac.caltech.edu/level5/March01/Carroll3/Carroll_contents.html.

5.13: Units in General Relativity

Analyzing units, also known as dimensional analysis, is one of the first things we learn in freshman physics. It's a useful way of checking our math, and it seems as though it ought to be straightforward to extend the technique to relativity. It certainly can be done, but it isn't quite as trivial as might be imagined, and it leads to some surprising new physical ideas.

One of our most common jobs is to change from one set of units to another, but in relativity it becomes nontrivial to define what we mean by the notion that our units of measurement change or don't change. We could, e.g., appeal to an atomic standard, but Dicke[24] points out that this could be problematic. Imagine, he says, that

> *... you are told by a space traveller that a hydrogen atom on Sirius has the same diameter as one on the earth. A few moments' thought will convince you that the statement is either a definition or else meaningless.*

To start with, we note that *abstract index notation* is more convenient than *concrete index notation* for these purposes. Concrete index notation assigns different units to different components of a tensor if we use coordinates, such as spherical coordinates (t, r, θ, ϕ), that don't all have units of length. In abstract index notation, a symbol like v^i stands for the whole vector, not for one of its components.

In concrete index notation, it also doesn't necessarily make sense to talk about rescaling. E.g., for polar coordinates in the Euclidean plane, the transformation (r, θ) → (2r, 2θ) doesn't have any interesting interpretation, and can't even be applied globally. In abstract index notation, we can say $v^i \to 2v^i$, and this simply means that the vector v^i has been scaled up by a factor of 2.

Since abstract index notation does not even offer us a notation for components, if we want to apply dimensional analysis we must define a system in which units are attributed to a tensor as a whole. Suppose we write down the abstract-index form of the equation for proper time:

$$ds^2 = g_{ab} dx^a dx^a \tag{5.13.1}$$

In abstract index notation, dx^a doesn't mean an infinitesimal change in a particular coordinate, it means an infinitesimal displacement vector.[25] This equation has one quantity on the left and three factors on the right. Suppose we assign these parts of the equation units $[ds] = L^\sigma$, $[g_{ab}] = L^{2\gamma}$, and $[dx^a] = [dx^b] = L^\xi$, where square brackets mean "the units of" and L stands for units of length. We then have $\sigma = \gamma + \xi$. Due to the ambiguities referred to above, we can pick any values we like for these three constants, as long as they obey this rule. I find $(\sigma, \gamma, \xi) = (1, 0, 1)$ to be natural and convenient, but Dicke, in the above-referenced paper, likes (1, 1, 0), while the mathematician Terry Tao advocates (0, $\mp$1, $\pm$1).

Suppose we raise and lower indices to form a tensor with r upper indices and s lower indices We refer to this as a tensor of rank (r, s). (We don't count contracted indices, e.g., $u^a v_a$ is a rank-(0, 0) scalar.) Since the metric is the tool we use for raising and lowering indices, and the units of the lower-index form of the metric are $L^{2\gamma}$, it follows that the units vary in proportion to $L^{\gamma(s-r)}$. In general, you can assign a physical quantity units L^u that are a product of two factors, a "kinematical" or purely geometrical factor L^k, where $k = \gamma(s - r)$, and a dynamical factor L^d . . ., which can depend on what kind of quantity it is, and where the . . . indicates that if your system of units has more than just one base unit, those can be in there as well. Dicke uses units with $\hbar$ = c = 1, for example, so there is only one base unit, and mass has units of inverse length and d_{mass} = −1. In general relativity it would be more common to use units in which G = c = 1, which instead give d_{mass} = +1.

> **Note**
>
> For a modern and rigorous development of differential geometry along these lines, see Nowik and Katz, arxiv.org/abs/1405.0984.

> **Example 24: The units of momentum**
>
> Consider the equation
>
> $$p^a = mv^a \tag{5.13.2}$$
>
> for the momentum of a material particle. Suppose we use specialrelativistic units in which c = 1, but because gravity isn't incorporated into the theory, G plays no special role, and it is natural to use a system of units in which there is a base unit of

mass M.

The kinematic units check out, because $k_p = k_m + k_v$:

$$\gamma(-1) = \gamma(0) + \gamma(-1) \qquad (5.13.3)$$

This is merely a matter of counting indices, and was guaranteed to check out as long as the indices were written in a grammatical way on both sides of the equation. What this check is essentially telling us is that if we were to establish Minkowski coordinates in a neighborhood of some point, and do a change of coordinates (t, x, y, z) → (αt, αx, αy, αz), then the quantities on both sides of the equation would vary under the tensor transformation laws according to the same exponent of α. For example, if we changed from meters to centimeters, the equation would still remain valid.

For the dynamical units, suppose that we use $(\sigma, \gamma, \xi) = (1, 0, 1)$, so that an infinitesimal displacement dx^a has units of length L, as does proper time ds. These two quantities are purely kinematic, so we don't assign them any dynamical units, and therefore the velocity vector $v^a = \frac{dx^a}{ds}$ also has no dynamical units. Our choice of a system of units gives [m] = M. We require that the equation $p^a = mv^a$ have dynamical units that check out, so:

$$M = 1 \cdot M \qquad (5.13.4)$$

We must also assign units of mass to the momentum.

A system almost identical to this one, but with different terminology, is given by Schouten.[26]

For practical purposes in checking the units of an equation, we can see from example 24 that worrying about the kinematic units is a waste of time as long as we have checked that the indices are grammatical. We can therefore give a simplified method that suffices for checking the units of any equation in abstract index notation.

1. We assign a tensor the same units that one of its concrete components *would* have if we were to adopt (local) Minkowski coordinates, in the system with $(\sigma, \gamma, \xi) = (1, 0, 1)$. These are the units we would automatically have imputed to it after learning special relativity but before learning about tensors or fancy coordinate transformations. Since $\gamma = 0$, the positions of the indices do not affect the result.
2. The units of a sum are the same as the units of the terms.
3. The units of a tensor product are the product of the units of the factors.

Our splitting of units into kinematic and dynamical parts can be understood as arising naturally from the following geometrical and physical considerations. In section 3.2, we introduced the notion of a *connection*, which is a rule that relates tensors living in one local region of spacetime to those in another region, depending on the path used for parallel transport. The connection is embodied concretely in the Christoffel symbols, and we need it in order to define sensible derivatives of vectors, because otherwise we lack the information needed in order to tell whether a vector is in fact constant, and only changing its components due to the way the coordinate system is defined. The connection and the metric embody a lot of the same geometrical information. If we know the metric, we can always find the connection (section 5.9).

We might then naturally ask whether it is possible to go in the other direction. Given the connection, can we find the metric? But this is clearly not true, because the connection doesn't carry any information about units of measurement, while the metric does. In fact, if the metric g results in a certain connection Γ, then so will the metric $\Omega^2 g$, where Ω is a real constant.[27] One way of thinking about the transformation g → $\Omega^2 g$ is that in the expression $ds^2 = g_{ab} dx^a dx^a$ for proper time, we scale up any clock reading s by a factor of Ω. This helps to explain Dicke's preference for the convention $(\sigma, \gamma, \xi) = (1, 1, 0)$, according to which the units are attributed to ds and g, while vectors are considered to be unitless. A further advantage of this system is that it can be adapted to concrete index notation, because we simply declare coordinates to be unitless names for points.

> **Note**
>
> If we multiplied g by a negative constant, then we would change the signature, e.g., from +−−− to −+++. Changing the signature would be particularly goofy in the context of Riemannian geometry, where it is customary to have a positive-definite metric.

The following table summarizes the factors by which various quantities change under rescaling of the lower-index metric and rescaling of local Minkowski coordinates x^μ. As above, r is the number of upper indices and s the number of lower indices. Entries

in lighter text follow from the more general rule. A curvature monomial of order p is an expression formed from the multiplication of p curvature tensors, possibly with contracted indices.

	$g_{ab} \to \Omega^2 g_{ab}$ (5.13.5)	$x^\mu \to \alpha x^\mu$ (5.13.6)
g	Ω^{s-r} (5.13.7)	α^{s-r} (5.13.8)
tensor density of rank (r, s) and weight w		α^{2w+r-s} (5.13.9)
Γ^a_{bc} (5.13.10)	1	α^{-1} (5.13.11)
curvature monomial of order p	Ω^{s-r-2p} (5.13.12)	α^{s-r} (5.13.13)

It makes sense that rescaling the metric doesn't change the Christoffel symbols, because it doesn't change the connection or the coordinates, and therefore shouldn't change the geodesic equation. Verifying the other entries in the table is a good exercise.

Example 25: A change of signature

Suppose that we change the signature of a metric from + − −− to − + ++ or vice versa. Although the notation Ω^2 was intended to imply that the signature of the metric would not be changed, nothing goes wrong in the logic if we take $\Omega^2 = -1$. According to the table, the lower-index form of the metric, with (r, s) = (0, 2) changes by a factor of −1, which is what we set out to do. A curvature polynomial of order p changes by a factor of $(-1)^p$. As a specific example, a cosmological model dominated by the cosmological constant (section 8.2) has Ricci scalar R = −12Λ in the + − −− signature used in this book, but R = +12Λ in the − + ++ signature.

Example 26: Curvature scalars for the Godel metric

The Ricci scalar R = R^a_a is a curvature monomial of order 1. Because it is a relativistic scalar, its value is invariant under a change of coordinates. A scalar constructed in this way from a curvature tensor is called a curvature scalar. In the system described above, it is a curvature monomial of order 1, and it is a tensor of rank (0, 0). It is a pure tensor, i.e., it is a tensor density in only the trivial sense, having weight w = 0.

The Kretschmann invariant K = $R^{abcd}R_{abcd}$, discussed in more detail in section 6.3, is a curvature monomial of order 2, with properties that are otherwise similar to the ones listed above for the Ricci scalar.

To have a specific example to talk about, let us consider the metric

$$ds^2 = dt^2 - dx^2 - dy^2 + \frac{1}{2}e^{2x}dz^2 - 2e^x dzdt. \qquad (5.13.14)$$

This is the historically and philosophically important Gödel metric, discussed in section 8.2. A calculation using Maxima gives R = 1 (+ − −− signature) and K = 3. (The fact that both of these are constant shows that the spacetime is highly symmetric, although this is not manifest when the metric is expressed in these coordinates.) Suppose that we recalibrate our clocks to use different units, changing the metric above according to $ds^2 \to \Omega^2 ds^2$. Then application of the rules given in the table tells us that R = Ω^{-2} and K = $3\Omega^{-4}$.

To round out our discussion of this approach, we state more precisely the relationship between the metric and the connection. Given a metric, there is a unique torsion-free connection. Given a torsion-free connection, there may or may not exist a metric that gives rise to that connection. If such a metric does exist, then except in exceptional cases that metric is unique up to a nonzero multiplicative constant. The reason for the uniqueness of the metric up to a constant factor is as follows. Suppose we fix the metric at one point on our manifold. Then by using the connection we can paralleltransport the metric tensor to other points on the manifold, so that defining it at one point has the effect of defining it everywhere. But there may be a lack of consistency, because parallel transport is path-dependent. In particular, if we transport the metric around a closed loop, we want to recover the original

metric. This consistency requirement is usually enough to rule out any freedom in defining the metric beyond a global scaling factor. A more complete treatment of this problem is given by Schmidt.[28]

An interesting exceptional case is flat spacetime. Because there is no curvature, parallel transport around a closed loop never changes the metric, so the consistency requirement is automatically satisfied, and we our freedom in choosing a metric is greater than just the ability to scale by a constant. In particular, some authors choose not to use natural units, so that instead of g = diag(1, −1, −1, −1) in Cartesian coordinates, one has g = diag(c^2, −1, −1, −1). In an approach where a change of units is represented by a change of coordinates, this change in the metric could be represented by (t, x, y, z) → ($\frac{t}{c}$, x, y, z). But in the convention followed by Dicke, we would take the coordinates to be immutable labels for points, and these would actually be physically different metrics, with different light cones.

A similar example in a Riemannian context is the Euclidean plane, in which the (trivial) connection is consistent any metric of the form given in example 9.

Finally, we note that it can be of interest to generalize the transformation g → Ω^2 g so that Ω can vary from point to point. This is called a conformal transformation. Conformal transformations can be used for a variety of purposes, including nontrivial physics (as in the Dicke paper) and techniques for visualization (section 7.3).

References

[24] "Mach's principle and invariance under transformation of units," Phys Rev 125 (1962) 2163

[26] Tensor Analysis for Physicists, ch. VI

[28] projecteuclid.org/euclid.cmp/1103858479

5.E: Curvature (Exercises)

1. Example 6 discussed some examples in electrostatics where the charge density on the surface of a conductor depends on the Gaussian curvature, when the curvature is positive. In the case of a knife-edge formed by two half-planes at an exterior angle $\beta > \pi$, there is a standard result[29] that the charge density at the edge blows up to infinity as $R^{\frac{\pi}{\beta-1}}$. Does this match up with the hypothesis that Gaussian curvature determines the charge density?

2. Show that for polar coordinates in a Euclidean plane, $\Gamma^r_{\phi\phi} = -r$ and $\Gamma^\phi_{r\phi} = \frac{1}{r}$.

3. In 1+1 dimensions, let the metric be $ds^2 = \frac{1}{t}\,dt^2 - t\,d\theta^2$, where θ is an angle running around the circle. Calculate all the nonvanishing Christoffel symbols by hand. These will be used in example 4 where we investigate some further properties of this interesting spacetime.

4. Partial derivatives commute with partial derivatives. Covariant derivatives don't commute with covariant derivatives. Do covariant derivatives commute with partial derivatives?

5. Show that if the differential equation is satisfied for one affine parameter λ, then it is also satisfied for any other affine parameter $\lambda' = a\lambda + b$, where a and b are constants.

6. Equation [2] gives a flat-spacetime metric in rotating polar coordinates. (a) Verify by explicit computation that this metric represents a flat spacetime. (b) Re-express the metric in rotating Cartesian coordinates, and check your answer by verifying that the Riemann tensor vanishes.

7. The purpose of this problem is to explore the difficulties inherent in finding anything in general relativity that represents a uniform gravitational field g. In example 11, we found, based on elementary arguments about the equivalence principle and photons in elevators, that gravitational time dilation must be given by e^Φ, where Φ = gz is the gravitational potential. This results in a metric

$$ds^2 = e^{2gz}dt^2 - dz^2. \tag{[1]}$$

On the other hand, example 19 derived the metric

$$ds^2 = (1+gz)^2 dt^2 - dz^2. \tag{[2]}$$

by transforming from a Lorentz frame to a frame whose origin moves with constant proper acceleration g. (These are known as Rindler coordinates.) Prove the following facts. None of the calculations are so complex as to require symbolic math software, so you might want to perform them by hand first, and then check yourself on a computer.

   a. The metrics [1] and [2] are approximately consistent with one another for z near 0.
   b. When a test particle is released from rest in either of these metrics, its initial proper acceleration is g.
   c. The two metrics are not exactly equivalent to one another under any change of coordinates.
   d. Both spacetimes are uniform in the sense that the curvature is constant. (In both cases, this can be proved without an explicit computation of the Riemann tensor.)

> **Note**
>
> The incompatibility between [1] and [2] can be interpreted as showing that general relativity does not admit any spacetime that has all the global properties we would like for a uniform gravitational field. This is related to Bell's spaceship paradox (example 15). Some further properties of the metric [1] are analyzed in section 7.5.

8. In a topological space T, the complement of a subset U is defined as the set of all points in T that are not members of U. A set whose complement is open is referred to as closed. On the real line, give (a) one example of a closed set and (b) one example of a set that is neither open nor closed. (c) Give an example of an inequality that defines an open set on the rational number line, but a closed set on the real line.

9. Prove that a double cone (e.g., the surface r = z in cylindrical coordinates) is not a manifold.

10. Prove that a torus is a manifold.

11. Prove that a sphere is not homeomorphic to a torus.

12. Curvature on a Riemannian space in 2 dimensions is a topic that goes back to Gauss and has a simple interpretation: the only intrinsic measure of curvature is a single number, the Gaussian curvature. What about 1+1 dimensions? The simplest metrics I can think of are of the form $ds^2 = dt^2 - f(t)dx^2$. (Something like $ds^2 = f(t)dt^2 - dx^2$ is obviously equivalent to Minkowski space under a change of coordinates, while $ds^2 = f(x)dt^2 - dx^2$ is the same as the original example except that we've swapped x and t.)

Playing around with simple examples, one stumbles across the seemingly mysterious fact that the metric $ds^2 = dt^2 - t^2dx^2$ is flat, while $ds^2 = dt^2 - tdx^2$ is not. This seems to require some simple explanation. Consider the metric $ds^2 = dt^2 - t\,pdx^2$.

   a. Calculate the Christoffel symbols by hand.

   b. Use a computer algebra system such as Maxima to show that the Ricci tensor vanishes only when p = 2.

> **Note**
>
> The explanation is that in the case p = 2, the x coordinate is expanding in proportion to the t coordinate. This can be interpreted as a situation in which our length scale is defined by a lattice of test particles that expands inertially. Since their motion is inertial, no gravitational fields are required in order to explain the observed change in the length scale; cf. the Milne universe.

References

[29] Jackson, *Classical Electrodynamics*

CHAPTER OVERVIEW

6: Vacuum Solutions

In this chapter we investigate general relativity in regions of space that have no matter to act as sources of the gravitational field. We will *not*, however, limit ourselves to calculating spacetimes in cases in which the entire *universe* has no matter. For example, we will be able to calculate general-relativistic effects in the region surrounding the earth, including a full calculation of the geodetic effect, which was estimated in Section 5.5 only to within an order of magnitude. We can have sources, but we just won't describe the metric in the regions where the sources exist, e.g., inside the earth. The advantage of accepting this limitation is that in regions of empty space, we don't have to worry about the details of the stress-energy tensor or how it relates to curvature. As should be plausible based on the physical motivation given in Section 5.1, the field equations in a vacuum are simply $R_{ab} = 0$.

Thumbnail: This artist's concept illustrates a supermassive black hole with millions to billions times the mass of our sun. Supermassive black holes are enormously dense objects buried at the hearts of galaxies. (Public Domain; NASA/JPL-Caltech).

6.1: Event Horizons

One seemingly trivial way to generate solutions to the field equations in vacuum is simply to start with a flat Lorentzian spacetime and do a change of coordinates. This might seem pointless, since it would simply give a new description (and probably a less convenient and descriptive one) of the same old, boring, flat spacetime. It turns out, however, that some very interesting things can happen when we do this.

Figure 6.1.1: A Swiss commemorative coin shows the vacuum field equation.

The Event Horizon of an Accelerated Observer

Consider the uniformly accelerated observer described in example 4 and example 19. Recalling these earlier results, we have for the ship's equation of motion in an inertial frame

$$x = \frac{1}{a}(\sqrt{1 + a^2 t^2} - 1),\qquad(6.1.1)$$

and for the metric in the ship's frame

$$g'_{t't'} = (1 + ax')^2 \qquad(6.1.2)$$
$$g'_{x'x'} = -1. \qquad(6.1.3)$$

Since this metric was derived by a change of coordinates from a flat-space metric, and the Ricci curvature is an intrinsic property, we expect that this one also has zero Ricci curvature. This is straightforward to verify. The nonvanishing Christoffel symbols are

$$\Gamma^{t'}_{x't'} = \frac{a}{1 + ax'} \qquad(6.1.4)$$

and

$$\Gamma^{x'}_{t't'} = a(1 + ax'). \qquad(6.1.5)$$

The only elements of the Riemann tensor that look like they might be nonzero are $R^{t'}_{t'x'x'}$ and $R^{x'}_{t'x't'}$, but both of these in fact vanish.

> ### Exercise 6.1.1
>
> Self-check: Verify these facts.

This seemingly routine exercise now leads us into some very interesting territory. Way back in section 1.1, we conjectured that not all events could be time-ordered: that is, that there might exists events in spacetime 1 and 2 such that 1 cannot cause 2, but neither can 2 cause 1. We now have enough mathematical tools at our disposal to see that this is indeed the case.

We observe that x(t) approaches the asymptote $x = \frac{t-1}{a}$. This asymptote has a slope of 1, so it can be interpreted as the world-line of a photon that chases the ship but never quite catches up to it. Any event to the left of this line can never have a causal relationship with any event on the ship's world-line. Spacetime, as seen by an observer on the ship, has been divided by a curtain into two causally disconnected parts. This boundary is called an *event horizon*. Its existence is relative to the world-line of a particular observer. An observer who is not accelerating along with the ship does consider an event horizon to exist. Although this particular example of the indefinitely accelerating spaceship has some physically implausible features (e.g., the ship would have to

run out of fuel someday), event horizons are real things. In particular, we will see in section 6.3 that black holes have event horizons.

Interpreting everything in the (t', x') coordinates tied to the ship, the metric's component g'$_{t't'}$ vanishes at $x' = -\frac{1}{a}$. An observer aboard the ship reasons as follows. If I start out with a head-start of $\frac{1}{a}$ relative to some event, then the timelike part of the metric at that event vanishes. If the event marks the emission of a material particle, then there is no possible way for that particle's world-line to have $ds^2 > 0$. If I were to detect a particle emitted at that event, it would violate the laws of physics, since material particles must have $ds^2 > 0$, so I conclude that I will never observe such a particle. Since all of this applies to any material particle, regardless of its mass m, it must also apply in the limit $m \to 0$, i.e., to photons and other massless particles. Therefore I can never receive a particle emitted from this event, and in fact it appears that there is no way for that event, or any other event behind the event horizon, to have any effect on me. In my frame of reference, it appears that light cones near the horizon are tipped over so far that their future light-cones lie entirely in the direction away from me.
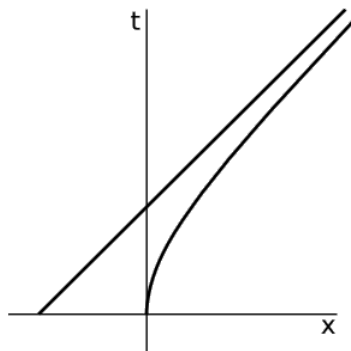
Figure 6.1.2 - A spaceship (curved worldline) moves with an acceleration perceived as constant by its passengers. The photon (straight world-line) comes closer and closer to the ship, but will never quite catch up.

We've already seen in Example 14 that a naive Newtonian argument suggests the existence of black holes; if a body is sufficiently compact, light cannot escape from it. In a relativistic treatment, this should be described as an *event horizon*.

Information Paradox

The existence of event horizons in general relativity has deep implications, and in particular it helps to explain why it is so difficult to reconcile general relativity with quantum mechanics, despite nearly a century of valiant attempts. Quantum mechanics has a property called unitarity. Mathematically, this says that if the state of a quantum mechanical system is given, at a certain time, in the form of a vector, then its state at some point in the future can be predicted by applying a unitary matrix to that vector. A unitary matrix is the generalization to complex numbers of the ordinary concept of an orthogonal matrix, and essentially it just represents a change of basis, in which the basis vectors have unit length and are perpendicular to one another.

To see what this means physically, consider the following nonexamples. The matrix

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \tag{6.1.6}$$

is not unitary, because its rows and columns are not orthogonal vectors with unit lengths. If this matrix represented the time-evolution of a quantum mechanical system, then its meaning would be that any particle in state number 1 would be left alone, but any particle in state 2 would disappear. Any information carried by particles in state 2 is lost forever and can never be retrieved. This also violates the time-reversal symmetry of quantum mechanics.

Another nonunitary matrix is:

$$\begin{pmatrix} 1 & 0 \\ 0 & \sqrt{2} \end{pmatrix} \tag{6.1.7}$$

Here, any particle in state 2 is increased in amplitude by a factor of $\sqrt{2}$, meaning that it is doubled in probability. That is, the particle is cloned. This is the opposite problem compared to the one posed by the first matrix, and it is equally problematic in terms of timereversal symmetry and conservation of information. Actually, if we could clone a particle in this way, it would violate the Heisenberg uncertainty principle. We could make two copies of the particle, and then measure the position of one copy and the

momentum of the other, each with unlimited precision. This would violate the uncertainty principle, so we believe that it cannot be done. This is known as the no-cloning theorem.[1]

> **Note**
>
> Ahn et al. have shown that the no-cloning theorem is violated in the presence of closed timelike curves: arxiv.org/abs/1008.0221v1

The existence of event horizons in general relativity violates unitarity, because it allows information to be destroyed. If a particle is thrown behind an event horizon, it can never be retrieved.

Radiation from Event Horizons

In interesting twist on the situation was introduced by Bill Unruh in 1976. Observer B aboard the accelerating spaceship believes in the equivalence principle, so she knows that the local properties of space at the event horizon would seem entirely normal and Lorentzian to a local observer A. (The same applies to a black hole's horizon.) In particular, B knows that A would see pairs of virtual particles being spontaneously created and destroyed in the local vacuum. This is simply a manifestation of the time-energy form of the uncertainty principle, $\Delta E \Delta t \leq h$. Now suppose that a pair of particles is created, but one is created in front of the horizon and one behind it. To A, these are virtual particles that will have to be annihilated within the time Δt, but according to B the one created in front of the horizon will eventually catch up with the spaceship, and can be observed there, although it will be red-shifted. The amount of redshift is given by

$$\sqrt{g'_{t't'}} = \sqrt{(1 + ax')^2}.$$ (6.1.8)

Say the pair is created right near the horizon, at $x' = -\frac{1}{a}$. By the uncertainty principle, each of the two particles is spread out over a region of space of size $\Delta x'$. Since these are photons, which travel at the speed of light, the uncertainty in position is essentially the same as the uncertainty in time. The forward-going photon's redshift comes out to be aΔx' = aΔt', which by the uncertainty principle should be at least $\frac{ha}{E}$, so that when the photon is observed by B, its energy is

$$E(\frac{ha}{E}) = ha.$$ (6.1.9)
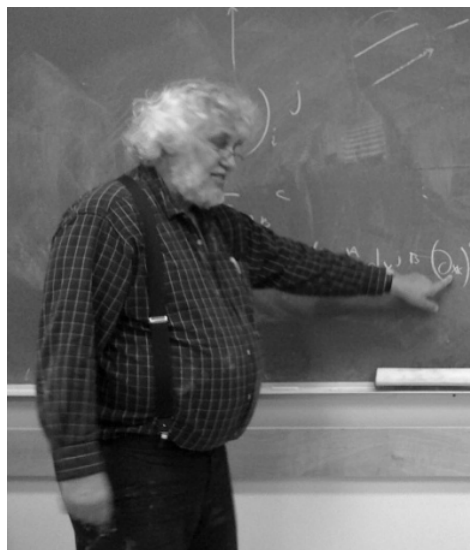
Figure $6.1.3$: Bill Unruh.

Now B sees a uniform background of photons, with energies of around ha, being emitted randomly from the horizon. They are being emitted from empty space, so it seems plausible to believe that they don't encode any information at all; they are completely random. A surface emitting a completely random (i.e., maximum-entropy) hail of photons is a black-body radiator, so we expect that the photons will have a black-body spectrum, with its peak at an energy of about ha. This peak is related to the temperature of the black body by E ∼ kT, where k is Boltzmann's constant. We conclude that the horizon acts like a black-body radiator with a temperature T ∼ $\frac{ha}{k}$. The more careful treatment by Unruh shows that the exact relation is $T = \frac{ha}{4\pi^2 k}$, or $\frac{ha}{4\pi^2 kc}$ in SI units.

An important observation here is that not only do different observers disagree about the number of quanta that are present (which is true in the case of ordinary Doppler shifts), but about the number of quanta in the vacuum as well. B sees photons that according to A do not exist.

Let's consider some real-world examples of large accelerations:

	acceleration (m/s^2)	temperature of horizon (K)
bullet fired from a gun	10^3	10^{-17}
electron in a CRT	10^7	10^{-13}
plasmas produced by intense laser pulses	10^{21}	10
proton in a helium nucleus	10^{27}	10^8

To detect Unruh radiation experimentally, we would ideally like to be able to accelerate a detector and let it detect the radiation. This is clearly impractical. The third line shows that it is possible to impart very large linear accelerations to subatomic particles, but then one can only hope to infer the effect of the Unruh radiation indirectly by its effect on the particles. As shown on the final line, examples of extremely large nonlinear accelerations are not hard to find, but the interpretation of Unruh radiation for nonlinear motion is unclear. A summary of the prospects for direct experimental detection of this effect is given by Rosu.[2] This type of experiment is clearly extremely difficult, but it is one of the few ways in which one could hope to get direct empirical insight, under controlled conditions, into the interface between gravity and quantum mechanics.

References

[2] http://xxx.lanl.gov/abs/gr-qc/9605032

6.2: The Schwarzschild Metric (Part 1)

We now set ourselves the goal of finding the metric describing the static spacetime outside a spherically symmetric, nonrotating, body of mass m. This problem was first solved by Karl Schwarzschild in 1915.[3] One byproduct of finding this metric will be the ability to calculate the geodetic effect exactly, but it will have more far-reaching consequences, including the existence of black holes. The problem we are solving is similar to calculating the spherically symmetric solution to Gauss's law in a vacuum. The solution to the electrical problem is of the form $\dfrac{\hat{\mathbf{r}}}{r^2}$, with an arbitrary constant of proportionality that turns out to be proportional to the charge creating the field. One big difference, however, is that whereas Gauss's law is linear, the equation

$$R_{ab} = 0 \tag{6.2.1}$$

is highly nonlinear, so that the solution cannot simply be scaled up and down in proportion to m.

The reason for this nonlinearity is fundamental to general relativity. For example, when the earth condensed out of the primordial solar nebula, large amounts of heat were produced, and this energy was then gradually radiated into outer space, decreasing the total mass of the earth. If we pretend, as in Figure 6.2.1, that this process involved the merging of only two bodies, each with mass m, then the net result was essentially to take separated masses m and m at rest, and bring them close together to form close-neighbor masses m and m, again at rest. The amount of energy radiated away was proportional to m², so the gravitational mass of the combined system has been reduced from $2m$ to $2m - (\ldots)m^2$, where . . . is roughly $\dfrac{G}{c^2 r}$. There is a nonlinear dependence of the gravitational field on the masses.
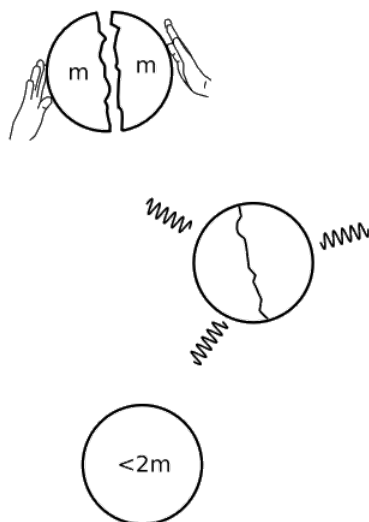
Figure 6.2.1: The field equations of general relativity are nonlinear.

Exercise 6.2.1

The signature of a metric is defined as the list of positive and negative signs that occur when it is diagonalized.[4] The equivalence principle requires that the signature be $+ - --$ (or $- + ++$, depending on the choice of sign conventions). Verify that any constant metric (including a metric with the "wrong" signature, e.g., 2+2 dimensions rather than 3+1) is a solution to the Einstein field equation in vacuum.

---

[4] See section 6.3 for a different but closely related use of the same term.

The correspondence principle tells us that our result must have a Newtonian limit, but the only variables involved are m and r, so this limit must be the one in which $\dfrac{r}{m}$ is large. Large compared to what? There is nothing else available with which to compare, so it can only be large compared to some expression composed of the unitless constants G and c. We have already chosen units such that $c = 1$, and we will now set $G = 1$ as well. Mass and distance are now comparable, with the conversion factor being $\dfrac{G}{c^2} = 7 \times 10^{-28} \,\mathrm{m/kg}$, or about a mile per solar mass. Since the earth's radius is thousands of times more

than a mile, and its mass hundreds of thousands of times less than the sun's, its $\dfrac{r}{m}$ is very large, and the Newtonian approximation is good enough for all but the most precise applications, such as the GPS network or the Gravity Probe B experiment.

The Zero-mass Case

First let's demonstrate the trivial solution with flat spacetime. In spherical coordinates, we have

$$ds^2 = dt^2 - dr^2 - r^2\theta^2 - r^2\sin^2\theta d\phi^2. \tag{6.2.2}$$

The nonvanishing Christoffel symbols (ignoring swaps of the lower indices) are:

$$\Gamma^\theta_{r\theta} = \frac{1}{r}$$
$$\Gamma^\phi_{r\phi} = \frac{1}{r}$$
$$\Gamma^r_{\theta\theta} = -r$$
$$\Gamma^r_{\phi\phi} = -r\sin^2\theta$$
$$\Gamma^\theta_{\phi\phi} = \sin\theta\cos\theta$$
$$\Gamma^\phi_{\theta\phi} = \cot\theta$$

Exercise 6.2.2

If we'd been using the $(-+++)$ metric instead of $(+---)$, what would have been the effect on the Christoffel symbols? What if we'd expressed the metric in different units, rescaling all the coordinates by a factor k?

Use of c tensor

In fact, when I calculated the Christoffel symbols above by hand, I got one of them wrong, and missed calculating one other because I thought it was zero. I only found my mistake by comparing against a result in a textbook. The computation of the Riemann tensor is an even bigger mess. It's clearly a good idea to resort to a computer algebra system here. Cadabra, which was discussed earlier, is specifically designed for coordinate-independent calculations, so it won't help us here. A good free and open-source choice is ctensor, which is one of the standard packages distributed along with the computer algebra system Maxima, introduced in section 2.5.

The following Maxima program calculates the Christoffel symbols found earlier.

```
1    load(ctensor);
2    ct_coords:[t,r,theta,phi];
3    lg:matrix([1,0,0,0],
4              [0,-1,0,0],
5              [0,0,-r^2,0],
6              [0,0,0,-r^2*sin(theta)^2]);
7    cmetric();
8    christof(mcs);
```

Line 1 loads the ctensor package. Line 2 sets up the names of the coordinates. Line 3 defines the g_{ab}, with 1g meaning "the version of g with lower indices." Line 7 tells Maxima to do some setup work with g_{ab}, including the calculation of the inverse matrix g^{ab}, which is stored in ug. Line 8 says to calculate the Christoffel symbols. The notation mcs refers to the tensor Γ'^a_{bc} with the indices swapped around a little compared to the convention Γ^a_{bc} followed in this book. On a Linux system, we put the program in a file flat.mac and run it using the command maxima -b flat.mac. The relevant part of the output is:

```
1                                                 1
2    (%t6)                           mcs        = -
3                                        2, 3, 3   r
4
5                                                 1
6    (%t7)                           mcs        = -
7                                        2, 4, 4   r
8
9    (%t8)                               mcs        = - r
10                                          3, 3, 2
11
12                                               cos(theta)
13   (%t9)                           mcs        = ----------
14                                      3, 4, 4   sin(theta)
15
16                                               2
17   (%t10)                      mcs        = - r sin (theta)
18                                  4, 4, 2
19
20   (%t11)                      mcs        = - cos(theta) sin(theta)
21                                  4, 4, 3
```

Adding the command ricci(true); at the end of the program results in the output THIS SPACETIME IS EMPTY AND/OR FLAT, which saves us hours of tedious computation. The tensor ric (which here happens to be zero) is computed, and all its nonzero elements are printed out. There is a similar command riemann(true); to compute the Riemann rensor riem. This is stored so that riem[i,j,k,l] is what we would call R^l_{ikj}. Note that l is moved to the end, and j and k are also swapped.

Geometrized Units

If the mass creating the gravitational field *isn't* zero, then we need to decide what units to measure it in. It has already proved very convenient to adopt units with $c = 1$, and we will now also set the gravitational constant $G = 1$. Previously, with only c set to 1, the units of time and length were the same, [T] = [L], and so were the units of mass and energy, [M] = [E]. With G = 1, all of these become the same units:

$$[T] = [L] = [M] = [E]. \tag{6.2.3}$$

Exercise 6.2.3

Verify this statement by combining Newton's law of gravity with Newton's second law of motion.

The resulting system is referred to as geometrized, because units like mass that had formerly belonged to the province of mechanics are now measured using the same units we would use to do geometry.

A Large-r Limit

Now let's think about how to tackle the real problem of finding the non-flat metric. Although general relativity lets us pick any coordinates we like, the spherical symmetry of the problem suggests using coordinates that exploit that symmetry. The flat-space coordinates θ and ϕ can stil be defined in the same way, and they have the same interpretation. For example, if we drop a test particle toward the mass from some point in space, its world-line will have constant θ and ϕ. The r coordinate is a little different. In curved spacetime, the circumference of a circle is not equal to 2π times the distance from the center to the circle; in fact, the discrepancy between these two is essentially the definition of the **Ricci curvature**. This gives us a choice of two logical ways to define r. We'll define it as the circumference divided by 2π, which has the advantage that the last two terms of the metric are the same as in flat space:

$$-r^2 d\theta^2 - r^2 \sin^2 \theta \, d\phi^2. \tag{6.2.4}$$

Since we're looking for static solutions, none of the elements of the metric can depend on t. Also, the solution is going to be symmetric under t → −t, $\theta \to -\theta$, and $\phi \to -\phi$, so we can't have any off-diagonal elements.[5] The result is that we have narrowed the metric down to something of the form

$$ds^2 = h(r)dt^2 - k(r)dr^2 - r^2 d\theta^2 - r^2 \sin^2 \theta d\phi^2, \qquad (6.2.5)$$

where both h and k approach 1 for r → ∞, where spacetime is flat.

> [5]For more about time-reversal symmetry, see later.

For guidance in how to construct h and k, let's consider the acceleration of a test particle at $r \gg m$, which we know to be $-\dfrac{m}{r^2}$, since nonrelativistic physics applies there. We have

$$\nabla_t v^r = \partial_t v^r + \Gamma^r_{tc} v^c. \qquad (6.2.6)$$

An observer free-falling along with the particle observes its acceleration to be zero, and a tensor that is zero in one coordinate system is zero in all others. Since the covariant derivative is a tensor, we conclude that $\nabla_t v^r = 0$ in all coordinate systems, including the (t, r, . . .) system we're using. If the particle is released from rest, then initially its velocity four-vector is (1, 0, 0, 0), so we find that its acceleration in (t, r) coordinates is $-\Gamma^r_{tt} = -\dfrac{1}{2}g^{rr}\partial_r g_{tt} = -\dfrac{1}{2}\dfrac{h'}{k}$. Setting this equal to $-\dfrac{m}{r^2}$, we find $\dfrac{h'}{k} = \dfrac{2m}{r^2}$ for r >> m. Since k ≈ 1 for large r, we have

$$h' \approx \frac{2m}{r^2} \; for \; r >> m. \qquad (6.2.7)$$

The interpretation of this calculation is as follows. We assert the equivalence principle, by which the acceleration of a free-falling particle can be said to be zero. After some calculations, we find that the rate at which time flows (encoded in h) is not constant. It is different for observers at different heights in a gravitational potential well. But this is something we had already deduced, without the index gymnastics, in example 7.

Integrating, we find that for large r, h = 1 − $\dfrac{2m}{r}$.

The Complete Solution

A Series Solution

We've learned some interesting things, but we still have an extremely nasty nonlinear differential equation to solve. One way to attack a differential equation, when you have no idea how to proceed, is to try a series solution. We have a small parameter $\dfrac{m}{r}$ to expand around, so let's try to write h and k as series of the form

$$h = \sum_{n=0}^{\infty} a_k \left(\frac{m}{r}\right)^n \qquad (6.2.8)$$

$$k = \sum_{n=0}^{\infty} b_k \left(\frac{m}{r}\right)^n \qquad (6.2.9)$$

We already know a_0, a_1, and b_0. Let's try to find b_1. In the following Maxima code I omit the factor of m in h_1 for convenience. In other words, we're looking for the solution for m = 1.

```
1   load(ctensor);
2   ct_coords:[t,r,theta,phi];
3   lg:matrix([(1-2/r),0,0,0],
4            [0,-(1+b1/r),0,0],
5            [0,0,-r^2,0],
6            [0,0,0,-r^2*sin(theta)^2]);
7   cmetric();
8   ricci(true);
```

I won't reproduce the entire output of the Ricci tensor, which is voluminous. We want all four of its nonvanishing components to vanish as quickly as possible for large values of r, so I decided to fiddle with R_{tt}, which looked as simple as any of them. It appears

to vary as r^{-4} for large r, so let's evaluate $\lim_{r \to \infty} (r^4 R_{tt})$:

```
9    limit(r^4*ric[1,1],r,inf);
```

The result is $\dfrac{(b_1 - 2)}{2}$, so let's set $b_1 = 2$. The approximate solution we've found so far (reinserting the m's),

$$ds^2 \approx \left(1 - \frac{2m}{r}\right) dt^2 - \left(1 + \frac{2m}{r}\right) dr^2 - r^2 d\theta^2 - r^2 \sin^2 \theta d\phi^2, \tag{6.2.10}$$

was first derived by Einstein in 1915, and he used it to solve the problem of the non-Keplerian relativistic correction to the orbit of Mercury, which was one of the first empirical tests of general relativity.

Continuing in this fashion, the results are as follows:

$a_0 = 1$	$b_0 = 1$
$a_1 = -2$	$b_1 = 2$
$a_2 = 0$	$b_2 = 4$
$a_3 = 0$	$b_2 = 8$

The Closed-form Solution

The solution is unexpectedly simple, and can be put into closed form. The approximate result we found for h was in fact exact. For k we have a geometric series $\dfrac{1}{(1 - \dfrac{2}{r})}$, and when we reinsert the factor of m in the only way that makes the units work, we get

$\dfrac{1}{(1 - \dfrac{2m}{r})}$. The result for the metric is

$$ds^2 = \left(1 - \frac{2m}{r}\right) dt^2 - \left(\frac{1}{1 - \dfrac{2m}{r}}\right) dr^2 - r^2 d\theta^2 - r^2 \sin^2 \theta d\phi^2. \tag{6.2.11}$$

This is called the Schwarzschild metric. A quick calculation in Maxima demonstrates that it is an exact solution for all r, i.e., the Ricci tensor vanishes everywhere, even at r < 2m, which is outside the radius of convergence of the geometric series.

Time-reversal Symmetry

The Schwarzschild metric is invariant under time reversal, since time occurs only in the form of dt^2, which stays the same under dt → − dt. This is the same time-reversal symmetry that occurs in Newtonian gravity, where the field is described by the gravitational acceleration g, and accelerations are time-reversal invariant.

Fundamentally, this is an example of general relativity's coordinate independence. The laws of physics provided by general relativity, such as the vacuum field equation, are invariant under any smooth coordinate transformation, and t → −t is such a coordinate transformation, so general relativity has time-reversal symmetry. Since the Schwarzschild metric was found by imposing time-reversalsymmetric boundary conditions on a time-reversal-symmetric differential equation, it is an equally valid solution when we time-reverse it. Furthermore, we expect the metric to be invariant under time reversal, unless spontaneous symmetry breaking occurs (see section 8.2).

This suggests that we ask the more fundamental question of what global symmetries general relativity has. Does it have symmetry under parity inversion, for example? Or can we take any solution such as the Schwarzschild spacetime and transform it into a frame of reference in which the source of the field is moving uniformly in a certain direction? Because general relativity is locally equivalent to special relativity, we know that these symmetries are locally valid. But it may not even be possible to define the corresponding global symmetries. For example, there are some spacetimes on which it is not even possible to define a global time coordinate. On such a spacetime, which is described as not time-orientable, there does not exist any smooth vector field that is everywhere timelike, so it is not possible to define past versus future light-cones at all points in space without having a

discontinuous change in the definition occur somewhere. This is similar to the way in which a Möbius strip does not allow an orientation of its surface (an "up" direction as seen by an ant) to be defined globally.

Suppose that our spacetime is time-orientable, and we are able to define coordinates (p, q, r, s) such that p is always the timelike coordinate. Because q → −q is a smooth coordinate transformation, we are guaranteed that our spacetime remains a valid solution of the field equations under this change. But that doesn't mean that what we've found is a symmetry under parity inversion in a plane. Our coordinate q is not necessarily interpretable as distance along a particular "q axis." Such axes don't even exist globally in general relativity. A coordinate does not even have to have units of time or distance; it could be an angle, for example, or it might not have any geometrical significance at all. Similarly, we could do a transformation q → q' = q + kp. If we think of q as measuring spatial position and p time, then this looks like a Galilean transformation, with k being the velocity. The solution to the field equations obtained after performing this transformation is still a valid solution, but that doesn't mean that relativity has Galilean symmetry rather than Lorentz symmetry. There is no sensible way to define a Galilean transformation acting on an entire spacetime, because when we talk about a Galilean transformation we assume the existence of things like global coordinate axes, which do not even exist in general relativity.

References

[3] "On the gravitational field of a point mass according to Einstein's theory," Sitzungsberichte der Koniglich Preussischen Akademie der Wissenschaften 1 (1916) 189. An English translation is available at http://arxiv.org/abs/physics/9905030v1.

This page titled 6.2: The Schwarzschild Metric (Part 1) is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Benjamin Crowell via source content that was edited to the style and standards of the LibreTexts platform.

6.3: The Schwarzschild Metric (Part 2)

Geodetic Effect

As promised in section 5.5, we now calculate the geodetic effect on Gravity Probe B, including all the niggling factors of 3 and π. To make the physics clear, we approach the actual calculation through a series of warmups.

Flat Space

As a first warmup, consider two spatial dimensions, represented by Euclidean polar coordinates (r, ϕ). Parallel-transport of a gyroscope's angular momentum around a circle of constant r gives

$$\nabla_\phi L^\phi = 0$$
$$\nabla_\phi L^r = 0.$$

Computing the covariant derivatives, we have

$$0 = \partial_\phi L^\phi + \Gamma^\phi_{\phi r} L^r$$
$$0 = \partial_\phi L^r + \Gamma^r_{\phi\phi} L^\phi.$$

The Christoffel symbols are $\Gamma^\phi_{\phi r} = \frac{1}{r}$ and $\Gamma^r_{\phi\phi} = -r$. This is all made to look needlessly complicated because L^ϕ and L^r are expressed in different units. Essentially the vector is staying the same, but we're expressing it in terms of basis vectors in the r and ϕ directions that are rotating. To see this more transparently, let r = 1, and write P for L^ϕ and Q for L^r, so that

$$P' = -Q$$
$$Q' = P,$$

which have solutions such as P = sin ϕ, Q = cos ϕ. For each orbit (2$\pi\pi$change in ϕ), the basis vectors rotate by 2ϕ, so the angular momentum vector once again has the same components. In other words, it hasn't really changed at all.

Spatial Curvature Only

The flat-space calculation above differs in two ways from the actual result for an orbiting gyroscope: (1) it uses a flat spatial geometry, and (2) it is purely spatial. The purely spatial nature of the calculation is manifested in the fact that there is nothing in the result relating to how quickly we've moved the vector around the circle. We know that if we whip a gyroscope around in a circle on the end of a rope, there will be a Thomas precession (section 2.5), which depends on the speed.

As our next warmup, let's curve the spatial geometry, but continue to omit the time dimension. Using the Schwarzschild metric, we replace the flat-space Christoffel symbol $\Gamma^r_{\phi\phi} = -r$ with −r+2m. The differential equations for the components of the L vector, again evaluated at r = 1 for convenience, are now

$$P' = -Q$$
$$Q' = (1 - \epsilon)P,$$

where ϵ = 2m. The solutions rotate with frequency $\omega' = \sqrt{1 - \epsilon}$. The result is that when the basis vectors rotate by 2π, the components no longer return to their original values; they lag by a factor of $\sqrt{1 - \epsilon} \approx 1 - m$. Putting the factors of r back in, this is $1 - \frac{m}{r}$. The deviation from unity shows that after one full revolution, the L vector no longer has quite the same components expressed in terms of the (r, ϕ) basis vectors.

To understand the sign of the effect, let's imagine a counterclockwise rotation. The (r, ϕ) rotate counterclockwise, so relative to them, the L vector rotates clockwise. After one revolution, it has not rotated clockwise by a full 2π, so its orientation is now slightly counterclockwise compared to what it was. Thus the contribution to the geodetic effect arising from spatial curvature is in the same direction as the orbit.

Comparing with the actual results from Gravity Probe B, we see that the direction of the effect is correct. The magnitude, however, is off. The precession accumulated over n periods is $\frac{2\pi nm}{r}$, or, in SI units, $\frac{2\pi nGm}{c^2 r}$. Using the data from section 2.5, we find $\Delta\theta = 2 \times 10^{-5}$ radians, which is too small compared to the data shown in Figure 5.5.2.

To reproduce the experimental results correctly, we need to include the time dimension. The angular momentum vector now has components (L^ϕ, L^r, L^t). The physical interpretation of the L^t component is obscure at this point; we'll return to this question later.

Writing down the total derivatives of the three components, and notating $\frac{dt}{d\phi}$ as ω^{-1}, we have

$$\frac{dL^\phi}{d\phi} = \partial_\phi L^\phi + \omega^{-1}\partial_t L^\phi \qquad (6.3.1)$$

$$\frac{dL^r}{d\phi} = \partial_\phi L^r + \omega^{-1}\partial_t L^r \qquad (6.3.2)$$

$$\frac{dL^t}{d\phi} = \partial_\phi L^t + \omega^{-1}\partial_t L^t \qquad (6.3.3)$$

Setting the covariant derivatives equal to zero gives

$$0 = \partial_\phi L^\phi + \Gamma^\phi_{\phi r} L^r$$
$$0 = \partial_\phi L^r + \Gamma^r_{\phi\phi} L^\phi$$
$$0 = \partial_t L^r + \Gamma^r_{tt} L^t$$
$$0 = \partial_t L^t + \Gamma^t_{tr} L^r.$$

Exercise 6.3.4

Self-check: There are not just four but six covariant derivatives that could in principle have occurred, and in these six covariant derivatives we could have had a total of 18 Christoffel symbols. Of these 18, only four are nonvanishing. Explain based on symmetry arguments why the following Christoffel symbols must vanish: $\Gamma^\phi_{\phi t}, \Gamma^t_{tt}$.

Putting all this together in matrix form, we have L' = ML, where

$$M = \begin{pmatrix} 0 & -1 & 0 \\ 1-\epsilon & 0 & -\frac{\epsilon(1-\epsilon)}{2\omega} \\ 0 & -\frac{\epsilon}{2\omega(1-\epsilon)} & 0 \end{pmatrix}. \qquad (6.3.4)$$

The solutions of this differential equation oscillate like $e^{i\Omega t}$, where $i\Omega$ is an eigenvalue of the matrix.

Exercise 6.3.5

Self-check: The frequency in the purely spatial calculation was found by inspection. Verify the result by applying the eigenvalue technique to the relevant 2 × 2 submatrix.

To lowest order, we can use the Newtonian relation $\omega^2 r = \frac{Gm}{r}$ and neglect terms of order ϵ^2, so that the two new off-diagonal matrix elements are both approximated as $\sqrt{\frac{\epsilon}{2}}$. The three resulting eigenfrequencies are zero and $\Omega = \pm[1 - \frac{3}{2})$m/r].

The presence of the mysterious zero-frequency solution can now be understood by recalling the earlier mystery of the physical interpretation of the angular momentum's L^t component. Our results come from calculating parallel transport, and parallel transport is a purely geometric process, so it gives the same result regardless of the physical nature of the four-vector. Suppose that we had instead chosen the velocity four-vector as our guinea pig. The definition of a geodesic is that it parallel-transports its own tangent vector, so the velocity vector has to stay constant. If we inspect the eigenvector corresponding to the zero-frequency eigenfrequency, we find a timelike vector that is parallel to the velocity four-vector. In our 2+1-dimensional space, the other two eigenvectors, which are spacelike, span the subspace of spacelike vectors, which are the ones that can physically be realized as the angular momentum of a gyroscope. These two eigenvectors, which vary as $e^{\pm i\Omega}$, can be superposed to make real-valued spacelike solutions that match the initial conditions, and these lag the rotation of the basis vectors by $\Delta\Omega = \frac{3}{2}$mr. This is greater than the purely spatial result by a factor of $\frac{3}{2}$. The resulting precession angle, over n orbits of Gravity Probe B, is $\frac{3\pi nGm}{c^2 r} = 3 \times 10^{-5}$ radians, in excellent agreement with experiment.

One will see apparently contradictory statements in the literature about whether Thomas precession occurs for a satellite: "The Thomas precession comes into play for a gyroscope on the surface of the Earth . . . , but not for a gyroscope in a freely moving satellite."[6] But: "The total effect, geometrical and Thomas, gives the well-known Fokker-de Sitter precession of $\frac{3\pi m}{r}$, in the same sense as the orbit."[7] The second statement arises from subtracting the purely spatial result from the 2+1-dimensional result, and noting that the absolute value of this difference is the same as the Thomas precession that *would* have been obtained if the gyroscope had been whirled at the end of a rope. In my opinion this is an unnatural way of looking at the physics, for two reasons. (1) The signs don't match, so one is forced to say that the Thomas precession has a different sign depending on whether the rotation is the result of gravitational or nongravitational forces. (2) Referring to observation, it is clearly artificial to treat the spatial curvature and Thomas effects separately, since neither one can be disentangled from the other by varying the quantities n, m, and r. For more discussion, see tinyurl.com/me3qf8o.

Orbits

The main event of Newton's *Principia Mathematica* is his proof of Kepler's laws. Similarly, Einstein's first important application in general relativity, which he began before he even had the exact form of the Schwarzschild metric in hand, was to find the non-Newtonian behavior of the planet Mercury. The planets deviate from Keplerian behavior for a variety of Newtonian reasons, and in particular there is a long list of reasons why the major axis of a planet's elliptical orbit is expected to gradually rotate. When all of these were taken into account, however, there was a remaining discrepancy of about 40 seconds of arc per century, or 6.6×10^{-7} radians per orbit. The direction of the effect was in the forward direction, in the sense that if we view Mercury's orbit from above the ecliptic, so that it orbits in the counterclockwise direction, then the gradual rotation of the major axis is also counterclockwise.

As a very rough hand-wavy explanation for this effect, consider the spatial part of the curvature of the spacetime surrounding the sun. This spatial curvature is positive, so a circle's circumference is less than 2π times its radius. We could imagine that this would cause Mercury to get back to a previously visited angular position before it has had time to complete its Newtonian cycle of radial motion. Arguments such as this one, however, should not be taken too seriously. A mathematical analysis is required.

Based on the examples in section 5.5, we expect that the effect will be of order $\frac{m}{r}$, where m is the mass of the sun and r is the radius of Mercury's orbit. This works out to be 2.5×10^{-8}, which is smaller than the observed precession by a factor of about 26.

Conserved Quantities

If Einstein had had a computer on his desk, he probably would simply have integrated the motion numerically using the geodesic equation. But it is possible to simplify the problem enough to attack it with pencil and paper, if we can find the relevant conserved quantities of the motion. Nonrelativistically, these are energy and angular momentum.

Consider a rock falling directly toward the sun. The Schwarzschild metric is of the special form

$$ds^2 = h(r)dt^2 - k(r)dr^2 - \ldots \tag{6.3.5}$$

The rock's trajectory is a geodesic, so it extremizes the proper time s between any two events fixed in spacetime, just as a piece of string stretched across a curved surface extremizes its length. Let the rock pass through distance r_1 in coordinate time t_1, and then through r_2 in t_2. (These should really be notated as $\Delta r_1, \ldots$ or $dr_1, \ldots$, but we avoid the Δ's or d's for convenience.) Approximating the geodesic using two line segments, the proper time is

$$s = s_1 + s_2$$
$$= \sqrt{h_1 t_1^2 - k_1 r_1^2} + \sqrt{h_2 t_2^2 - k_2 r_2^2}$$
$$= \sqrt{h_1 t_1^2 - k_1 r_1^2} + \sqrt{h_2 (T - t_1)^2 - k_2 r_2^2},$$

where $T = t_1 + t_2$ is fixed. If this is to be extremized with respect to t_1, then $\frac{ds}{dt_1} = 0$, which leads to

$$0 = \frac{h_1 t_1}{s_1} - \frac{h_2 t_2}{s_2}, \tag{6.3.6}$$

which means that

$$h\frac{dt}{ds} = g_{tt}\frac{dx^t}{ds} = \frac{dx_t}{ds} \tag{6.3.7}$$

is a constant of the motion. Except for an irrelevant factor of m, this is the same as p_t, the timelike component of the covariant momentum vector. We've already seen that in special relativity, the timelike component of the momentum four-vector is interpreted as the mass-energy E, and the quantity p_t has a similar interpretation here. Note that no special assumption was made about the form of the functions h and k. In addition, it turns out that the assumption of purely radial motion was unnecessary. All that really mattered was that h and k were independent of t. Therefore we will have a similar conserved quantity p^μ any time the metric's components, expressed in a particular coordinate system, are independent of x^μ. (This is generalized in section 7.1.) In particular, the Schwarzschild metric's components are independent of ϕ as well as t, so we have a second conserved quantity p_ϕ, which is interpreted as angular momentum.
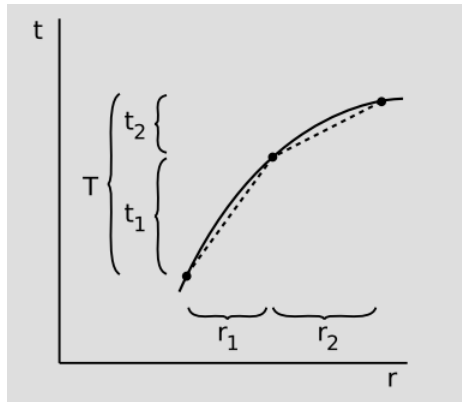

Figure 6.3.2 - Proof that if the metric's components are independent of t, the geodesic of a test particle conserves p_t.

Writing these two quantities out explicitly in terms of the contravariant coordinates, in the case of the Schwarzschild spacetime, we have

$$E = \left(1 - \frac{2m}{r}\right)\frac{dt}{ds} \tag{6.3.8}$$

and

$$L = r^2 \frac{d\phi}{ds} \tag{6.3.9}$$

for the conserved energy per unit mass and angular momentum per unit mass.

In interpreting the energy per unit mass E, it is important to understand that in the general-relativistic context, there is no useful way of separating the rest mass, kinetic energy, and potential energy into separate terms, as we could in Newtonian mechanics. E includes contributions from all of these, and turns out to be less than the contribution due to the rest mass (i.e., less than 1) for a planet orbiting the sun. It turns out that E can be interpreted as a measure of the additional gravitational mass that the solar system possesses as measured by a distant observer, due to the presence of the planet. It then makes sense that E is conserved; by analogy with Newtonian mechanics, we would expect that any gravitational effects that depended on the detailed arrangement of the masses within the solar system would decrease as $\frac{1}{r^4}$, becoming negligible at large distances and leaving a constant field varying as $\frac{1}{r^2}$.

One way of seeing that it doesn't make sense to split E into parts is that although the equation given above for E involves a specific set of coordinates, E can actually be expressed as a Lorentz-invariant scalar (see section 7.1). This property makes E especially interesting and useful (and different from the energy in Newtonian mechanics, which is conserved but not frame-independent). On the other hand, the kinetic and potential energies depend on the velocity and position. These are completely dependent on the coordinate system, and there is nothing physically special about the coordinate system we've used here. Suppose a particle is falling directly toward the earth, and an astronaut in a space-suit is free-falling along with it and monitoring its progress. The astronaut judges the particle's kinetic energy to be zero, but other observers say it's nonzero, so it's clearly not a Lorentz scalar. And suppose the astronaut insists on defining a potential energy to go along with this kinetic energy. The potential energy must be decreasing, since the particle is getting closer to the earth, but then there is no way that the sum of the kinetic and potential energies could be constant.

Perihelion Advance

For convenience, let the mass of the orbiting rock be 1, while m stands for the mass of the gravitating body.

The unit mass of the rock is a third conserved quantity, and since the magnitude of the momentum vector equals the square of the mass, we have for an orbit in the plane $\theta = \frac{\pi}{2}$,

$$
\begin{aligned}
1 &= g^{tt}p_t^2 - g^{rr}p_r^2 - g^{\phi\phi}p_\phi^2 \\
&= g^{tt}p_t^2 g_{rr}(p^r)^2 - g^{\phi\phi}p_\phi^2 \\
&= \frac{1}{1-\frac{2m}{r}}E^2 - \frac{1}{1-\frac{2m}{r}}\left(\frac{dr}{ds}\right)^2 - \frac{1}{r^2}L^2.
\end{aligned}
$$

Rearranging terms and writing $\dot{r}$ for $\frac{dr}{ds}$, this becomes

$$
\dot{r}^2 = E^2 - \left(1-\frac{2m}{r}\right)\left(d\frac{1+L^2}{r^2}\right) \tag{6.3.10}
$$

or

$$
\dot{r}^2 = E^2 - U^2 \tag{6.3.11}
$$

where

$$
U^2 = \left(1-\frac{2m}{r}\right)\left(1+\frac{L^2}{r^2}\right). \tag{6.3.12}
$$

There is a varied and strange family of orbits in the Schwarzschild field, including bizarre knife-edge trajectories that take several nearly circular turns before suddenly flying off. We turn our attention instead to the case of an orbit such as Mercury's which is nearly Newtonian and nearly circular.

Nonrelativistically, a circular orbit has radius $r = \frac{L^2}{m}$ and period $T = \frac{2\pi L^3}{m^2}$.

Relativistically, a circular orbit occurs when there is only one turning point at which $\dot{r} = 0$. This requires that E^2 equal the minimum value of U^2, which occurs at

$$
\begin{aligned}
r &= \frac{L^2}{2m}\left(1+\sqrt{1-\frac{12m^2}{L^2}}\right) \\
&\approx \frac{L^2}{m}(1-\epsilon),
\end{aligned}
$$

where $\epsilon = 3\left(\frac{m}{L}\right)^2$. A planet in a nearly circular orbit oscillates between perihelion and aphelion with a period that depends on the curvature of U^2 at its minimum. We have

$$
\begin{aligned}
k &= \frac{d^2(U^2)}{dr^2} \\
&= \frac{d^2}{dr^2}\left(1-\frac{2m}{r}+\frac{L^2}{r^2}-\frac{2mL^2}{r^3}\right) \\
&= -\frac{4m}{r^3}+\frac{6L^2}{r^4}-\frac{24mL^2}{r^5} \\
&= 2L^{-6}m^4(1+2\epsilon)
\end{aligned}
$$

The period of the oscillations is

$$
\begin{aligned}
\Delta s_{osc} &= 2\pi\sqrt{\frac{2}{k}} \\
&= 2\pi L^3 m^{-2}(1-2\epsilon).
\end{aligned}
$$

The period of the azimuthal motion is

$$\Delta s_{az} = \frac{2\pi r^2}{L}$$
$$= 2\pi L^3 m^{-2}(1 - 2\epsilon).$$

The periods are slightly mismatched because of the relativistic correction terms. The period of the radial oscillations is longer, so that, as expected, the perihelion shift is in the forward direction. The mismatch is $\epsilon\Delta s$, and because of it each orbit rotates the major axis by an angle $2\pi\epsilon = 6\pi(\frac{m}{L})^2 = \frac{6\pi m}{r}$. Plugging in the data for Mercury, we obtain 5.8×10^{-7} radians per orbit, which agrees with the observed value to within about 10%. Eliminating some of the approximations we've made brings the results in agreement to within the experimental error bars, and Einstein recalled that when the calculation came out right, "for a few days, I was beside myself with joyous excitement."

Further attempts were made to improve on the precision of this historically crucial test of general relativity. Radar now gives the most precise orbital data for Mercury. At the level of about one part per thousand, however, an effect creeps in due to the oblateness of the sun, which is difficult to measure precisely.

In 1974, astronomers J.H. Taylor and R.A. Hulse of Princeton, working at the Arecibo radio telescope, discovered a binary star system whose members are both neutron stars. The detection of the system was made possible because one of the neutron stars is a pulsar: a neutron star that emits a strong radio pulse in the direction of the earth once per rotational period. The orbit is highly elliptical, and the minimum separation between the two stars is very small, about the same as the radius of our sun. Both because the r is small and because the period is short (about 8 hours), the rate of perihelion advance per unit time is very large, about 4.2 degrees per year. The system has been compared in great detail with the predictions of general relativity,8 giving extremely good agreement, and as a result astronomers have been confident enough to reason in the opposite direction and infer properties of the system, such as its total mass, from the general-relativistic analysis. The system's orbit is decaying due to the radiation of energy in the form of gravitational waves, which are predicted to exist by relativity.

Deflection of Light

As discussed in section 5.5, one of the first tests of general relativity was Eddington's measurement of the deflection of rays of light by the sun's gravitational field. The deflection measured by Eddington was 1.6 seconds of arc. For a light ray that grazes the sun's surface, the only physically relevant parameters are the sun's mass m and radius r. Since the deflection is unitless, it can only depend on $\frac{m}{r}$, the unitless ratio of the sun's mass to its radius. Expressed in SI units, this is $\frac{Gm}{c^2 r}$, which comes out to be about 10^{-6}. Roughly speaking, then, we expect the order of magnitude of the effect to be about this big, and indeed 10^{-6} radians comes out to be in the same ball-park as a second of arc. We get a similar estimate in Newtonian physics by treating a photon as a (massive) particle moving at speed c.

It is possible to calculate a precise value for the deflection using methods very much like those used to determine the perihelion advance from earlier. However, some of the details would have to be changed. For example, it is no longer possible to parametrize the trajectory using the proper time s, since a light ray has ds = 0; we must use an affine parameter. Let us instead use this an an example of the numerical technique for solving the geodesic equation, first demonstrated in section 5.9. Modifying our earlier program, we have the following:

```
1    import math
2
3    # constants, in SI units:
4    G = 6.67e-11            # gravitational constant
5    c = 3.00e8              # speed of light
6    m_kg = 1.99e30          # mass of sun
7    r_m = 6.96e8            # radius of sun
8
9    # From now on, all calculations are in units of the
10   # radius of the sun.
11
12   # mass of sun, in units of the radius of the sun:
13   m_sun = (G/c**2)*(m_kg/r_m)
14   m = 1000.*m_sun
15   print "m/r=",m
16
17   # Start at point of closest approach.
18   # initial position:
19   t=0
20   r=1 # closest approach, grazing the sun's surface
21   phi=-math.pi/2
22   # initial derivatives of coordinates w.r.t. lambda
23   vr = 0
24   vt = 1
25   vphi = math.sqrt((1.-2.*m/r)/r**2)*vt # gives ds=0, lightlike
26
27   l = 0     # affine parameter lambda
28   l_max = 20000.
29   epsilon = 1e-6 # controls how fast lambda varies
30   while l<l_max:
31     dl = epsilon*(1.+r**2) # giant steps when farther out
32     l = l+dl
33     # Christoffel symbols:
34     Gttr = m/(r**2-2*m*r)
35     Grtt = m/r**2-2*m**2/r**3
36     Grrr = -m/(r**2-2*m*r)
37     Grphiphi = -r+2*m
38     Gphirphi = 1/r
39     # second derivatives:
40     #  The factors of 2 are because we have, e.g., G^a_{bc}=G^a_{cb}
41     at   = -2.*Gttr*vt*vr
42     ar   = -(Grtt*vt*vt + Grrr*vr*vr + Grphiphi*vphi*vphi)
43     aphi = -2.*Gphirphi*vr*vphi
44     # update velocity:
45     vt = vt + dl*at
46     vr = vr + dl*ar
47     vphi = vphi + dl*aphi
48     # update position:
49     r = r + vr*dl
50     t = t + vt*dl
51     phi = phi + vphi*dl
52
53   # Direction of propagation, approximated in asymptotically flat coords.
54   # First, differentiate (x,y)=(r cos phi,r sin phi) to get vx and vy:
55   vx = vr*math.cos(phi)-r*math.sin(phi)*vphi
56   vy = vr*math.sin(phi)+r*math.cos(phi)*vphi
57   prop = math.atan2(vy,vx) # inverse tan of vy/vx, in the proper quadr
58   prop_sec = prop*180.*3600/math.pi
59   print "final direction of propagation = %6.2f arc-seconds" % prop_se
```

At line 14, we take the mass to be 1000 times greater than the mass of the sun. This helps to make the deflection easier to calculate accurately without running into problems with rounding errors. Lines 17-25 set up the initial conditions to be at the point of closest approach, as the photon is grazing the sun. This is easier to set up than initial conditions in which the photon approaches from far away. Because of this, the deflection angle calculated by the program is cut in half. Combining the factors of 1000 and one half, the final result from the program is to be interpreted as 500 times the actual deflection angle.

The result is that the deflection angle is predicted to be 870 seconds of arc. As a check, we can run the program again with $m = 0$; the result is a deflection of -8 seconds, which is a measure of the accumulated error due to rounding and the finite increment used for λ.

Dividing by 500, we find that the predicted deflection angle is 1.74 seconds, which, expressed in radians, is exactly $\frac{4Gm}{c^2 r}$. The unitless factor of 4 is in fact the correct result in the case of small deflections, i.e., for $\frac{m}{r} \ll 1$.

Although the numerical technique has the disadvantage that it doesn't let us directly prove a nice formula, it has some advantages as well. For one thing, we can use it to investigate cases for which the approximation $\frac{m}{r} \ll 1$ fails. For $\frac{m}{r} = 0.3$, the numerical technique gives a deflection of 222 degrees, whereas the weak-field approximation $\frac{4Gm}{c^2 r}$ gives only 69 degrees. What is happening here is that we're getting closer and closer to the event horizon of a black hole. Black holes are the topic of section 6.3, but it should be intuitively reasonable that something wildly nonlinear has to happen as we get close to the point where the light wouldn't even be able to escape.

The precision of Eddington's original test was only about ± 30%, and has never been improved on significantly with visible-light astronomy. A better technique is radio astronomy, which allows measurements to be carried out without waiting for an eclipse. One merely has to wait for the sun to pass in front of a strong, compact radio source such as a quasar. These techniques have now verified the deflection of light predicted by general relativity to a relative precision of about 10^{-5}.[9]

References

[6] Misner, Thorne, and Wheeler, *Gravitation*, p. 1118

[7] Rindler, *Essential Relativity*, 1969, p. 141

[9] For a review article on this topic, see Clifford Will, "The Confrontation between General Relativity and Experiment," relativity.livingreviews.org/...es/lrr-2006-3/.

6.4: Black Holes (Part 1)

Singularities

A provocative feature of the Schwarzschild metric is that it has elements that blow up at $r = 0$ and at $r = 2m$. If this is a description of the sun, for example, then these singularities are of no physical significance, since we only solved the Einstein field equation for the vacuum region outside the sun, whereas $r = 2m$ would lie about 3 km from the sun's center. Furthermore, it is possible that one or both of these singularities is nothing more than a spot where our coordinate system misbehaves. This would be known as a *coordinate singularity*. For example, the metric of ordinary polar coordinates in a Euclidean plane has $g^{\theta\theta} \to \infty$ as $r \to 0$.

One way to test whether a singularity is a coordinate singularity is to calculate a scalar measure of curvature, whose value is independent of the coordinate system. We can take the trace of the Ricci tensor, R_a^a, known as the *scalar curvature* or *Ricci scalar*, but since the Ricci tensor is zero, it's not surprising that that is zero. A different scalar we can construct is the product $R^{abcd} R_{abcd}$ of the Riemann tensor with itself. This is known as the *Kretschmann invariant*. The Maxima command lriemann(true) displays the nonvanishing components of R$_{abcd}$ The component that misbehaves the most severely at r = 0 is R$_{trrt}$ = $\frac{2m}{r^3}$. Because of this, the Kretschmann invariant blows up like r^{-6} as $r \to 0$. This shows that the singularity at r = 0 is a real, physical singularity.

The singularity at $r = 2m$, on the other hand, turns out to be only a coordinate singularity. To prove this, we have to use some technique other than constructing scalar measures of curvature. Even if every such scalar we construct is finite at $r = 2m$, that doesn't prove that every such scalar we *could* construct is also well behaved. We can instead search for some other coordinate system in which to express the solution to the field equations, one in which no such singularity appears. A partially successful change of coordinates for the Schwarzschild metric, found by Eddington in 1924, is (see problem 8):

$$t \to t' = t - 2m \ln(r - 2m). \tag{6.4.1}$$

This makes the covariant metric finite at $r = 2m$, although the contravariant metric still blows up there. A more complicated change of coordinates that completely eliminates the singularity at $r = 2m$ was found by Eddington and Finkelstein in 1958, establishing that the singularity was only a coordinate singularity. Thus, if an observer is so unlucky as to fall into a black hole, he will not be subjected to infinite tidal stresses — or infinite anything — at $r = 2m$. He may not notice anything special at all about his local environment. (Or he may already be dead because the tidal stresses at $r > 2m$, although finite, were nevertheless great enough to kill him.)

Event Horizon

Even though $r = 2m$ isn't a real singularity, interesting things do happen there. For $r < 2m$, the sign of g_{tt} becomes negative, while g_{rr} is positive. In our $+ - --$ signature, this has the following interpretation. For the world-line of a material particle, ds^2 is supposed to be the square of the particle's proper time, and it must always be positive. If a particle had a constant value of r, for $r < 2m$, it would have $ds^2 < 0$, which is impossible.

The timelike and spacelike characters of the r and t coordinates have been swapped, so r acts like a time coordinate.

Thus for an object compact enough that $r = 2m$ is exterior, $r = 2m$ is an event horizon: future light cones tip over so far that they do not allow causal relationships to connect with the spacetime outside. In relativity, event horizons do not occur only in the context of black holes; their properties, and some of the implications for black holes, have already been discussed in section 6.1.

The gravitational time dilation in the Schwarzschild field, relative to a clock at infinity, is given by the square root of the g_{tt} component of the metric. This goes to zero at the event horizon, meaning that, for example, a photon emitted from the event horizon will be infinitely redshifted when it reaches an observer at infinity. This makes sense, because the photon is then undetectable, just as it would be if it had been emitted from *inside* the event horizon.

Infalling Matter

If matter is falling into a black hole, then due to time dilation an observer at infinity "sees" that matter as slowing down more and more as it approaches the horizon. This has some counterintuitive effects. A radially infalling particle has

$$\frac{d^2 r}{dt^2} > 0 \tag{6.4.2}$$

once it falls past a certain point, which could be interpreted as a gravitational repulsion. The observer at infinity may also be led to describe the black hole as consisting of an empty, spherical shell of matter that never quite made it through the horizon. If asked what holds the shell up, the observer could say that it is held up by gravitational repulsion.

There is actually nothing wrong with any of this, but one should realize that it is only one possible description in one possible coordinate system. An observer hovering just outside the event horizon sees a completely different picture, with matter falling past at velocities that approach the speed of light as it comes to the event horizon. If an atom emits a photon from the event horizon, the hovering observer sees it as being infinitely red-shifted, but explains the red-shift as a kinematic one rather than a gravitational one.

We can imagine yet a third observer, one who free-falls along with the infalling matter. According to this observer, the gravitational field is always zero, and it takes only a finite time to pass through the event horizon.

If a black hole has formed from the gravitational collapse of a cloud of matter, then some of our observers can say that "right now" the matter is located in a spherical shell at the event horizon, while others can say that it is concentrated at an infinitely dense singularity at the center. Since simultaneity isn't well defined in relativity, it's not surprising that they disagree about what's happening "right now." Regardless of where they say the matter is, they all agree on the spacetime curvature. In fact, Birkhoff's theorem tells us that any spherically symmetric vacuum spacetime is Schwarzschild in form, so it doesn't matter where we say the matter is, as long as it's distributed in a spherically symmetric way and surrounded by vacuum.

A particularly nice way of summarizing and understanding these issues is with the use of a Penrose diagram, as discussed in section 7.3.

Expected Formation

Einstein and Schwarzschild did not believe, however, that any of these features of the Schwarzschild metric were more than a mathematical curiosity, and the term "black hole" was not invented until the 1967, by John Wheeler. There is quite a bit of evidence these days that our universe does contain objects that have undergone complete gravitational collapse, in the sense that their mass M is contained within a radius $r \lesssim M$ (in geometrized units). These objects are probably black holes, although doubts have been raised recently as to whether they are in fact other objects such as naked singularities.[10] Supposing that black holes do exist, there is also the question of what sizes they come in.

We might expect naively that since gravity is an attractive force, there would be a tendency for any primordial cloud of gas or dust to spontaneously collapse into a black hole. But clouds of less than about $0.1 M_\odot$ (0.1 solar masses) form planets, which achieve a permanent equilibrium between gravity and internal pressure. Heavier objects initiate nuclear fusion, but those with masses above about $100 M_\odot$ are immediately torn apart by their own solar winds. In the range from 0.1 to $100 M_\odot$, stars form. As discussed in section 4.4, those with masses greater than about a few $M_\odot$ are expected to form black holes when they die. We therefore expect, on theoretical grounds, that the universe should contain black holes with masses ranging from a few solar masses to a few tens of solar masses.

Observational Evidence

A black hole is expected to be a very compact object, with a strong gravitational field, that does not emit any of its own light. A bare, isolated black hole would be difficult to detect, except perhaps via its lensing of light rays that happen to pass by it. But if a black hole occurs in a binary star system, it is possible for mass to be transferred onto the black hole from its companion, if the companion's evolution causes it to expand into a giant and intrude upon the black hole's gravity well. The infalling gas would then get hot and emit radiation before disappearing behind the event horizon (Figure 6.4.1). The object known as Cygnus X-1 is the best-studied example. This X-ray-emitting object was discovered by a rocket-based experiment in 1964. It is part of a double-star system, the other member being a blue supergiant. They orbit their common center of mass with a period of 5.6 days. The orbit is nearly circular, and has a semimajor axis of about 0.2 times the distance from the earth to the sun. Applying Kepler's law of periods to these data constrains the sum of the masses, and knowledge of stellar structure fixes the mass of the supergiant. The result is that the mass of Cygnus X-1 is greater than about 10 solar masses, and this is confirmed by multiple methods. Since this is far above the Tolman-Oppenheimer-Volkoff limit, Cygnus X-1 is believed to be a black hole, and its X-ray emissions are interpreted as the radiation from the disk of superheated material accreting onto it from its companion. It is believed to have more than 90% of the maximum possible spin for a black hole of its mass.[11]
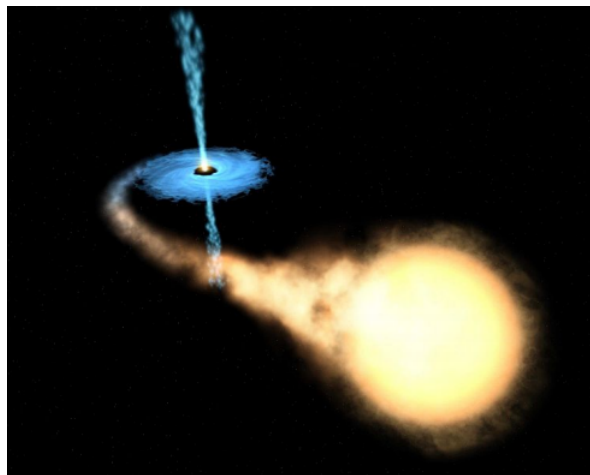
Figure 6.4.1: A black hole accretes matter from a companion star. (Public Domain; NASA).

Around the turn of the 21st century, new evidence was found for the prevalence of supermassive black holes near the centers of nearly all galaxies, including our own. Near our galaxy's center is an object called Sagittarius A*, detected because nearby stars orbit around it. The orbital data show that Sagittarius A* has a mass of about four million solar masses, confined within a sphere with a radius less than 2.2×10^7 km. There is no known astrophysical model that could prevent the collapse of such a compact object into a black hole, nor is there any plausible model that would allow this much mass to exist in equilibrium in such a small space, without emitting enough light to be observable.

The existence of supermassive black holes is surprising. Gas clouds with masses greater than about 100 solar masses cannot normally form stable stars, so supermassive black holes cannot be the end-point of the evolution of heavy stars. Mergers of multiple stars to form more massive objects are generally statistically unlikely, since a star is such a small target in relation to the distance between the stars. Once astronomers were confronted with the empirical fact of their existence, a variety of mechanisms was proposed for their formation. Little is known about which of these mechanisms is correct, although the existence of quasars in the early universe is interpreted as evidence that mass accreted rapidly onto supermassive black holes in the early stages of the evolution of the galaxies. As of 2016, an explanation getting a lot of attention is that in the early universe, there was a brief period in which the ambient conditions allowed the creation of supermassive black holes by direct collapse.[12]

A skeptic could object that although Cygnus X-1 and Sagittarius A* are more compact than is believed possible for a neutron star, this does not necessarily prove that they are black holes. Indeed, speculative theories have been proposed in which exotic objects could exist that are intermediate in compactness between black holes and neutron stars. These hypothetical creatures have names like black stars, gravastars, quark stars, boson stars, Q-balls, and electroweak stars. Although there is no evidence that these theories are right or that these objects exist, we are faced with the question of how to determine whether a given object is really a black hole or one of these other species. The defining characteristic of a black hole is that it has an event horizon rather than a physical surface. We currently have two ways of probing the structure of these stars at the radii where general relativity predicts the existence of an event horizon.

If an object is not a black hole, then by conservation of energy any matter that falls onto it must release its gravitational potential energy when it hits that surface. Cygnus X-1 has a copious supply of matter falling onto it from its supergiant companion, and Sagittarius A* likewise accretes a huge amount of gas from the stellar wind of nearby stars. By analyzing millimeter and infrared very-long-baseline-interferometry observations, Broderick, Loeb, and Narayan[13] have shown that if Sagittarius A* had a surface, then the luminosity of this surface must be less than 0.3% of the luminosity of the accretion disk. But this is not physically possible, because there are fundamental limits on the efficiency with which the gas can radiate away its energy before hitting the surface. We can therefore conclude that Sagittarius A* must have an event horizon. Its event horizon may be imaged directly in the near future.[14]

A second approach is through the observation of gravitational waves. As discussed in more detail in ch. 9, 2016 saw the first direct observation of gravitational waves. The waveform that was detected (Figure 9.2.2) fits very well with the predictions of general relativity for the merger of two black holes. It seems very unlikely that a waveform with this time-scale and characteristic shape could have been produced unless general relativity's description of black holes is correct in detail.

Singularities and Cosmic Censorship

Informal Ideas

Since we observe that black holes really do exist, maybe we should take the singularity at r = 0 seriously. Physically, it says that the mass density and tidal forces blow up to infinity there.

Generally when a physical theory says that observable quantities blow up to infinity at a particular point, it means that the theory has reached the point at which it can no longer make physical predictions. For instance, Maxwell's theory of electromagnetism predicts that the electric field blows up like r^{-2} near a point charge, and this implies that infinite energy is stored in the field within a finite radius around the charge. Physically, this can't be right, because we know it only takes 511 keV of energy to create an electron out of nothing, e.g., in nuclear beta decay. The paradox is resolved by quantum electrodynamics, which modifies the description of the vacuum around the electron to include a sea of virtual particles popping into and out of existence.

In the case of a black hole singularity, it is possible that quantum mechanical effects at the Planck scale prevent the formation of a singularity. Unfortunately, we are unlikely to find any empirical evidence about this, since black holes always seem to come clothed in event horizons, so we outside observers cannot extract any data about the singularity inside. Even if we take a suicidal trip into a black hole, we get no data about the singularity, because the singularity in the Schwarzschild metric is spacelike, not timelike, and therefore it always lies in our future light cone, never in our past.

In a way, the inaccessibility of singularities is a good thing. If a singularity exists, it is a point at which all the known laws of physics break down, and physicists therefore have no way of predicting anything about its behavior. There is likewise no great crisis for physics due to the Big Bang singularity or the Big Crunch singularity that occurs in some cosmologies in which the universe recollapses; we have no reasonable expectation of being able to make and test predictions or retrodictions that extend beyond the beginning or end of the universe.

What would be a crushing blow to the enterprise of physics would be a singularity that could sit on someone's desk. As John Earman of the University of Pittsburgh puts it, anything could pop out of such a "naked" singularity (defined formally later), including green slime or your lost socks.

Penrose's *cosmic censorship* conjecture states that the laws of physics prevent the formation of naked singularities from nonsingular and generic initial conditions. "Generic" is a necessary addition to Penrose's original 1969 formulation, since Choptuik showed in 1993 that certain perfectly fine-tuned initial conditions allowed collapse to a naked singularity.[15] As of 2017, evidence is accumulating that cosmic censorship is false. This is discussed at greater length later.

Formal Definitions

The remainder of this subsection provides a more formal exposition of the definitions relating to singularities. It can be skipped without loss of continuity.

The reason we care about singularities is that they indicate an incompleteness of the theory, and the theory's inability to make predictions. One of the simplest things we could ask any theory to do would be to predict the trajectories of test particles. For example, Maxwell's equations correctly predict the motion of an electron in a uniform magnetic field, but they fail to predict the motion of an electron that collides head-on with a positron. It might have been natural for someone in Maxwell's era (assuming they were informed about the existence of positrons and told to assume that both particles were pointlike) to guess that the two particles would scatter through one another at $\theta = 0$, their velocities momentarily becoming infinite. But it would have been equally natural for this person to refuse to make a prediction.

Similarly, if a particle hits a black hole singularity, we should not expect general relativity to make a definite prediction. It doesn't, because the geodesic equation breaks down.

We would therefore like to define a singularity as a situation in which the geodesics of test particles can't be extended indefinitely. But what does "indefinitely" mean? If the test particle is a photon, then the metric length of its world-line is zero. We get around this by defining length in terms of an affine parameter.

> **Definition**
>
> A spacetime is said to be geodesically incomplete if there exist timelike or lightlike geodesics that cannot be extended beyond some finite affine parameter into the past or future.

This is also a pretty good working definition of what we mean when we say that a spacetime contains a singularity, although it may not be optimal for all purposes.[16] The Schwarzschild spacetime has a singularity at r = 0, but not at the event horizon, since geodesics continue smoothly past the event horizon. Cosmological spacetimes contain a Big Bang singularity which prevents geodesics from being extended beyond a certain point in the *past*.

Actual singularities involving geodesic incompleteness are to be distinguished from coordinate singularities, which are not really singularities at all. In the Schwarzschild spacetime, as described in Schwarzschild's original coordinates, some components of the metric blow up at the event horizon, but this is not an actual singularity. This coordinate system can be replaced with a different one in which the metric is well behaved.

> ## Example 1: A harmless blow-up
>
> Let's define coordinates (t, y) in the region of spacetime where you're sitting and reading this book. Let (0, 0) be your current time and position, and for convenience let this be an inertial frame (so that your motion is not geodesic). The Riemann tensor, expressed in these coordinates, has a component $R_{tyyt} = \frac{2Gm}{r^3}$, where m and r are the mass and radius of the earth. This has the finite value of 1.5×10^{-6} s^{-2}, which expresses the strength of a tidal effect near the earth's surface.
>
> Now define a new coordinate u = y^3. Applying the tensor transformation law, we have
>
> $$R_{tuut} = R_{tyyt}\left(\frac{\partial y}{\partial u}\right)^2, \tag{6.4.3}$$
>
> which is infinite at y = 0. This example demonstrates that we cannot test for a singularity by looking for a blow-up of the components of a curvature tensor at certain coordinates or as the coordinates approach some limit.

There are two types of singularities: curvature singularities and non-curvature singularities.

The big bang and black hole singularities are examples of curvature singularities, which can often be recognized because there are scalar measures of curvature such as $R^{abcd}R_{abcd}$, known as the Kretschmann invariant, that blow up. These indicate that tidal forces blow up to infinity, and would destroy any observer.

The reason curvature scalars are useful as tests for a curvature singularity is that since they're scalars, they can't diverge in one coordinate system but stay finite in another (cf. example 1). A sufficient condition for a singularity to be a curvature singularity is if timelike or lightlike geodesics can only be extended to some finite affine parameter, and some curvature scalar (not necessarily every such scalar) approaches infinity as we approach this value of the affine parameter.

But we should not expect this to be a necessary condition for a curvature singularity. Example 2 below shows that the most commonly occurring curvature scalars may not be enough to catch the presence of a singularity. This is not too surprising, since curvature scalars do not suffice to tell us everything there is to know about the curvature of a spacetime (example 3).

> ## Example 2: Incompleteness with finite curvature scalars
>
> Consider the 1+1-dimensional spacetime described by the metric
>
> $$ds^2 = A(dt^2 - dx^2)$$
> $$A = \frac{1}{1+e^t},$$
>
> with $-\infty < x < \infty$ and $-\infty < t < \infty$. For large negative t it is indistinguishable from Minkowski space. The following Maxima code computes its Riemann tensor and the scalar curvature R and the Kretchmann invariant K.

```
1   load(ctensor);
2   dim:2;
3   ct_coords:[t,x];
4   u:1/(1+exp(t));
5   lg:matrix([u,0],
6      [0,-u]);
7   cmetric();
8   ricci(true);
9   lriemann(true);
10  uriemann(true);
11  scurvature();/* scalar curvature */
12  rinvariant(); /* Kretchmann */
```

The results for the two curvature scalars are

$$R = (1+e^{-t})^{-1} \qquad (6.4.4)$$

and

$$K = (1+e^{-t})^{-2}, \qquad (6.4.5)$$

both of which are finite everywhere; they go from 0 at large negative times to 1 at large positive times.

From these results we would not imagine that there was any singularity present, but blow-ups of curvature scalars are only a sufficient condition for geodesic incompleteness, not a necessary one. Consider the timelike curve x = 0, which by symmetry is a geodesic. If we integrate the proper time along this geodesic, we get a finite limit as t → ∞. Since proper time qualifies as an affine parameter, this geodesic is incomplete.

But it is not so obvious that this spacetime is "really" singular. It is possible that we could smoothly extend it beyond t = +∞. If so, then the singularity at t = +∞ would be a kind of fake singularity, of the type that we could obtain simply by chopping off the part of Minkowski space with t ≥ 0.

Example 3: Vanishing curvature scalars

We remarked above that curvature scalars do not in general suffice to tell us everything about the curvature of a spacetime. In fact, there is an entire class of curved spacetimes such that *every* curvature invariant vanishes everywhere. Schmidt[17] gives the example

$$ds^2 = dudv - a^2(u)dw^2, \qquad (6.4.6)$$

where a is an arbitrary nonlinear function. The eigenvalues of this metric are 1, −1, and −a^2, so its signature is + − −, i.e., this is general relativity in 2 + 1 dimensions. A computation shows that the space is not flat, since, e.g., $R_{uu} = -\frac{a''}{a}$. The u and v directions are lightlike, so this metric represents a wavelike disturbance traveling at the speed of light. (Since the Ricci tensor doesn't vanish, this isn't a vacuum solution, and we don't have a gravitational wave in vacuum. Such waves, as described in ch. 9, are transverse and can only exist in 3 + 1 or more dimensions.)

The lightlike character of u and v motivates us to consider coordinate transformations of the form (u, v) → (uD, $\frac{v}{D}$), because in the case a = 0, which is flat, this would be a Lorentz boost with a Doppler shift factor D. In the case where D approaches zero, we are chasing the wave at a velocity approaching c, so the wave Doppler-shifts to undetectability. All components of the Riemann tensor, as well as their derivatives, approach zero.

Now consider any curvature scalar I that is expressible as a continuous function of the Riemann tensor and its derivatives. By continuity, I approaches zero as D → 0. But curvature scalars are scalars, so they are invariant under coordinate transformations. It therefore follows that I = 0 identically, regardless of the value of D. Thus we have a spacetime that, although curved, has no nonvanishing curvature scalars anywhere.

---

[17] "Why do all the curvature invariants of a gravitational wave vanish?," arxiv.org/abs/gr-qc/9404037

References

[10] See sec. 6.3, and, e.g., Joshi et al., arxiv.org/abs/1304.7331.

[11] Gou et al., "The Extreme Spin of the Black Hole in Cygnus X-1," http://arxiv.org/abs/1106.3690

[12] See, e.g., http://arxiv.org/abs/1402.5675

[13] arxiv.org/abs/0903.1105

[14] arxiv.org/abs/0906.4040

[15] Phys. Rev. Lett. 70, p. 9

[16] Geroch, "What is a singularity in general relativity?," Ann Phys 48 (1968) 526.

6.5: Black Holes (Part 2)

Singularities can also occur without any blow-up in the curvature. An example of this is a conical singularity, Figure 6.3.2. (Cf. ⬚ Figure 5.9.2) In 2+1-dimensional relativity, curvature vanishes identically in the case of a vacuum, and the only kind of curvature singularity we can have is a non-curvature singularity. Another example of a non-curvature singularity is provided by the Taub-NUT family of spacetimes, in which some lightlike geodesics spiral in toward a horizon, but tidal forces do not blow up at the horizon. There is no clear reason to expect that non-curvature singularities could actually exist in our universe, but neither is there any proof that they cannot be formed by natural processes.
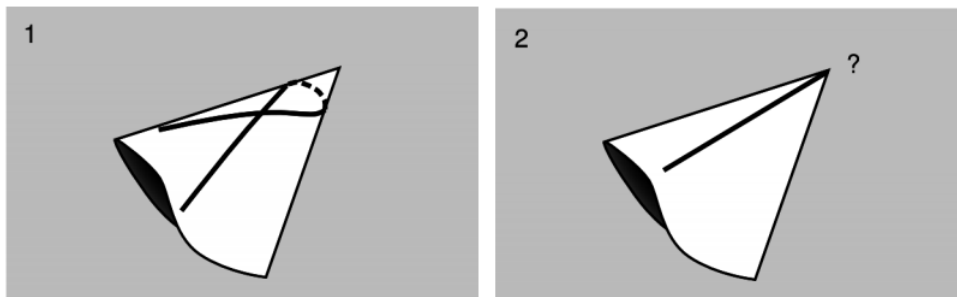

Figure 6.5.2: A conical singularity. The cone has zero intrinsic curvature everywhere except at its tip. Geodesic 1 can be extended infinitely far, but geodesic 2 cannot; since the metric is undefined at the tip, there is no sensible way to define how geodesic 2 should be extended.

Example 4: A non-curvature singularity

Consider the metric

$$ds^2 = \frac{1}{t}\,dt^2 - t\,d\theta^2 \tag{6.5.1}$$

in 1+1 dimensions, where θ is an angle running around the circle. This is a simplified version of a Taub-NUT spacetime. Lightlike geodesics have ds = 0, so $\frac{dt}{t} = \pm\,d\theta$, and θ = (const) $\pm$ ln($\pm$t), where the two signs can be chosen independently. Single out the geodesic $\theta = \ln t$, which is defined only for t > 0. It wraps around the circle infinitely many times as t goes to zero, and we would like to know whether it is incomplete there. If the affine parameter goes to infinity as t approaches zero, then the geodesic is not incomplete.

The nonvanishing Christoffel symbols are $\Gamma^t_{tt} = -\frac{1}{2t}, \Gamma^\theta_{\theta t} = \Gamma^\theta_{t\theta} = \frac{1}{2t}$, and $\Gamma^t_{\theta\theta} = \frac{t}{2}$ (Problem 3). The resulting geodesic equations are

$$\ddot{t} = \frac{\dot{t}^2}{2t} - \frac{t}{2}\dot{\theta}^2$$

$$\ddot{\theta} = -\frac{\dot{t}}{\dot{\theta}}t,$$

where dots represent differentiation with respect to the affine parameter λ. Implicit differentiation of the equation θ = ln t gives $\dot{\theta} = \frac{\dot{t}}{t}$, and plugging this in to the first geodesic equation results in $\ddot{t} = 0$. We can therefore take t = λ. (We could also take t = aλ + b, which would result in a different and equally valid affine parameter.) If λ had gone to $-\infty$ as t went to zero, then we would have demonstrated that the geodesic was complete. It approaches a finite limit instead, which suggests, but does not prove, that it is incomplete.

The change of coordinates $\theta \to \theta - \ln t$ allows the counterclockwise lightlike geodesics to be continued through t = 0. (Because this transformation is not a diffeomorphism, it is not just a renaming of points but an actual physical change in the structure of the spacetime; it is equivalent to cutting apart the halves with t < 0 and t > 0 and gluing them back together in a different way.) The corresponding geodesics in the clockwise direction, however, remain incomplete. A different change of coordinates extends the clockwise but not the counterclockwise ones. In all cases there are incomplete geodesics, so it still appears that we do have a singularity. Since curvature singularities don't exist in less than 3+1 dimensions, this is a non-curvature singularity. (A calculation also shows that this particular spacetime is flat.)

A singularity is not considered to be a point or set of points in a spacetime; it's more like a hole in the topology of the manifold. For example, the Big Bang didn't occur at a point or set of points. A singularity represents a breakdown in the metric, and without a metric we may not even be able to tell the difference between one point and many. For more on these issues, see the discussions of boundary constructions in section 7.3. There is a sense in which a black hole singularity is not a thing at all, and has no definable characteristics; section 7.4.

> ### Example 5: One point, or many?
>
> Suppose I have a two-dimensional space with coordinates (u, v), and I ask you whether S = {(u, v)|v = 0} is a point or a curve, while refusing to divulge what metric I have in mind. You'd probably say S was a curve, and if the metric was $ds^2 = du^2 + dv^2$, you'd be right. On the other hand, if the metric was $ds^2 = v^2 du^2 + dv^2$, S would be a point.
>
> This was an example where there were two possible metrics we could imagine. At a singularity, it's even worse. There is *no* possible metric that we can extend to the singularity.

Because a singularity isn't a point or a point-set, we can't define its timelike or spacelike character in quite the way we would with, say, a curve. A timelike singularity, also referred to as a locally naked singularity, is one such that an observer with a timelike worldline can have the singularity sometimes in his future light-cone and sometimes in his past light-cone.[18]

Schwarzschild and Big Bang singularities are spacelike. (Note that in the Schwarzschild metric, the Schwarzschild r and t coordinates swap their timelike and spacelike characters inside the event horizon.) The definition of a timelike singularity is local. A timelike singularity would be one that you could have sitting on your desk, where you could look at it and poke it with a stick.

A naked singularity is one from which timelike or lightlike worldlines can originate and then escape to infinity. The Schwarzschild metric's singularity is not naked. This notion is global.

Evidence Accumulating Against Cosmic Censorship

As of 2017, evidence is accumulating that cosmic censorship is false. Back in 1969 when Roger Penrose first formulated the hypothesis, relativists had been strongly influenced by a 1939 calculation by Oppenheimer and Snyder for the gravitational collapse of a uniform, spherical cloud of "dust," meaning material particles that act like a pressureless ideal fluid (see example 1). (Cf. section 4.4 on the Tolman-Oppenheimer-Volkoff limit, derived earlier the same year.) Even though Oppenheimer and Snyder were too timid to continue the calculation past the formation of an event horizon, their result was not taken seriously for years, the notion of a runaway gravitational collapse being too distant from the state of the art in terms of observation. But later workers did complete the calculation. They found that a singularity developed, but that the horizon formed early enough to cloak it, so that no timelike or lightlike geodesic from the singularity could escape to a distant observer. This was consistent with a weak version of the cosmic censorship hypothesis, that a (globally) naked singularity cannot form from gravitational collapse.

But to interpret the result as evidence for cosmic censorship was misleading. With hindsight, there are clear Newtonian reasons to suspect that a perfectly homogeneous cloud has properties that are a little too special. In the Newtonian version the internal gravitational field is proportional to r. Starting from rest at r, a particle has to travel a distance r to reach the center, but since the acceleration is proportional to r, the time needed to reach the center is the same for all particles. There is a Newtonian singularity of infinite density, and this occurs at the *same* time for all particles, which is after the formation of a surface from which the escape velocity has any fixed value, such as c. Therefore in Newtonian terms, cosmic censorship holds, but it holds *only* because of the perfect homogeneity of the cloud.

In fact, the general-relativistic version of inhomogeneous gravitational collapse had already been worked out around 1933 by Lemaître, Tolman, and Bondi, again for the case of a spherical cloud, but now with a density profile $\rho(r)$. This family of metrics, called the Lemaître-Tolman-Bondi metrics, is general enough to include models of cosmological expansion as well as models of local gravitational collapse. Tolman applied the collapse model to the formation of "nebulae," i.e., galaxies, in the early universe, but did not follow the evolution of the collapse to its ultrarelativistic *dénouement*, as Oppenheimer and Snyder had. When one does so, dealing with some technical obstacles and imposing some constraints for physical reasonableness, it turns out that in *most* cases, the result is a locally naked singularity.[19] That is, fine-tuning is required in order to produce something more like a standard black hole. It remains to be seen whether this holds true when the constraint of perfect spherical symmetry is relaxed.

This does not necessarily mean on the face of it that cosmic censorship is dead, since spacetimes with spherical symmetry are themselves finely tuned in some sense, but it is rather a dramatic development, since people had imagined for 75 years, based on the Oppenheimer-Snyder calculations for homogeneous dust, that a black hole was the generic result of runaway gravitational

collapse. Cosmic censorship is in a sense impossible to disprove, since part of the research program is to find the most appropriate definition of the conjecture, but these results suggest that if it is to be true, then it has to be weakened so much as to be of little interest. In general, a meaningful definition of what it means to violate weak cosmic censorship should probably include something like the following ingredients.

1. The initial conditions do not make available an infinite amount of energy within a finite region.
2. The initial conditions do not contain singularities.
3. Incomplete lightlike geodesics can arrive at a distant observer.
4. Such a violation still occurs if we impose small perturbations on the initial data.
5. The forms of matter are physically realistic.

If we do not impose something like condition 1, then we can set up initial conditions that are of no interest because they are unrealistic. For this reason, one usually studies spacetimes that are asymptotically flat.[20] Condition 2 expresses the idea that any singularities that occur should be new ones formed by gravitational collapse. The censorship violation is expressed by condition 3. The notion of a distant observer can be further formalized by requiring that such a geodesic arrive at null infinity, I +; see p. 265. If 4 is omitted, then clear counterexamples to censorship are known. However, it is not known whether there is an appropriately rigorous way to define "small perturbations" here.[21] Realistic matter fields, 5, are expected, for example, not to have negative mass.[22]

Because weak cosmic censorship seems to be violated if described by these five conditions, people have started looking for additional conditions that could salvage the conjecture.

Wald[23] suggests adding a sixth requirement. He proposes that the types of matter be further restricted to ones having the property that if the metric is fixed, rather than dynamical as in general relativity, then no singularities occur. This seems to me to be much too strong a condition, and there are indications that it is not sufficient.

> **Note**
>
> [20] Asymptotic flatness was introduced informally in section 4.5 and is defined in detail in section 7.4. It may also be necessary to impose a requirement that the matter fields fall off at some rate as we go to infinity.
>
> [21] In technical terms, we do not have any topology or measure defined on the set of all possible initial conditions. In actual work to date, people have selected some set of possible initial conditions, described by some small number of adjustable parameters, and have then tried to test condition 4 using a seemingly natural topology and measure defined on the space of those parameters.
>
> [22] More rigorously, we expect them to satisfy suitable energy conditions, section 8.1.

Another proposal is along the following lines. When a naked singularity occurs, then we have a region of spacetime for which the singularity is inside the past lightcone. The lightlike surface constituting the boundary of this region is called a Cauchy horizon. An observer who passes beyond the Cauchy horizon can observe arbitrary information, i.e., phenomena not predicted by any laws of physics, and infinite fluxes of energy. Roger Penrose has, however, pointed out that in certain illustrative cases, there is a tendency for energy from the entire spacetime prior to the singularity to be focused onto the Cauchy horizon. The result could then be that such an observer is destroyed when passing through the Cauchy horizon. In other words, the Cauchy horizon actually turns into a singularity. Penrose's mechanism appears to fail, however, for a spacetime with a positive cosmological constant, which is what we actually have in our universe.

Hawking Radiation

Radiation from Black Holes

Since event horizons are expected to emit blackbody radiation, a black hole should not be entirely black; it should radiate. This is called Hawking radiation. Suppose observer B just outside the event horizon blasts the engines of her rocket ship, producing enough acceleration to keep from being sucked in. By the equivalence principle, what she observes cannot depend on whether the acceleration she experiences is actually due to a gravitational field. She therefore detects radiation, which she interprets as coming from the event horizon below her. As she gets closer and closer to the horizon, the acceleration approaches infinity, so the intensity and frequency of the radiation grows without limit.

A distant observer A, however, sees a different picture. According to A, B's time is extremely dilated. A sees B's acceleration as being only $\sim 1/m$, where m is the mass of the black hole; A does not perceive this acceleration as blowing up to infinity as B approaches the horizon. When A detects the radiation, it is extremely red-shifted, and it has the spectrum that one would expect for a horizon characterized by an acceleration $a \sim 1/m$. The result for a 10-solar-mass black hole is $T \sim 10^{-8}$ K, which is so low that the black hole is actually absorbing more energy from the cosmic microwave background radiation than it emits.

Direct observation of black-hole radiation is therefore probably only possible for black holes of very small masses. These may have been produced soon after the big bang, or it is conceivable that they could be created artificially, by advanced technology. If blackhole radiation does exist, it may help to resolve the information paradox, since it is possible that information that goes into a black hole is eventually released via subtle correlations in the black-body radiation it emits.

Particle Physics

Hawking radiation has some intriguing properties from the point of view of particle physics. In a particle accelerator, the list of particles one can create in appreciable quantities is determined by coupling constants. In Hawking radiation, however, we expect to see a representative sampling of all types of particles, biased only by the fact that massless or low-mass particles are more likely to be produced than massive ones. For example, it has been speculated that some of the universe's dark matter exists in the form of "sterile" particles that do not couple to any force except for gravity. Such particles would never be produced in particle accelerators, but would be seen in Hawking radiation. Based on present knowledge of particle physics, the main components of Hawking radiation, for all but the most microscopic black holes, are expected to be photons and gravitons, which would compete on roughly equal terms, depending on the angular momentum of the black hole.[24]

Hawking radiation would violate many cherished conservation laws of particle physics. Let a hydrogen atom fall into a black hole. We've lost a lepton and a baryon, but if we want to preserve conservation of lepton number and baryon number, we cover this up with a fig leaf by saying that the black hole has simply increased its lepton number and baryon number by +1 each. But eventually the black hole evaporates, and the evaporation is probably mostly into zero-mass particles such as photons. Once the hole has evaporated completely, our fig leaf has evaporated as well. There is now no physical object to which we can attribute the +1 units of lepton and baryon number.

Black-hole Complementarity

A very difficult question about the relationship between quantum mechanics and general relativity occurs as follows. In our example above, observer A detects an extremely red-shifted spectrum of light from the black hole. A interprets this as evidence that the space near the event horizon is actually an intense maelstrom of radiation, with the temperature approaching infinity as one gets closer and closer to the horizon. If B returns from the region near the horizon, B will agree with this description. But suppose that observer C simply drops straight through the horizon. C does not feel any acceleration, so by the equivalence principle C does not detect any radiation at all. Passing down through the event horizon, C says, "A and B are liars! There's no radiation at all." A and B, however, C see as having entered a region of infinitely intense radiation. "Ah," says A, "too bad. C should have turned back before it got too hot, just as I did." This is an example of a principle we've encountered before, that when gravity and quantum mechanics are combined, different observers disagree on the number of quanta present in the vacuum. We are presented with a paradox, because A and B believe in an entirely different version of reality that C. A and B say C was fricasseed, but C knows that that didn't happen. One suggestion is that this contradiction shows that the proper logic for describing quantum gravity is nonaristotelian, as described in section 2.3. This idea, suggested by Susskind et al., goes by the name of *black-hole complementarity*, by analogy with Niels Bohr's philosophical description of wave-particle duality as being "complementary" rather than contradictory. In this interpretation, we have to accept the fact that C experiences a qualitatively different reality than A and B, and we comfort ourselves by recognizing that the contradiction can never become too acute, since C is lost behind the event horizon and can never send information back out.

Black Holes in d Dimensions

It has been proposed that our universe might actually have not d = 4 dimensions but some higher number, with the d − 4 "extra" ones being spacelike, and curled up on some small scale ρ so that we don't see them in ordinary life. One candidate for such a scale ρ is the *Planck length*, and we then have to talk about theories of quantum gravity such as string theory. On the other hand, it could be the 1 TeV electroweak scale; the motivation for such an idea is that it would allow the unification of electroweak interactions with gravity. This idea goes by the name of "large extra dimensions" — "large" because ρ is bigger than the Planck length. In fact, in such theories the Planck length is the electroweak unification scale, and the number normally referred to as the Planck length is not really the Planck length.[25]

In d dimensions, there are d−1 spatial dimensions, and a surface of spherical symmetry has d − 2. In the Newtonian weak-field limit, the density of gravitational field lines falls off like $\frac{m}{r^{d-2}}$ with distance from a source m, and we therefore find that Newton's law of gravity has an exponent of −(d − 2). If d ≠ 3, we can integrate to find that the gravitational potential varies as $\Phi \sim -mr^{-(d-3)}$. Passing back to the weak-field limit of general relativity, the equivalence principle dictates that the g_{tt} term of the metric be approximately $1 + 2\Phi$, so we find that the metric has the form

$$ds^2 \approx (1 - 2mr^{-(d-3)})dt^2 - (\ldots)dr^2 - r^2 d\theta^2 - r^2 \sin^2\theta\, d\phi^2. \qquad (6.5.2)$$

This looks like the Schwarzschild form with no other change than a generalization of the exponent, and in fact Tangherlini showed in 1963 that for d > 4, one obtains the exact solution simply by applying the same change of exponent to g_{rr} as well.[26]

If large extra dimensions do exist, then this is the actual form of any black-hole spacetime for r << ρ, where the background curvature of the extra dimensions is negligible. Since the exponents are all changed, gravitational forces become stronger than otherwise expected at small distances, and it becomes easier to make black holes. It has been proposed that if large extra dimensions exist, microscopic black holes would be observed at the Large Hadron Collider. They would immediately evaporate into Hawking radiation, with an experimental signature of violating the standard conservation laws of particle physics. As of 2010, the empirical results seem to be negative.[27]

The reasoning given above fails in the case of d = 3, i.e., 2+1- dimensional spacetime, both because the integral of r^{-1} is not r^0 and because the Tangherlini-Schwarzschild metric is not a vacuum solution. As shown in problem 12, there is no counterpart of the Schwarzschild metric in 2+1 dimensions. This is essentially because for d = 3 mass is unitless, so given a source having a certain mass, there is no way to set the distance scale at which Newtonian weak-field behavior gives way to the relativistic strong field. Whereas for d ≥ 4, Newtonian gravity is the limiting case of relativity, for d = 3 they are unrelated theories. In fact, the relativistic theory of gravity for d = 3 is somewhat trivial. Spacetime does not admit curvature in vacuum solutions,[28] so that the only nontrivial way to make non-Minkowski 2+1-dimensional spacetimes is by gluing together Minkowski pieces in various topologies, like gluing pieces of paper to make things like cones and M¨obius strips. 2+1-dimensional gravity has conical singularities, but not Schwarzschild-style ones that are surrounded by curved spacetime.

If black-hole solutions exist in d dimensions, then one can extend such a solution to d+1 dimensions with cylindrical symmetry, forming a "black string." The nonexistence of d = 3 black holes implies that black string solutions do not exist in our own d = 4 universe. However, different considerations arise in a universe with a negative cosmological constant. There are then 2+1-dimensional solutions known as BTZ black holes.[29] Since our own universe has a positive cosmological constant, not a negative one, we still find that black strings cannot exist.

References

[18] Penrose, Gravitational radiation and gravitational collapse; Proceedings of the Symposium, Warsaw, 1973. Dordrecht, D. Reidel Publishing Co. pp. 82-91, free online at adsabs.harvard.edu/full/1974IAUS...64...82P

[19] Joshi and Malafarina, arxiv.org/abs/1405.1146

[23] "Gravitational Collapse and Cosmic Censorship," arxiv.org/abs/gr-qc/ 9710068

[24] Dong, arxiv.org/abs/1511.05642

[25] Kanti, arxiv.org/abs/hep-ph/0402168

[26] Emparan and Reall, "Black Holes in Higher Dimensions," relativity. livingreviews.org/Articles/lrr-2008-6/

[27] http://arxiv.org/abs/1012.3375

[28] arxiv.org/abs/gr-qc/0503022v4

[29] arxiv.org/abs/gr-qc/9506079v1

6.6: Degenerate Solutions

This section can be omitted on a first reading.

At the event horizon of the Schwarzschild spacetime, the timelike and spacelike roles of the Schwarzschild r and t coordinates get swapped around, so that the signs in the metric change from $+---$ to $-+--$. In discussing cases like this, it becomes convenient to define a new usage of the term "signature," as s = p − q, where p is the number of positive signs and q the number of negative ones. This can also be represented by the pair of numbers (p, q). The example of the Schwarzschild horizon is not too disturbing, both because the funny behavior arises at a singularity that can be removed by a change of coordinates and because the signature stays the same. An observer who free-falls through the horizon observes that the local properties of spacetime stay the same, with |s| = 2, as required by the equivalence principle.

But this only makes us wonder whether there are other examples in which an observer would actually detect a change in the metric's signature. We are encouraged to think of the signature as something empirically observable because, for example, it has been proposed that our universe may have previously unsuspected additional spacelike dimensions, and these theories make testable predictions. Since we do not notice the extra dimensions in ordinary life, they would have to be wrapped up into a cylindrical topology. Some such theories, like string theory, are attempts to create a theory of quantum gravity, so the cylindrical radius is assumed to be on the order of the Planck length, which corresponds quantum-mechanically to an energy scale that we will not be able to probe using any foreseeable technology. But it is also possible that the radius is large — a possibility that goes by the name of "large extra dimensions" — so that we could see an effect at the Large Hadron Collider. Nothing in the formulation of the Einstein field equations requires a 3+1 (i.e., (1, 3)) signature, and they work equally well if the signature is instead 4+1, 5+1, ... Newton's inverse-square law of gravity is described by general relativity as arising from the three-dimensional nature of space, so on small scales in a theory with n large extra dimensions, the $1/r^2$ behavior changes over to $1/r^{2+n}$, and it becomes possible that the LHC could produce microscopic black holes, which would immediately evaporate into Hawking radiation in a characteristic way.

So it appears that the signature of spacetime is something that is not knowable *a priori*, and must be determined by experiment. When a thing is supposed to be experimentally observable, general relativity tells us that it had better be coordinate-independent. Is this so? A proposition from linear algebra called *Sylvester's law of inertia* encourages us to believe that it is. The theorem states that when a real matrix A is diagonalized by a real, nonsingular change of basis (a similarity transformation $S^{-1}AS$), the number of positive, negative, and zero diagonal elements is uniquely determined. Since a change of coordinates has the effect of applying a similarity transformation on the metric, it appears that the signature is coordinate-independent.

This is not quite right, however, as shown by the following paradox. The coordinate invariance of general relativity tells us that if all clocks, everywhere in the universe, were to slow down simultaneously (with simultaneity defined in any way we like), there would be no observable consequences. This implies that the spacetime

$$ds^2 = -t\,dt^2 - d\ell^2, \tag{6.6.1}$$

where

$$d\ell^2 = dx^2 + dy^2 + dz^2, \tag{6.6.2}$$

is empirically indistinguishable from a flat spacetime. Starting from t = −∞, the positive g_{tt} component of the metric shrinks uniformly, which should be harmless. We can indeed verify by direct evaluation of the Riemann tensor that this is a flat spacetime (problem 10). But for t > 0 the signature of the metric switches from $+---$ to $----$, i.e., from Lorentzian (|s| = 2) to Euclidean (|s| = 4). This is disquieting. For t < 0, the metric is a perfectly valid description of our own universe (which is approximately flat). Time passes, and there is no sign of any impending disaster. Then, suddenly, at some point in time, the entire structure of spacetime undergoes a horrible spasm. This is a paradox, because we could just as well have posed our initial conditions using some other coordinate system, in which the metric had the familiar form $ds^2 = dt^2 - d\ell^2$. General relativity is supposed to be agnostic about coordinates, but a choice of coordinate leads to a differing prediction about the signature, which is a coordinate-independent quantity.

We are led to the resolution of the paradox if we explicitly construct the coordinate transformation involved. In coordinates (t, x, y, z), we have $ds^2 = -t\,dt^2 - d\ell^2$. We would like to find the relationship between t and some other coordinate u such that we recover the familiar form $ds^2 = du^2 - d\ell^2$ for the metric. The tensor transformation law gives

$$g_{tt} = \left(\frac{\partial u}{\partial t} \right)^2 g_{uu}$$

$$-t = \left(\frac{\partial u}{\partial t} \right)^2$$

with solution

$$u = \pm \frac{2}{3} t^{\frac{3}{2}}, \quad t < 0. \tag{6.6.3}$$
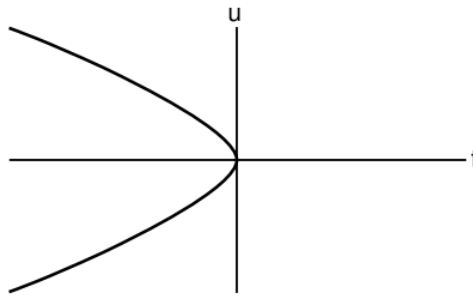
There is no solution for t > 0.

Figure 6.6.1 - The change of coordinates is degenerate at t = 0.

If physicists living in this universe, at t < 0, for some reason choose t as their time coordinate, there is in fact a way for them to tell that the cataclysmic event at t = 0 is not a reliable prediction. At t = 0, their metric's time component vanishes, so its signature changes from + − −− to 0 − −−. At that moment, the machinery of the standard tensor formulation of general relativity breaks down. For example, one can no longer raise indices, because g^{ab} is the matrix inverse of g_{ab}, but g_{ab} is not invertible. Since the field equations are ultimately expressed in terms of the metric using machinery that includes raising and lowering of indices, there is no way to apply them at t = 0. They do not make a false prediction of the end of the world; they fail to make any prediction at all. Physicists accustomed to working in terms of the t coordinate can simply throw up their hands and say that they have no way to predict anything at t > 0. But they already know that their spacetime is one whose observables, such as curvature, are all constant with respect to time, so they should ask why this perfect symmetry is broken by singling out t = 0. There is physically nothing that should make one moment in time different than any other, so choosing a particular time to call t = 0 should be interpreted merely as an arbitrary choice of the placement of the origin of the coordinate system. This suggests to the physicists that all of the problems they've been having are not problems with any physical meaning, but merely problems arising from a poor choice of coordinates. They carry out the calculation above, and discover the u time coordinate. Expressed in terms of u, the metric is well behaved, and the machinery of prediction never breaks down.

The paradox posed earlier is resolved because Sylvester's law of inertia only applies to a *nonsingular* transformation S. If S had been singular, then the S^{-1} referred to in the theorem wouldn't even have existed. But the transformation from u to t has $\frac{\partial t}{\partial u}$ = 0 at u = t = 0, so it is singular. This is all in keeping with the general philosophy of coordinate-invariance in relativity, which is that only smooth, one-to-one coordinate transformations are allowed. Someone who has found a lucky coordinate like u, and who then contemplates transforming to t, should realize that it isn't a good idea, because the transformation is not smooth and one-to-one. Someone who has started by working with an unlucky coordinate like t finds that the machinery breaks down at t = 0, and concludes that it would be a good idea to search for a more useful set of coordinates. This situation can actually arise in practical calculations.

What about our original question: could the signature of spacetime actually change at some boundary? The answer is now clear. Such a change of signature is something that could conceivably have intrinsic physical meaning, but if so, then the standard formulation of general relativity is not capable of making predictions about it. There are other formulations of general relativity, such as Ashtekar's, that are ordinarily equivalent to Einstein's, but that are capable of making predictions about changes of signature. However, there is more than one such formulation, and they do not agree on their predictions about signature changes.

6.E: Vacuum Solutions (Exercises)

1. Show that in geometrized units, power is unitless. Find the equivalent in watts of a power that equals 1 in geometrized units.

2. The metric of coordinates (θ, ϕ) on the unit sphere is $ds^2 = d\theta^2 + \sin^2\theta d\phi^2$. (a) Show that there is a singular point at which $g^{ab} \to \infty$. (b) Verify directly that the scalar curvature R = R^a_a constructed from the trace of the Ricci tensor is never infinite. (c) Prove that the singularity is a coordinate singularity.

3. (a) Space probes in our solar system often use a slingshot maneuver. In the simplest case, the probe is scattered gravitationally through an angle of 180 degrees by a planet. Show that in some other frame such as the rest frame of the sun, in which the planet has speed u toward the incoming probe, the maneuver adds 2u to the speed of the probe. (b) Suppose that we replace the planet with a black hole, and the space probe with a light ray. Why doesn't this accelerate the ray to a speed greater than c?

4. An observer outside a black hole's event horizon can never observe a test particle falling past the event horizon and later hitting the singularity. We could therefore wonder whether general relativity's predictions about the interior of a black hole, and the singularity in particular, are even a testable scientific theory. However, the observer could herself fall into the black hole. The question is then whether she would reach the singularity within a finite proper time; if so, then it is observable to her. The purpose of this problem is to prove that this is so, using the techniques of section 6.2. Suppose for simplicity that the observer starts at rest far away from the black hole, and falls directly inward toward it. (a) In the notation of section 6.2, what are the values of E and L in this case? (b) Find the function r(s), i.e., the observer's Schwarzschild radial coordinate as a function of her proper time, and show that she does reach the singularity in finite proper time.

5. The curve given parametrically by $(\cos^3 t, \sin^3 t)$ is called an astroid. The arc length along this curve is given by s = $(\frac{3}{2})\sin^2 t$, and its curvature by k = $-(\frac{2}{3})\csc^2 t$. By rotating this astroid about the x axis, we form a surface of revolution that can be described by coordinates (t, ϕ), where ϕ is the angle of rotation. (a) Find the metric on this surface. (b) Identify any singularities, and classify them as coordinate or intrinsic singularities.

6. (a) Section 3.5 gave a flat-spacetime metric in rotating polar coordinates,

$$ds^2 = (1 - \omega^2 r^2)dt^2 - dr^2 - r^2 d\theta'^2 - 2\omega r^2 d\theta' dt. \qquad (6.E.1)$$

Identify the two values of r at which singularities occur, and classify them as coordinate or non-coordinate singularities.
(b) The corresponding spatial metric was found to be

$$ds^2 = -dr^2 - \frac{r^2}{1 - \omega^2 r^2}d\theta'^2. \qquad (6.E.2)$$

Identify the two values of r at which singularities occur, and classify them as coordinate or non-coordinate singularities.
(c) Consider the following argument, which is intended to provide an answer to part b without any computation. In two dimensions, there is only one measure of curvature, which is equivalent (up to a constant of proportionality) to the Gaussian curvature. The Gaussian curvature is proportional to the angular deficit ϵ of a triangle. Since the angular deficit of a triangle in a space with negative curvature satisfies the inequality $-\pi < \epsilon < 0$, we conclude that the Gaussian curvature can never be infinite. Since there is only one measure of curvature in a two-dimensional space, this means that there is no non-coordinate singularity. Is this argument correct, and is the claimed result consistent with your answers to part b?

7. The first experimental verification of gravitational redshifts was a measurement in 1925 by W.S. Adams of the spectrum of light emitted from the surface of the white dwarf star Sirius B. Sirius B has a mass of $0.98 M_{\odot}$ and a radius of 5.9×10^6 m. Find the redshift.

8. Show that, as claimed in section 6.3, applying the change of coordinates t' = t−2m ln(r−2m) to the Schwarzschild metric results in a metric for which g_{rr} and $g_{t't'}$ never blow up, but that $g^{t't'}$ does blow up.

9. Use the geodesic equation to show that, in the case of a circular orbit in a Schwarzschild metric, $\frac{d^2t}{ds^2}$ = 0. Explain why this makes sense.

10. Verify by direct calculation, as asserted in section 6.4, that the Riemann tensor vanishes for the metric $ds^2 = -t\, dt^2 - d\ell^2$, where $d\ell^2 = dx^2 + dy^2 + dz^2$.

11. Suppose someone proposes that the vacuum field equation of general relativity isn't R_{ab} = 0 but rather R_{ab} = k, where k is some constant that describes an innate tendency of spacetime to have tidal distortions. Explain why this is not a good proposal.

12. Prove, as claimed in section 6.3, that in 2+1 dimensions, with a vanishing cosmological constant, there is no nontrivial Schwarzschild metric.

13. In section 6.2, I argued that there is no way to define a time-reversal operation in general relativity so that it applies to all spacetimes. Why can't we define it by picking some arbitrary spacelike surface that covers the whole universe, flipping the

velocity of every particle on that surface, and evolving a new version of the spacetime backward and forward from that surface using the field equations?

This page titled 6.E: Vacuum Solutions (Exercises) is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Benjamin Crowell via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER OVERVIEW

7: Symmetries

This chapter is not required in order to understand the later material.

Thumbnail: Penrose diagram for flat spacetime.

7.1: Killing Vectors

The Schwarzschild metric is an example of a highly symmetric spacetime. It has continuous symmetries in space (under rotation) and in time (under translation in time). In addition, it has discrete symmetries under spatial reflection and time reversal. In Section 6.2, we saw that the two continuous symmetries led to the existence of conserved quantities for the trajectories of test particles, and that these could be interpreted as mass-energy and angular momentum.

Generalizing, we want to consider the idea that a metric may be invariant when every point in spacetime is systematically shifted by some infinitesimal amount. For example, the Schwarzschild metric is invariant under $t \rightarrow t + dt$. In coordinates

$$(x^0, x^1, x^2, x^3) = (t, r, \theta, \phi), \tag{7.1.1}$$

we have a vector field $(dt, 0, 0, 0)$ that defines the time-translation symmetry, and it is conventional to split this into two factors, a finite vector field $\boldsymbol{\xi}$ and an infinitesimal scalar, so that the displacement vector is

$$\boldsymbol{\xi} dt = (1, 0, 0, 0) dt. \tag{7.1.2}$$

Such a field is called a *Killing vector field*, or simply a *Killing vector*, after Wilhelm Killing. When all the points in a space are displaced as specified by the Killing vector, they flow without expansion or compression. The path of a particular point, such as the dashed line in Figure 7.1.1, under this flow is called its orbit. Although the term "Killing vector" is singular, it refers to the entire field of vectors, each of which differs in general from the others. For example, the $\boldsymbol{\xi}$ shown in Figure 7.1.1 has a greater magnitude than a $\boldsymbol{\xi}$ near the neck of the surface.
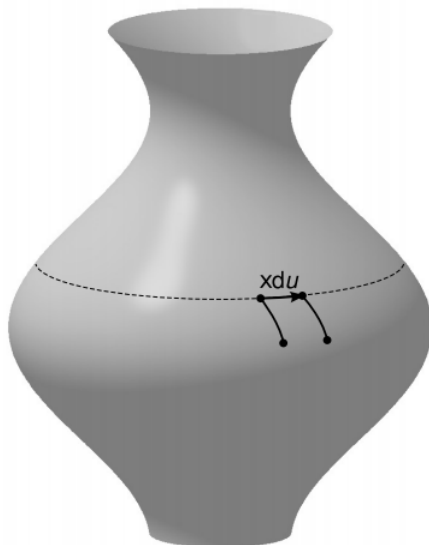
Figure 7.1.1: The two-dimensional space has a symmetry which can be visualized by imagining it as a surface of revolution embedded in three-space. Without reference to any extrinsic features such as coordinates or embedding, an observer on this surface can detect the symmetry, because there exists a vector field ξ du such that translation by ξ du doesn't change the distance between nearby points.

Figure 7.1.2: Wilhelm Killing (1847-1923).

The infinitesimal notation is designed to describe a continuous symmetry, not a discrete one. For example, the Schwarzschild spacetime also has a discrete time-reversal symmetry t → −t. This can't be described by a Killing vector, because the displacement in time is not infinitesimal.

> ### Example 7.1.1: The Euclidean plane
>
> The Euclidean plane has two Killing vectors corresponding to translation in two linearly independent directions, plus a third Killing vector for rotation about some arbitrarily chosen origin O. In Cartesian coordinates, one way of writing a complete set of these is is
>
> $$\xi_1 = (1, 0)$$
> $$\xi_2 = (0, 1)$$
> $$\xi_3 = (-y, x).$$
>
> A theorem from classical geometry[1] states that any transformation in the Euclidean plane that preserves distances and handedness can be expressed either as a translation or as a rotation about some point. The transformations that do not preserve handedness, such as reflections, are discrete, not continuous. This theorem tells us that there are no more Killing vectors to be found beyond these three, since any translation can be accomplished using ξ_1 and ξ_2, while a rotation about a point P can be done by translating P to O, rotating, and then translating O back to P.
>
> ---
>
> [1] Coxeter, *Introduction to Geometry*, ch. 3

In the example of the Schwarzschild spacetime, the components of the metric happened to be independent of t when expressed in our coordinates. This is a sufficient condition for the existence of a Killing vector, but not a necessary one. For example, it is possible to write the metric of the Euclidean plane in various forms such as

$$ds^2 = dx^2 + dy^2 \qquad (7.1.3)$$

and

$$ds^2 = dr^2 + r^2 \, d\phi^2. \qquad (7.1.4)$$

The first form is independent of x and y, which demonstrates that x → x + dx and y → y + dy are Killing vectors, while the second form gives us $\phi \to \phi + d\phi$. Although we may be able to find a particular coordinate system in which the existence of a Killing vector is manifest, its existence is an intrinsic property that holds regardless of whether we even employ coordinates. In general, we define a Killing vector not in terms of a particular system of coordinates but in purely geometrical terms: a space has a Killing vector $\boldsymbol{\xi}$ if translation by an infinitesimal amount $\boldsymbol{\xi}$ du doesn't change the distance between nearby points. Statements such as "the spacetime has a timelike Killing vector" are therefore intrinsic, since both the timelike property and the property of being a Killing vector are coordinate-independent.
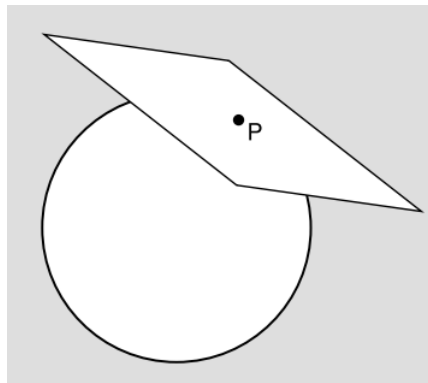
Figure $7.1.3$: Vectors at a point P on a sphere can be visualized as occupying a Euclidean plane that is particular to P.

Killing vectors, like all vectors, have to live in some kind of vector space. On a manifold, this vector space is particular to a given point, Figure $7.1.3$. A different vector space exists at every point, so that vectors at different points, occupying different spaces, can be compared only by parallel transport. Furthermore, we really have *two* such spaces at a given point, a space of contravariant vectors and a space of covariant ones. These are referred to as the tangent and cotangent spaces.The infinitesimal displacements we've been discussing belong to the contravariant (upper-index) space, but by lowering and index we can just as well discuss them as covariant vectors. The customary way of notating Killing vectors makes use of the fact, mentioned in passing in Section 5.10, that the partial derivative operators $\partial_0, \partial_1, \partial_2, \partial_3$ form the basis for a vector space. In this notation, the Killing vector of the Schwarzschild metric we've been discussing can be notated simply as

$$\boldsymbol{\xi} = \partial_t. \tag{7.1.5}$$

The partial derivative notation, like the infinitesimal notation, implicitly refers to continuous symmetries rather than discrete ones. If a discrete symmetry carries a point P_1 to some distant point P_2, then P_1 and P_2 have two different tangent planes, so there is not a uniquely defined notion of whether vectors $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_1$ at these two points are equal — or even approximately equal. There can therefore be no well-defined way to construe a statement such as, "P_1 and P_2 are separated by a displacement $\boldsymbol{\xi}$." In the case of a continuous symmetry, on the other hand, the two tangent planes come closer and closer to coinciding as the distance s between two points on an orbit approaches zero, and in this limit we recover an approximate notion of being able to compare vectors in the two tangent planes. They can be compared by parallel transport, and although parallel transport is path-dependent, the difference bewteen paths is proportional to the area they enclose, which varies as s^2, and therefore becomes negligible in the limit s → 0.

Exercise 7.1.1

Find another Killing vector of the Schwarzschild metric, and express it in the tangent-vector notation.

It can be shown that an equivalent condition for a field to be a Killing vector is

$$\nabla_a \boldsymbol{\xi}_b + \nabla_b \boldsymbol{\xi}_a = 0. \tag{7.1.6}$$

This relation, called the Killing equation, is written without reference to any coordinate system, in keeping with the coordinate-independence of the notion.

When a spacetime has more than one Killing vector, any linear combination of them is also a Killing vector. This means that although the existence of certain types of Killing vectors may be intrinsic, the exact choice of those vectors is not.

Example 7.1.2: Euclidean translations

The Euclidean plane has two translational Killing vectors (1, 0) and (0, 1), i.e., ∂_x and ∂_y. These same vectors could be expressed as (1, 1) and (1, −1) in coordinate system that was rescaled and rotated by 45 degrees.

Example 7.1.3: a cylinder

The local properties of a cylinder, such as intrinsic flatness, are the same as the local properties of a Euclidean plane. Since the definition of a Killing vector is local and intrinsic, a cylinder has the same three Killing vectors as a plane, if we consider only a patch on the cylinder that is small enough so that it doesn't wrap all the way around. However, only two of these — the

translations — can be extended to form a smooth vector field on the entire surface of the cylinder. These might be more naturally notated in (ϕ, z) coordinates rather than (x, y), giving ∂_z and ∂_ϕ.
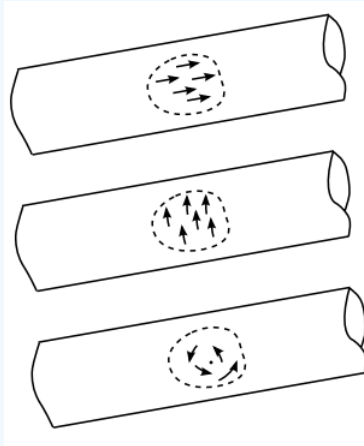

Figure 7.1.4: A cylinder has three local symmetries, but only two that can be extended globally to make Killing vectors.

Example 7.1.4: a sphere

A sphere is like a plane or a cylinder in that it is a two-dimensional space in which no point has any properties that are intrinsically different than any other. We might expect, then, that it would have two Killing vectors. Actually it has three, ξ_x, ξ_y, and ξ_z, corresponding to infinitesimal rotations about the x, y, and z axes. To show that these are all independent Killing vectors, we need to demonstrate that we can't, for example, have $\xi_x = c_1 \xi_y + c_2 \xi_z$ for some constants c_1 and c_2. To see this, consider the actions of ξ_y and ξ_z on the point P where the x axis intersects the sphere. (References to the axes and their intersection with the sphere are extrinsic, but this is only for convenience of description and visualization.) Both ξ_y and xi_z move P around a little, and these motions are in orthogonal directions, whereas ξ_x leaves P fixed. This proves that we can't have $\xi_x = c_1 \xi_y + c_2 \xi_z$. All three Killing vectors are linearly independent.

This example shows that linear independence of Killing vectors can't be visualized simply by thinking about the vectors in the tangent plane at one point. If that were the case, then we could have at most two linearly independent Killing vectors in this two-dimensional space. When we say "Killing vector" we're really referring to the Killing vector *field*, which is defined everywhere on the space.

Example 7.1.5: Proving nonexistence of Killing vectors

- Find all Killing vectors of these two metrics:

$$ds^2 = e^{-x} dx^2 + e^x dy^2$$
$$ds^2 = dx^2 + x^2 dy^2.$$

- Since both metrics are manifestly independent of y, it follows that ∂_y is a Killing vector for both of them. Neither one has any other manifest symmetry, so we can reasonably conjecture that this is the only Killing vector either one of them has. However, one can have symmetries that are not manifest, so it is also possible that there are more.

One way to attack this would be to use the Killing equation to find a system of differential equations, and then determine how many linearly independent solutions there were.

But there is a simpler approach. The dependence of these metrics on x suggests that the spaces may have intrinsic properties that depend on x; if so, then this demonstrates a lower symmetry than that of the Euclidean plane, which has three Killing vectors. One intrinsic property we can check is the scalar curvature R. The following Maxima code calculates R for the first metric.

```
1    load(ctensor);
2    dim:2;
3    ct_coords:[x,y];
4    lg:matrix([exp(-x),0],[0,exp(x)]);
5    cmetric();
6    R:scurvature(); /* scalar curvature */
```

The result is R = −e^x, which demonstrates that points that differ in x have different intrinsic properties. Since the flow of a Killing field ξ can never connect points that have different properties, we conclude that $\xi_x = 0$. If only ξ_y can be nonzero, the Killing equation $\nabla_a \xi_b + \nabla_b \xi_a = 0$ simplifies to $\nabla_x \xi_y = \nabla_y \xi_y = 0$. These equations constrain both $\partial_x \xi_y$ and $\partial_y \xi_y$, which means that given a value of ξ_y at some point in the plane, its value everywhere else is determined. Therefore the only possible Killing vectors are scalar multiples of the Killing vector already found. Since we don't consider Killing vectors to be distinct unless they are linearly independent, the first metric only has one Killing vector.

A similar calculation for the second metric shows that R = 0, and an explicit calculation of its Riemann tensor shows that in fact the space is flat. It is simply the Euclidean plane written in funny coordinates. This metric has the same three Killing vectors as the Euclidean plane.

It would have been tempting to leap to the wrong conclusion about the second metric by the following reasoning. The signature of a metric is an intrinsic property. The metric has signature ++ everywhere in the plane except on the y axis, where it has signature +0. This shows that the y axis has different intrinsic properties than the rest of the plane, and therefore the metric must have a lower symmetry than the Euclidean plane. It can have at most two Killing vectors, not three. This contradicts our earlier conclusion. The resolution of this paradox is that this metric has a removable degeneracy of the same type as the one described in section 6.4. As discussed in that section, the signature is invariant only under *nonsingular* transformations, but the transformation that converts these coordinates to Cartesian ones is singular.

Inappropriate Mixing of Notational Systems

Confusingly, it is customary to express vectors and dual vectors by summing over basis vectors like this:

$$\mathbf{v} = v^\mu \partial_\mu$$
$$\boldsymbol{\omega} = \omega_\mu dx^\mu.$$

This is an abuse of notation, driven by the desire to have up-down pairs of indices to sum according to the usual rules of the Einstein notation convention. But by that convention, a quantity like $\mathbf{v}$ or $\boldsymbol{\omega}$ with no indices is a scalar, and that's not the case here. The products on the right are not tensor products, i.e., the indices aren't being contracted.

This muddle is the result of trying to make the Einstein notation do too many things at once and of trying to preserve a clumsy and outdated system of notation and terminology originated by Sylvester in 1853. In pure abstract index notation, there are not six flavors of objects as in the two equations above but only two: vectors like v^a and dual vectors like ω_a. The Sylvester notation is the prevalent one among mathematicians today, because their predecessors committed themselves to it a century before the development of alternatives like abstract index notation and birdtracks. The Sylvester system is inconsistent with the way physicists today think of vectors and dual vectors as being defined by their transformation properties, because Sylvester considers $\mathbf{v}$ and $\boldsymbol{\omega}$ to be invariant.

Mixing the two systems leads to the kinds of notational clashes described above. As a particularly absurd example, a physicist who is asked to suggest a notation for a vector will typically pick up a pen and write v$^\mu$. We are then led to say that a vector is written in a concrete basis as a linear combination of dual vectors ∂_μ!

Conservation Laws

Whenever a spacetime has a Killing vector, geodesics have a constant value of v$^b \xi_b$, where v^b is the velocity four-vector. For example, because the Schwarzschild metric has a Killing vector $\boldsymbol{\xi} = \partial_t$, test particles have a conserved value of v^t, and therefore we also have conservation of p^t, interpreted as the mass-energy.

Example 7.1.6: Energy-momentum in flat 1+1 spacetime

A flat 1+1-dimensional spacetime has Killing vectors ∂_x and ∂_t. Corresponding to these are the conserved momentum and mass-energy, p and E. If we do a Lorentz boost, these two Killing vectors get mixed together by a linear transformation, corresponding to a transformation of p and E into a new frame.

In addition, one can define a globally conserved quantity found by integrating the flux density $P^a = T^{ab}\xi_b$ over the boundary of any compact orientable region.[2] In case of a flat spacetime, there are enough Killing vectors to give conservation of energy-momentum and angular momentum.

Note

Hawking and Ellis, *The Large Scale Structure of Space-Time*, p. 62, give a succinct treatment that describes the flux densities and proves that Gauss's theorem, which ordinarily fails in curved spacetime for a non-scalar flux, holds in the case where the appropriate Killing vectors exist. For an explicit description of how one can integrate to find a scalar mass-energy, see Winitzki, *Topics in General Relativity*, section 3.1.5, available for free online.

This page titled 7.1: Killing Vectors is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Benjamin Crowell via source content that was edited to the style and standards of the LibreTexts platform.

7.2: Spherical Symmetry

A little more work is required if we want to link the existence of Killing vectors to the existence of a specific symmetry such as spherical symmetry. When we talk about spherical symmetry in the context of Newtonian gravity or Maxwell's equations, we may say, "The fields only depend on r," implicitly assuming that there is an r coordinate that has a definite meaning for a given choice of origin. But coordinates in relativity are not guaranteed to have any particular physical interpretation such as distance from a particular origin. The origin may not even exist as part of the spacetime, as in the Schwarzschild metric, which has a singularity at the center. Another possibility is that the origin may not be unique, as on a Euclidean two-sphere like the earth's surface, where a circle centered on the north pole is also a circle centered on the south pole; this can also occur in certain cosmological spacetimes that describe a universe that wraps around on itself spatially.

We therefore define spherical symmetry as follows. A spacetime S is spherically symmetric if we can write it as a union $S = \cup s_{r,t}$ of nonintersecting subsets $s_{r,t}$, where each s has the structure of a twosphere, and the real numbers r and t have no preassigned physical interpretation, but $s_{r,t}$ is required to vary smoothly as a function of them. By "has the structure of a two-sphere," we mean that no intrinsic measurement on s will produce any result different from the result we would have obtained on some two-sphere. A two-sphere has only two intrinsic properties:

1. it is spacelike, i.e., locally its geometry is approximately that of the Euclidean plane;
2. it has a constant positive curvature.

If we like, we can require that the parameter r be the corresponding radius of curvature, in which case t is some timelike coordinate.

To link this definition to Killing vectors, we note that condition 2 is equivalent to the following alternative condition: (2') The set s should have three Killing vectors (which by condition 1 are both spacelike), and it should be possible to choose these Killing vectors such that algebraically they act the same as the ones constructed explicitly in example 4 in section 7.1. As an example of such an algebraic property, Figure $7.2.1$ shows that rotations are noncommutative.
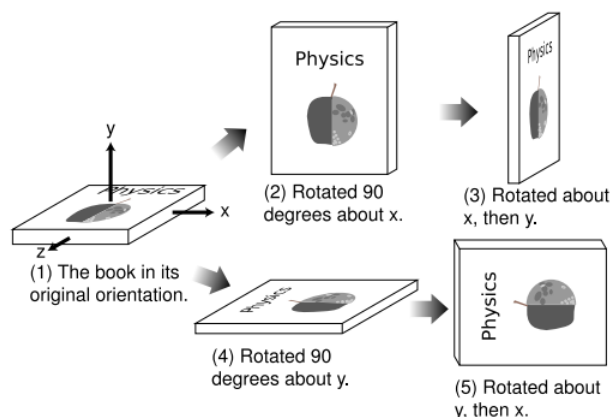
Figure $7.2.1$: Performing the rotations in one order gives one result, 3, while reversing the order gives a different result, 5.

Example 7: A cylinder is not a sphere

- Show that a cylinder does not have the structure of a two-sphere.
- The cylinder passes condition 1. It fails condition 2 because its Gaussian curvature is zero. Alternatively, it fails condition 2' because it has only two independent Killing vectors (example 3).

Example 8: A plane is not a sphere

- Show that the Euclidean plane does not have the structure of a two-sphere.
- Condition 2 is violated because the Gaussian curvature is zero. Or if we wish, the plane violates 2' because ∂_x and ∂_y commute, but none of the Killing vectors of a 2-sphere commute.

7.3: Penrose Diagrams and Causality

We can't directly visualize a four-dimensional manifold. When a spacetime has a symmetry, however, we may be able to visualize the relevant properties the whole thing by considering a lower-dimensional part of it. By analogy, if we wanted to visualize the structure of the earth's interior, we might draw a diagram showing a two-dimensional section through its center. In fact, we could get rid of two dimensions and simply draw a diagram of a single radial line running from the earth's core to its surface; each point on this line would then represent a sphere. If we do this in general relativity, for a spacetime that is spherically symmetric, then we can reduce the four-dimensional to a two-dimensional one, with each point representing a two-sphere. By applying some further tricks, we will see that we can end up with a very convenient and useful visualization called a Penrose diagram, also known as a Penrose-Carter diagram or causal diagram.

Flat Spacetime

As a warmup, Figure 7.3.1 shows a Penrose diagram for flat (Minkowski) spacetime. The diagram looks $1 + 1$-dimensional, but the convention is that spherical symmetry is assumed, so two more dimensions are hidden, and we're really portraying $3 + 1$ dimensions. A typical point on the interior of the diamond region represents a 2-sphere. On this type of diagram, light cones look just like they would on a normal spacetime diagram of Minkowski space, but distance scales are highly distorted. The diamond represents the entire spacetime, with the distortion fitting this entire infinite region into that finite area on the page. Despite the distortion, the diagram shows lightlike surfaces as 45-degree diagonals. Spacelike and timelike geodesics, however, are distorted, as shown by the curves in the diagram.
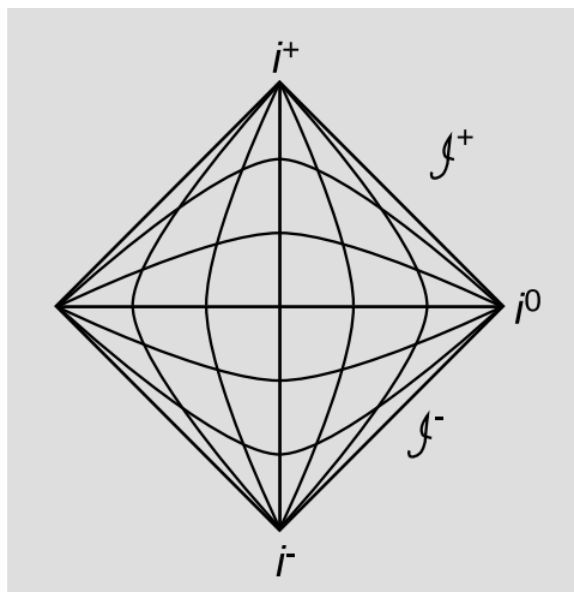
Figure 7.3.1 - Penrose diagram for flat spacetime.

The distortion becomes greater as we move away from the center of the diagram, and becomes infinite near the edges. Because of this infinite distortion, the points i^- and i^+ actually represent 3-spheres. All timelike curves start at i^- and end at i^+, which are idealized points at infinity, like the vanishing points in perspective drawings. We can think of i^+ as the "Elephants' graveyard," where massive particles go when they die. Similarly, lightlike curves end on $\mathscr{I}^+$ (which includes its mirror image on the left), referred to as *null infinity*.[3] The point at i^0 is an infinitely distant endpoint for spacelike curves. Because of the spherical symmetry, the left and right halves of the diagram are redundant.

It is possible to make up explicit formulae that translate back and forth between Minkowski coordinates and points on the diamond, but in general this is not necessary. In fact, the utility of the diagrams is that they let us think about causal relationships in coordinate-independent ways. A light cone on the diagram looks exactly like a normal light cone.

Since this particular spacetime is homogeneous, it makes no difference what spatial location on the diagram we pick as our axis of symmetry. For example, we could arbitrarily pick the left-hand corner, the central timelike geodesic (drawn straight) or one of the other timelike geodesics (represented as if it were curved).

Schwarzschild Spacetime

Figure 7.3.2 is a Penrose diagram for the Schwarzschild spacetime, i.e., a spacetime that looks like Minkowski space, except that it has one eternal black hole in it. This is a black hole that did not form by gravitational collapse. This spacetime isn't homogeneous; it has a specific location that is its center of spherical symmetry, and this is the vertical line on the left marked r = 0. The triangle is the spacetime inside the event horizon; we could have copied it across the r = 0 line if we had so desired, but the copies would have been redundant.
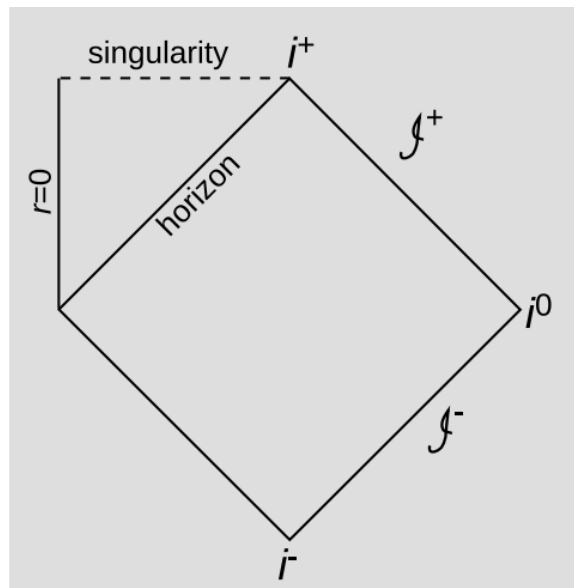
Figure 7.3.2 - *Penrose diagram for Schwarzschild spacetime: a black hole that didn't form by gravitational collapse.*

The Penrose diagram makes it easy to reason about causal relationships. For example, we can see that if a particle reaches a point inside the event horizon, its entire causal future lies inside the horizon, and all of its possible future world-lines intersect the singularity. The horizon is a lightlike surface, which makes sense, because it's defined as the boundary of the set of points from which a light ray could reach $\mathscr{I}^{+}$.

Astrophysical Black Hole

Figure 7.3.3 is a Penrose diagram for a black hole that has formed by gravitational collapse. Using this type of diagram, we can succinctly address one of the most vexing FAQs about black holes. (Cf. section 6.3, where we took a more cumbersome approach without Penrose diagrams.) If a distant observer watches the collapsing cloud of matter from which the black hole forms, her optical observations will show that the light from the matter becomes more and more gravitationally redshifted, and if she wishes, she can interpret this as an example of gravitational time dilation. As she waits longer and longer, the light signals from the infalling matter take longer and longer to arrive. The redshift approaches infinity as the matter approaches the horizon, so the light waves ultimately become too low in energy to be detectable by any given instrument. Furthermore, her patience (or her lifetime) will run out, because the time on her clock approaches infinity as she waits to get signals from matter that is approaching the horizon. This is all exactly as it should be, since the horizon is by definition the boundary of her observable universe. (A light ray emitted from the horizon will end up at i$^+$, which is an end-point of timelike world-lines reached only by observers who have experienced an infinite amount of proper time.)
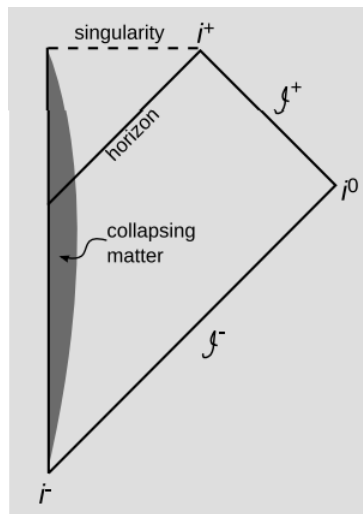
Figure 7.3.3 - Penrose diagram for a black hole formed by gravitational collapse.

People who are bothered by these issues often acknowledge the external *unobservability* of matter passing through the horizon, and then want to pass from this to questions like, "Does that mean the black hole never really forms?" This presupposes that our distant observer has a uniquely defined notion of simultaneity that applies to a region of space stretching from her own position to the interior of the black hole, so that she can say what's going on inside the black hole "now." But the notion of simultaneity in general relativity is even more limited than its counterpart in special relativity. Not only is simultaneity in general relativity observer-dependent, as in special relativity, but it is also local rather than global.

In Figure 7.3.4, E is an event on the world-line of an observer. The spacelike surface S_1 is one possible "now" for this observer. According to this surface, no particle has ever fallen in and reached the horizon; every such particle has a world-line that intersects S_1, and therefore it's still on its way in.
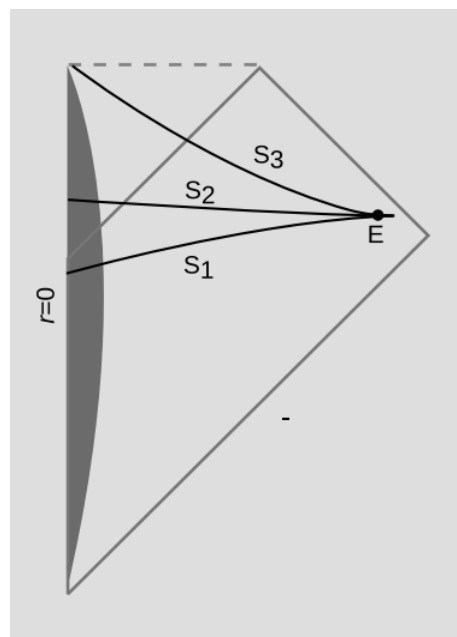
Figure 7.3.4 - According to the distant observer, does the infalling matter ever reach the singularity?

S_2 is another possible "now" for the same observer at the same time. According to this definition of "now," all the particles have passed the event horizon, but none have hit the singularity yet. Finally, S_3 is a "now" according to which all the particles have hit the singularity.

If this was special relativity, then we could decide which surface was the correct notion of simultaneity for the observer, based on the observer's state of motion. But in general relativity, this only works locally (which is why I made all three surfaces coincide near E). There is no well-defined way of deciding which is the correct way of globally extending this notion of simultaneity.

Although it may seem strange that we can't say whether the singularity has "already" formed according to a distant observer, this is really just an inevitable result of the fact that the singularity is spacelike. The same thing happens in the case of a Schwarzschild spacetime, which we think of as a description of an eternal black hole, i.e., one that has always existed and always will. On the similar Penrose diagram for an eternal black hole, we can still draw a spacelike surface like S_1 or S_2, representing a definition of "now" such that the singularity doesn't exist yet.

Penrose Diagrams in General

Ideally we would like to generalize the procedure for drawing Penrose diagrams so that we would be able to uniquely determine one for any spacetime. This turns out to be not so clear-cut. The procedure would go something like this:

1. Make an n-dimensional section or projection, where usually, but not always, n = 2.
2. Do a transformation to reduce the resulting manifold to a flat one of finite size.
3. Adjoin idealized surfaces and points at infinity.

At step 1, we want to take advantage of any symmetries, such as rotational symmetry, so that the final result will be informative, be representative of the whole spacetime, and accurately depict causal relationships in the original spacetime. If the original spacetime has a low degree of symmetry (e.g., a spacetime containing three black holes arranged in a triangle), then this might require n > 2. At this step we also need to make sure that lightlike geodesics in the original space correspond properly to lightlike geodesics in the submanifold.

For step 2, we have already given a geometrical characterization of the type of transformation we have in mind, which is called a *conformal transformation*. It turns out to be possible to encapsulate this idea in a simple analytic way. Given a spacetime with a metric g, we define a fictitious metric $\tilde{g} = \Omega^2 g$, where Ω is a nonzero real number that varies from point to point. (Cf. sec. 5.11, where Ω was constant.) The idea here is that g and $\tilde{g}$ agree on where the light cone is, but they disagree on the measurement of distances and times. The same manifold equipped with the fictitious metric $\tilde{g}$ is the one being drawn on the page when we make a Penrose diagram. We let $\Omega \to 0$ as we approach the idealized boundary regions like i^0 and $\mathscr{I}^+$, and this is what causes the Penrose diagram to take up finite space on the page.

It is not possible in general to do what is required in step 2 by making a conformal transformation to change a manifold into a flat one. A manifold that can be flattened in this way is called conformally flat. All two-dimensional manifolds are conformally flat, so in the n = 2 case this is guaranteed. For n > 3 we will usually not have conformal flatness if there are gravitational waves or tidal forces present.

The most problematic part, surprisingly, is step 3. This topic goes under the general heading of "boundary constructions." Reviews are available on this topic.[4] There are a number of more or less specific techniques for constructing a boundary, with an alphabet soup of names including the g-boundary, c-boundary, b-boundary, and a-boundary. As someone who is not a specialist in this subfield, the impression I get is that this is an area of research that has turned out badly and has never produced any useful results, but work continues, and it is possible that at some point the smoke will clear. As a simple example of what one would like to get, but doesn't get, from these studies, it would seem natural to ask how many dimensions there are in a black hole singularity. (See example 4 for a discussion of why this is a nontrivial question.) Different answers come back from the different methods. For example, the b-boundary approach says that both black-hole and cosmological singularities are zero-dimensional points, while in the c-boundary method (which was designed to harmonize with Penrose diagrams) they are three-surfaces (as one would imagine from the Penrose diagrams).

Global Hyperbolicity

Causality refers to our vaguely defined feeling that the world should have an orderly progression of cause and effect. Making this notion more precise is surprisingly difficult. Penrose diagrams, and their associated concepts, are essentially representations of the causal structure of spacetime, and these turn out to be helpful in putting together one of the more satisfying attempts to define causality. This definition is called global hyperbolicity. The obscure terminology is related to the classification of partial differential equations.

Some definitions are required as a preliminary. Consider a set S of events in spacetime. S is *bounded* if it does not include any of the idealized points on the Penrose diagram that we have added at infinity.[5] S is *closed* if it contains its own boundary.[6] S is *compact* if it is closed and bounded.

Example 9: Compact and noncompact light cones

Figure 7.3.5 shows a spacetime containing a black hole that forms by gravitational collapse. Point P is inside the event horizon, Q outside. Consider the following four point-sets:
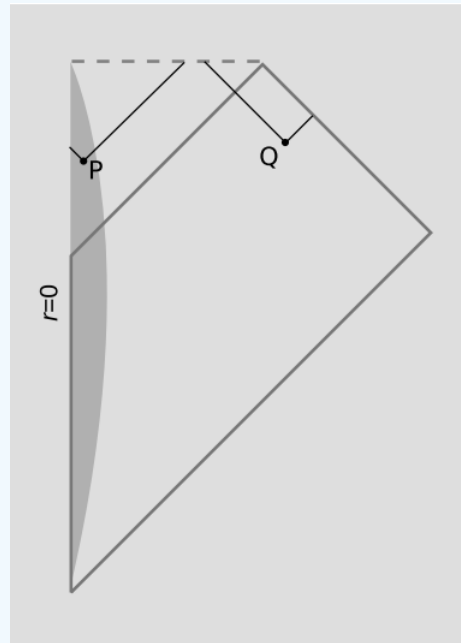

Figure 7.3.5

I^+ (P), called the *chronological future* of P, is the interior of P's future-directed light cone.

J^+ (P) is like I^+ (P), but also includes events that are on the boundary of the light cone, i.e., events that cannot be connected to P by a timelike curve but that can be connected to it by a lightlike curve. We call this the *causal future* of P, since it is the set of events that could be caused by P.

I^+ (Q) and J^+ (Q) are the analogous sets built on Q.

Of these four sets, only J^+ (P) is compact. I^+ (P) is noncompact because it is not closed. I^+ (Q) and J^+ (Q) are noncompact because they are not bounded; they include idealized points at infinity that lie in i^+ and $\mathscr{I}^+$.

In addition to the notation introduced in example 9, we will need the similar notations I^- and J^- for the corresponding past light-cones.

Definition

A spacetime is globally hyperbolic if: (1) there are no closed, timelike curves (CTCs),[7] and (2) given any two events P and Q, the intersection of J^+(P) and J^-(Q) is compact. (Condition 2 is required only when P and Q are points in the manifold, not boundary points.)

[7] For precision, the condition needs to be made a little stronger. We want no closed, non-spacelike curves, and we also want it to be impossible for curves to exist that are arbitrarily close to being such curves, in the sense that for any event, there exists a neighborhood around it that can never be revisited.

In a globally hyperbolic spacetime, initial-value problems always have unique solutions. That is, we can pick a spacelike surface and give the value of a wave on that surface, and the wave equation will then have a unique solution. Such a surface is called a Cauchy surface.

We can readily verify by inspection of the Penrose digrams that the spacetimes described earlier in this section are globally hyperbolic. Condition 2 implies that the intersection doesn't contain any singularities or points at infinity. Although black hole spacetimes do contain singularities, the spacelike nature of these singularities implies that they can never lie in the intersection of light cones as referred to in the definition. Therefore such spacetimes are globally hyperbolic.

Example 10: Global hyperbolicity
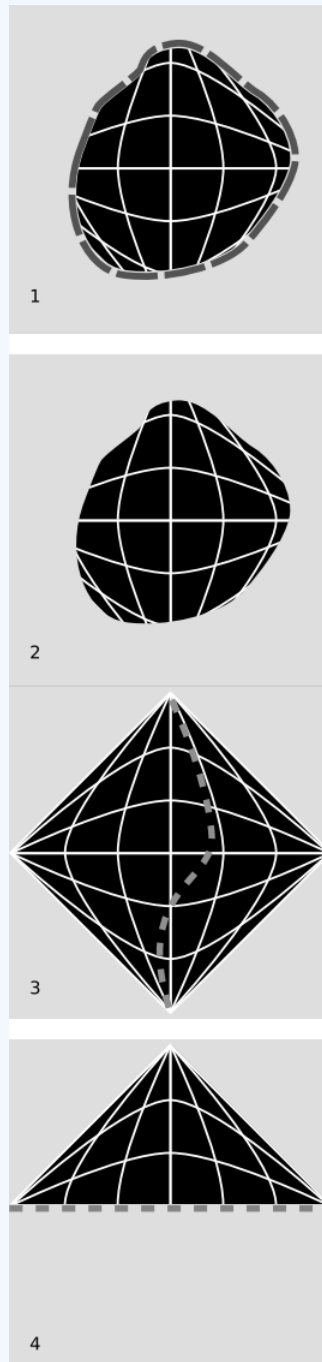
Figure 7.3.6

Figure 7.3.6 (1) shows a piece cut out of Minkowski space. The dashed outline is meant to indicate that the piece doesn't include its boundary. This spacetime is not globally hyperbolic. For certain choices of events P and Q, the intersection J^+ (P) $\cap$ J^-(Q) could extend out to the cut at the edge. Since the spacetime doesn't include its boundary, this intersection would not be compact. It's easy to see why causality fails in this spacetime. If we pick a spacelike surface near the bottom of the diagram, it would only cut through a small part of the bottom of the spacetime. At later times, the spacetime grows at a rate that is greater than c. Therefore such a surface cannot be used as a Cauchy surface; given the initial conditions on this surface, we cannot predict what will happen in the parts of the universe that are outside its causal future.

In Figure 7.3.6 (2) we have the same example, but now the boundary is included. This set is not a manifold, which excludes it from consideration as a spacetime in general relativity. Figure 7.3.6 (3) is a picture of Minkowski space with a timelike singularity in it (dashed line). Singularities are not point-sets in the manifold, so topologically, this is like Minkowski space with a single timelike curve surgically removed. Global hyperbolicity fails because the intersection J^+ (P) $\cap$ J^-(Q) could surround the singularity, and would not be compact because it would not include its boundary at the singularity. This violation of global hyperbolicity indicates a failure of causality in such a spacetime (see section 6.3).

By cutting off the lower half of the diamond representing Minkowski spacetime, we obtain Figure 7.3.6 (4). The dashed line indicates that the boundary is not included, and therefore this is a manifold. It is also globally hyperbolic. This example suggests that global hyperbolicity does not necessarily capture everything we might ever want to describe in a definition of causality. If a paleontologist living in this spacetime finds a dinosaur fossil embedded in a rock, she will naturally infer that a dinosaur lived at some point in the past, causing the fossil to exist. But perhaps this is not the case — the hypothetical dinosaur might be one that would have existed before the boundary. This creationism-flavored violation of causality is of a different flavor than the situation we would have had if the bottom edge of the diagram had been a big bang singularity; in that case, we would have had a knowable reason why chains of cause and effect could not be extended back into the past beyond a certain time.

References

[3] See also section 7.4

[4] Ashley, "Singularity theorems and the abstract boundary construction," digitalcollections.anu.edu.a...dle/1885/46055. GarciaParrado and Senovilla, "Causal structures and causal boundaries," http:// arxiv.org/abs/gr-qc/0501069.

7.4: Static and Stationary Spacetimes (Part 1)

Stationary Spacetimes

When we set out to describe a generic spacetime, the Alice in Wonderland quality of the experience is partly because coordinate invariance allows our time and distance scales to be arbitrarily rescaled, but also partly because the landscape can change from one moment to the next. The situation is drastically simplified when the spacetime has a timelike Killing vector. Such a spacetime is said to be stationary. Two examples are flat spacetime and the spacetime surrounding the rotating earth (in which there is a frame-dragging effect). Non-examples include the solar system, cosmological models, gravitational waves, and a cloud of matter undergoing gravitational collapse.

Can Alice determine, by traveling around her spacetime and carrying out observations, whether it is stationary? If it's not, then she might be able to prove it. For example, suppose she visits a certain region and finds that the *Kretchmann invariant* $R^{abcd}R_{abcd}$ varies with time in her frame of reference. Maybe this is because an asteroid is coming her way, in which case she could readjust her velocity vector to match that of the asteroid. Even if she can't see the asteroid, she can still try to find a velocity that makes her local geometry stop changing in this particular way. If the spacetime is truly stationary, then she can always "tune in" to the right velocity vector in this way by searching systematically. If this procedure ever fails, then she has proved that her spacetime is not stationary.

> **Exercise 7.4.1**
>
> Why is the timelike nature of the Killing vector important in this story?

Proving that a spacetime is stationary is harder. This is partly just because spacetime is infinite, so it will take an infinite amount of time to check everywhere. We aren't inclined to worry too much about this limitation on our geometrical knowledge, which is of a type that has been familiar since thousands of years ago, when it upset the ancient Greeks that the parallel postulate could only be checked by following lines out to an infinite distance. But there is a new type of limitation as well. The Schwarzschild spacetime is not stationary according to our definition. In the coordinates used in section 6.2, ∂_t is a Killing vector, but is only timelike for r > 2m; for r < 2m it is spacelike. Although the solution describes a black hole that is going to sit around forever without changing, no observer can ever verify that fact, because once she strays inside the horizon she must follow a timelike world-line, which will end her program of observation within some finite time.

Isolated Systems

Asymptotic flatness

This unfortunate feature of our definition of stationarity — its empirical unverifiability — is something that in general we just have to live with. But there is an alternative in the special case of an isolated system, such as our galaxy or a black hole. It may be a good approximation to ignore distant matter, modeling such a system with a spacetime that is asymptotically flat. The notion of asymptotic flatness was introduced informally in section 4.5. Formulating the definition of this term rigorously and in a coordinate-invariant way involves a large amount of technical machinery, since we are not guaranteed to be presented in advance with a special, physically significant set of coordinates that would lead directly to a quantitative way of defining words like "nearby." The essential idea is that a spacetime is asymptotically flat if it is possible to perform a conformal transformation in such a way that the result has idealized regions at infinity i^0, $\mathscr{I}^+$ and $\mathscr{I}^-$ (but not i^+ and i^-) that look like those of Minkowski space. The reader who wants to see the full machinery presented can find presentations in various places, such as Hawking and Ellis, ch. 11 of Wald, or the online review article "Conformal Infinity" at livingreviews.org.

Asymptotically stationary spacetimes

In the case of an asymptotically flat spacetime, we say that it is also asymoptotically stationary if it has a Killing vector that becomes timelike far away. Some authors (e.g., Ludvigsen) define "stationary" to mean what I'm calling "asymoptotically stationary," others (Hawking and Ellis) define it the same way I do, and still others (Carroll) are not self-consistent. The Schwarzschild spacetime is asymptotically stationary, but not stationary.

A Stationary Field with No Other Symmetries

Consider the most general stationary case, in which the only Killing vector is the timelike one. The only ambiguity in the choice of this vector is a rescaling; its direction is fixed. At any given point in space, we therefore have a notion of being at rest, which is to have a velocity vector parallel to the Killing vector. An observer at rest detects no time-dependence in quantities such as tidal forces.

Points in space thus have a permanent identity. The gravitational field, which the equivalence principle tells us is normally an elusive, frame-dependent concept, now becomes more concrete: it is the proper acceleration required in order to stay in one place. We can therefore use phrases like "a stationary field," without the usual caveats about the coordinate-dependent meaning of "field."

Space can be sprinkled with identical clocks, all at rest. Furthermore, we can compare the rates of these clocks, and even compensate for mismatched rates, by the following procedure. Since the spacetime is stationary, experiments are reproducible. If we send a photon or a material particle from a point A in space to a point B, then identical particles emitted at later times will follow identical trajectories. The time lag between the arrival of two such particls tells an observer at B the amount of time at B that corresponds to a certain interval at A. If we wish, we can adjust all the clocks so that their rates are matched. An example of such rate-matching is the GPS satellite system, in which the satellites' clocks are tuned to 10.22999999543 MHz, matching the ground-based clocks at 10.23 MHz. (Strictly speaking, this example is out of place in this subsection, since the earth's field has an additional azimuthal symmetry.)

It is tempting to conclude that this type of spacetime comes equipped with a naturally preferred time coordinate that is unique up to a global affine transformation t → at + b. But to construct such a time coordinate, we would have to match not just the rates of the clocks, but also their phases. The best method relativity allows for doing this is Einstein synchronization (Appendix 1), which involves trading a photon back and forth between clocks A and B and adjusting the clocks so that they agree that each clock gets the photon at the mid-point in time between its arrivals at the other clock. The trouble is that for a general stationary spacetime, this procedure is not transitive: synchronization of A with B, and of B with C, does not guarantee agreement between A with C. This is because the time it takes a photon to travel clockwise around triangle ABCA may be different from the time it takes for the counterclockwise itinerary ACBA. In other words, we may have a Sagnac effect, which is generally interpreted as a sign of rotation. Such an effect will occur, for example, in the field of the rotating earth, and it cannot be eliminated by choosing a frame that rotates along with the earth, because the surrounding space experiences a frame-dragging effect, which falls off gradually with distance.

Although a stationary spacetime does not have a uniquely preferred time, it does prefer some time coordinates over others. In a stationary spacetime, it is always possible to find a "nice" t such that the metric can be expressed without any t-dependence in its components.

A Stationary Field with Additional Symmetries

Most of the results given above for a stationary field with no other symmetries also hold in the special case where additional symmetries are present. The main difference is that we can make linear combinations of a particular timelike Killing vector with the other Killing vectors, so the timelike Killing vector is not unique. This means that there is no preferred notion of being at rest. For example, in a flat spacetime we cannot define an observer to be at rest if she observes no change in the local observables over time, because that is true for any inertial observer. Since there is no preferred rest frame, we can't define the gravitational field in terms of that frame, and there is no longer any preferred definition of the gravitational field.

Static Spacetimes

In addition to synchronizing all clocks to the same frequency, we might also like to be able to match all their phases using Einstein synchronization, which requires transitivity. Transitivity is framedependent. For example, flat spacetime allows transitivity if we use the usual coordinates. However, if we change into a rotating frame of reference, transitivity fails (see section 3.5). If coordinates exist in which a particular spacetime has transitivity, then that spacetime is said to be static. In these coordinates, the metric is diagonalized, and since there are no space-time cross-terms like dx dt in the metric, such a spacetime is invariant under time reversal. Roughly speaking, a static spacetime is one in which there is no rotation.

Birkhoff's Theorem

Birkhoff's theorem, proved below, states that in the case of spherical symmetry, the vacuum field equations have a solution, the Schwarzschild spacetime, which is unique up to a choice of coordinates and the value of m. Let's enumerate the assumptions that

went into our derivation of the Schwarzschild metric in section 6.2. These were: (1) the vacuum field equations, (2) spherical symmetry, (3) asymptotic staticity, (4) a certain choice of coordinates, and (5) $\Lambda = 0$. Birkhoff's theorem says that the assumption of staticity was not necessary. That is, even if the mass distribution contracts and expands over time, the exterior solution is still the Schwarzschild solution. Birkhoff's theorem holds because gravitational waves are transverse, not longitudinal (see section 9.2), so the mass distribution's radial throbbing cannot generate a gravitational wave.

Proof of Birkhoff's theorem

Spherical symmetry guarantees that we can introduce coordinates r and t such that the surfaces of constant r and t have the structure of a sphere with radius r. On one such surface we can introduce colatitude and longitude coordinates θ and ϕ. The (θ, ϕ) coordinates can be extended in a natural way to other values of r by choosing the radial lines to lie in the direction of the covariant derivative vector[8] $\nabla_a r$, and this ensures that the metric will not have any nonvanishing terms in dr $d\theta$ or dr $d\phi$, which could only arise if our choice had broken the symmetry between positive and negative values of $d\theta$ and $d\phi$. Just as we were free to choose any way of threading lines of constant (θ, ϕ, t) between spheres of different radii, we can also choose how to thread lines of constant (θ, ϕ, r) between different times, and this can be done so as to keep the metric free of any time-space cross-terms such as $d\theta$ dt. The metric can therefore be written in the form[9]

$$ds^2 = h(t,r)dt^2 - k(t,r)dr^2 - r^2(d\theta^2 + \sin^2\theta d\phi^2). \tag{7.4.1}$$

This has to be a solution of the vacuum field equations, $R_{ab} = 0$, and in particular a quick calculation with Maxima shows that $R_{rt} = -\frac{\partial_t k}{k^2 r}$, so k must be independent of time. With this restriction, we find

\[R_{rr} = − \frac{\partial_{r} h}{hkr} − \frac{1}{r^{2}} − \frac{1}{kr^{2}} = 0,\]

and since k is time-independent, $\frac{\partial_r h}{h}$ is also time-independent. This means that for a particular time to, the function $f(r) = h(t_o, r)$ has some universal shape set by a differential equation, with the only possible ambiguity being an over-all scaling that depends on to. But since h is the timetime component of the metric, this scaling corresponds physically to a situation in which every clock, all over the universe, speeds up and slows down in unison. General relativity is coordinate-independent, so this has no observable effects, and we can absorb it into a redefinition of t that will cause h to be time-independent. Thus the metric can be expressed in the time-independent diagonal form

$$ds^2 = h(r)dt^2 - k(r)dr^2 - r^2(d\theta^2 + \sin^2\theta d\phi^2). \tag{7.4.2}$$

We have already solved the field equations for a metric of this form and found as a solution the Schwarzschild spacetime.[10] Since the metric's components are all independent of t, ∂_t is a Killing vector, and it is timelike for large r, so the Schwarzschild spacetime is asymptotically static.

---

[8] It may seem backwards to start talking about the covariant derivative of a particular coordinate before a complete coordinate system has even been introduced. But (excluding the trivial case of a flat spacetime), r is not just an arbitrary coordinate, it is something that an observer at a certain point in spacetime can determine by mapping out a surface of geometrically identical points, and then determining that surface's radius of curvature. Another worry is that it is possible for $\nabla_a r$ to misbehave on certain surfaces, such as the event horizon of the Schwarzschild spacetime, but we can simply require that radial lines remain continuous as they pass through these surfaces.

[9] On the same surfaces referred to in the preceding footnote, the functions h and k may to go to 0 or ∞. These turn out to be nothing more serious than coordinate singularities.

[10] The Schwarzschild *spacetime* is the uniquely defined geometry found by removing the coordinate singularities from this form of the Schwarzschild *metric*.

7.5: Static and Stationary Spacetimes (Part 2)

No-hair Theorems

Birkhoff's theorem is similar to a set of theorems called no-hair theorems describing black holes. The most general no-hair theorem states that a black hole is completely characterized by its mass, charge, and angular momentum. Other than these three numbers, nobody on the outside can recover any information that was possessed by the matter and energy that were sucked into the black hole.

It has been proposed[11] that the no-hair theorem for nonzero angular momentum and zero charge could be tested empirically by observations of Sagittarius A*. If the observations are consistent with the no-hair theorem, it would be taken as supporting the validity of general relativity and the interpretation of this object as a supermassive black hole. If not, then there are various possibilities, including a failure of general relativity to be the correct theory of strong gravitational fields, or a failure of one of the theorem's other assumptions, such as the nonexistence of closed timelike curves in the surrounding universe.

The no-hair theorems say that relativity only has a small repertoire of types of black-holes, defined as regions of space that are causally disconnected from the universe, in the sense that future light-cones of points in the region do not extend to infinity.[12] That is, a black hole is defined as a region hidden behind an event horizon, and since the definition of an event horizon is dependent on the observer, we specify an observer infinitely far away. Birkhoff's theorem has a somewhat different structure than those of the no-hair theorems, since it assumes a symmetry and proves the existence of an event horizon (if the vacuum region is extended to small enough radii), whereas the no-hair theorems assume an event horizon and prove the form of the metric, including its symmetries.

> **null infinity**
>
> For a more formal statement of this, see Hawking and Ellis, "The Large Scale Structure of Space-Time," p. 315. Essentially, the region must be a connected region on a spacelike three-surface, and there must be no lightlike world-lines that connect points in that region to null infinity. Null infinity was introduced briefly in section 7.3 is defined formally using conformal techniques, but basically refers to points that are infinitely far away in both space and time, and have the two infinities *equal* in a certain sense, so that a free light ray could end up there. The definition is based on the assumption that the surrounding spacetime is asymptotically flat, since otherwise null infinity can't be defined. It is not actually necessary to assume a singularity as part of the definition; the no-hair theorems guarantee that one exists.

The no-hair theorems cannot classify naked singularities, i.e., those not hidden behind horizons. The role of naked singularities in relativity is the subject of the cosmic censorship hypothesis, which is an open problem. The theorems do not rule out the Big Bang singularity, because we cannot define the notion of an observer infinitely far from the Big Bang. We can also see that Birkhoff's theorem does not prohibit the Big Bang, because cosmological models are not vacuum solutions with $\Lambda = 0$. Black string solutions are not ruled out by Birkhoff's theorem because they would lack spherical symmetry, so we need the arguments given in section 6.3 to show that they don't exist.

We saw on example 4 and section 7.3 that there is no clearly defined way to treat a singularity as a geometrical object, and that this ambiguity extends even to such seemingly straightforward questions as how many dimensions it has. Geometrically, as Gertrude Stein said about Oakland, there's "no there there." We could also ask whether a black hole singularity has any *physical properties*. If so, then the no-hair theorems would limit the list of such properties to at most three. But we cannot ascribe these properties to the singularity itself. Rather, they are properties of some large region of the spacetime, measurable by an observer at asymptotic infinity. Such an observer cannot say whether a black hole's mass is a property of the singularity; she cannot even say whether the singularity exists "now." In this sense a black hole singularity is not an "it." Asking about "its" properties is like asking what time it is when the tip of the minute hand is at the center of the clock. The dial only exists around the circumference of the circle, not at its center.

The Gravitational Potential

When Pound and Rebka made the first observation of gravitational redshifts, these shifts were interpreted as evidence of gravitational time dilation, i.e., a mismatch in the rates of clocks. We are accustomed to connecting these two ideas by using the expression $e^{-\Delta\Phi}$ for the ratio of the rates of two clocks (example 11), where Φ is a function of the spatial coordinates, and this is in fact the most general possible definition of a gravitational potential Φ in relativity. Since a stationary field allows us to compare

rates of clocks, it seems that we should be able to define a gravitational potential for any stationary field. There is a problem, however, because when we talk about a potential, we normally have in mind something that has encoded within it all there is to know about the field. We would therefore expect to be able to find the metric from the potential. But the example of the rotating earth shows that this need not be the case for a general stationary field. In that example, there are effects like frame-dragging that clearly cannot be deduced from Φ; for by symmetry, Φ is independent of azimuthal angle, and therefore it cannot distinguish between the direction of rotation and the contrary direction. In a static spacetime, these rotational effects don't exist; a static vacuum spacetime can be described completely in terms of a single scalar potential plus information about the spatial curvature.

There are two main reasons why relativity does not offer a gravitational potential with the same general utility as its Newtonian counterpart.

The Einstein field equations are nonlinear. Therefore one cannot, in general, find the field created by a given set of sources by adding up the potentials. At best this is a possible weak-field approximation. In particular, although Birkhoff's theorem is in some ways analogous to the Newtonian shell theorem, it cannot be used to find the metric of an arbitrary spherically symmetric mass distribution by breaking it up into spherical shells.

It is also not meaningful to talk about any kind of gravitational potential for spacetimes that aren't static or stationary. For example, consider a cosmological model describing our expanding universe. Such models are usually constructed according to the Copernican principle that no position in the universe occupies a privileged place. In other words, they are homogeneous in the sense that they have Killing vectors describing arbitrary translations and rotations. Because of this high degree of symmetry, a gravitational potential for such a model would have to be independent of position, and then it clearly could not encode any information about the spatial part of the metric. Even if we were willing to make the potential a function of time, $\Phi(t)$, the results would still be nonsense. The gravitational potential is defined in terms of rate-matching of clocks, so a potential that was purely a function of time would describe a situation in which all clocks, everywhere in the universe, were changing their rates in a uniform way. But this is clearly just equivalent to a redefi- nition of the time coordinate, which has no observable consequences because general relativity is coordinate-invariant. A corollary is that in a cosmological spacetime, it is not possible to give a natural prescription for deciding whether a particular redshift is gravitational (measured by Φ) or kinematic, or some combination of the two (see also section 8.2).

References

[11] Johannsen and Psaltis, http://arxiv.org/abs/1008.3902v1

7.6: The Uniform Gravitational Field Revisited

This section gives a somewhat exotic example. It is not necessary to read it in order to understand the later material.

In problem 7, we made a wish list of desired properties for a uniform gravitational field, and found that they could not all be satisfied at once. That is, there is no global solution to the Einstein field equations that uniquely and satisfactorily embodies all of our Newtonian ideas about a uniform field. We now revisit this question in the light of our new knowledge.

The 1+1-dimensional metric

$$ds^2 = e^{2gz}dt^2 - dz^2 \qquad\qquad (7.6.1)$$

is the one that uniquely satisfies our expectations based on the equivalence principle (example 11), and it is a vacuum solution. We might logically try to generalize this to 3+1 dimensions as follows:

$$ds^2 = e^{2gz}dt^2 - dx^2 - dy^2 - dz^2. \qquad\qquad (7.6.2)$$

But a funny thing happens now — simply by slapping on the two new Cartesian axes x and y, it turns out that we have made our vacuum solution into a non-vacuum solution, and not only that, but the resulting stress-energy tensor is unphysical (problem 8).

One way to proceed would be to relax our insistence on making the spacetime one that exactly embodies the equivalence principle's requirements for a uniform field.[13] This can be done by taking $g_{tt} = e^{2\Phi}$, where Φ is not necessarily equal to 2gz. By requiring that the metric be a 3+1 vacuum solution, we arrive at a differential equation whose solution is $\Phi = \ln(z + k_1) + k_2$, which recovers the flat-space metric that we found in example 19 by applying a change of coordinates to the Lorentz metric.

> **Note**
>
> Thanks to physicsforums.com user Mentz114 for suggesting this approach and demonstrating the following calculation

What if we want to carry out the generalization from 1+1 to 3+1 without violating the equivalence principle? For physical motivation in how to get past this obstacle, consider the following argument made by Born in 1920.[14] Take a frame of reference tied to a rotating disk, as in the example from which Einstein originally took much of the motivation for creating a geometrical theory of gravity (section 3.5). Clocks near the edge of the disk run slowly, and by the equivalence principle, an observer on the disk interprets this as a gravitational time dilation. But this is not the only relativistic effect seen by such an observer. Her rulers are also Lorentz contracted as seen by a non-rotating observer, and she interprets this as evidence of a non-Euclidean spatial geometry. There are some physical differences between the rotating disk and our default conception of a uniform field, specifically in the question of whether the metric should be static (i.e., lacking in cross-terms between the space and time variables). But even so, these considerations make it natural to hypothesize that the correct 3+1-dimensional metric should have transverse spatial coefficients that decrease with height.

With this motivation, let's consider a metric of the form

$$ds^2 = e^{2z}dt^2 - e^{-2jz}dx^2 - e^{-2kz}dy^2 - dz^2, \qquad\qquad (7.6.3)$$

where j and k are constants, and I've taken g = 1 for convenience.[15] The following Maxima code calculates the scalar curvature and the Einstein tensor:

```
1   load(ctensor);
2   ct_coords:[t,x,y,z];
3   lg:matrix([exp(2*z),0,0,0],
4             [0,-exp(-2*j*z),0,0],
5             [0,0,-exp(-2*k*z),0],
6             [0,0,0,-1]
7   );
8   cmetric();
9   scurvature();
10  leinstein(true);
```

> **Note**
>
> A metric of this general form is referred to as a Kasner metric. One usually sees it written with a logarithmic change of variables, so that z appears in the base rather than in the exponent.

The output from line 9 shows that the scalar curvature is constant, which is a necessary condition for any spacetime that we want to think of as representing a uniform field. Inspecting the Einstein tensor output by line 10, we find that in order to get G_{xx} and G_{yy} to vanish, we need j and k to be $\frac{1 \pm \sqrt{3}i}{2}$. By trial and error, we find that assigning the complex-conjugate values to j and k makes G_{tt} and G_{zz} vanish as well, so that we have a vacuum solution. This solution is, unfortunately, complex, so it is not of any obvious value as a physically meaningful result. Since the field equations are nonlinear, we can't use the usual trick of forming real-valued superpositions of the complex solutions. We could try simply taking the real part of the metric. This gives $g_{xx} = e^{-z} \cos \sqrt{3}z$ and $g_{yy} = e^{-z} \sin \sqrt{3}z$, and is unsatisfactory because the metric becomes degenerate (has a zero determinant) at $z = \frac{n\pi}{2\sqrt{3}}$, where n is an integer.

It turns out, however, that there is a very similar solution, found by Petrov in 1962,[16] that is real-valued. The Petrov metric, which describes a spacetime with cylindrical symmetry, is:

$$ds^2 = -dr^2 - e^{-2r}dz^2 + e^r[2\sin\sqrt{3}r\,d\phi\,dt - \cos\sqrt{3}r(d\phi^2 - dt^2)] \tag{7.6.4}$$

Note that it has many features in common with the complex oscillatory solution we found above. There are transverse length contractions that decay and oscillate in exactly the same way. The presence of the $d\varphi\,dt$ term tells us that this is a non-static, rotating solution — exactly like the one that Einstein and Born had in mind in their prototypical example! We typically obtain this type of effect due to frame dragging by some rotating massive body (see section 4.5), and the Petrov solution can indeed be interpreted as the spacetime that exists in the vacuum on the exterior of an infinite, rigidly rotating cylinder of "dust" (see example 13).

The complicated Petrov metric might seem like the furthest possible thing from a uniform gravitational field, but in fact it is about the closest thing general relativity provides to such a field. We first note that the metric has Killing vectors $\partial_z, \partial_\phi$, and ∂_r, so it has at least three out of the four translation symmetries we expect from a uniform field. By analogy with electromagnetism, we would expect this symmetry to be absent in the radial direction, since by Gauss's law the electric field of a line of charge falls off like $\frac{1}{r}$. But surprisingly, the Petrov metric is also uniform radially. It is possible to give the fourth killing vector explicitly (it is $\partial_r + z\partial_z + (\frac{1}{2})(\sqrt{3}t - \phi)\partial_\phi - (\frac{1}{2})(\sqrt{3}\phi + t)\partial_t)$, but it is perhaps more transparent to check that it represents a field of constant strength (problem 4).

For insight into this surprising result, recall that in our attempt at constructing the Cartesian version of this metric, we ran into the problem that the metric became degenerate at $z = \frac{n\pi}{2\sqrt{3}}$. The presence of the $d\phi\,dt$ term prevents this from happening in Petrov's cylindrical version; two of the metric's diagonal components can vanish at certain values of r, but the presence of the off-diagonal component prevents the determinant from going to zero. (The determinant is in fact equal to −1 everywhere.) What is happening physically is that although the labeling of the ϕ and t coordinates suggests a time and an azimuthal angle, these two coordinates are in fact treated completely symmetrically. At values of r where the cosine factor equals 1, the metric is diagonal, and has signature (t, ϕ, r, z) = (+, −, −, −), but when the cosine equals −1, this becomes (−, +, −, −), so that ϕ is now the timelike coordinate. This

perfect symmetry between ϕ and t is an extreme example of frame-dragging, and is produced because of the specially chosen rate of rotation of the dust cylinder, such that the velocity of the dust at the outer surface is exactly c (or approaches it).

Classically, we would expect that a test particle released close enough to the cylinder would be pulled in by the gravitational attraction and destroyed on impact, while a particle released farther away would fly off due to the centrifugal force, escaping and eventually approaching a constant velocity. Neither of these would be anything like the experience of a test particle released in a uniform field. But consider a particle released at rest in the rotating frame at a radius r_1 for which $\cos \sqrt{3}r_1 = 1$, so that t is the timelike coordinate. The particle accelerates (let's say outward), but at some point it arrives at an r_2 where the cosine equals zero, and the $\phi - t$ part of the metric is purely of the form $d\phi\, dt$. At this location, we can define local coordinates $u = \phi - t$ and $v = \phi + t$, so that the metric depends only on $du^2 - dv^2$. One of the coordinates, say u, is now the timelike one. Since our particle is material, its world-line must be timelike, so it is swept along in the $-\phi$ direction. Gibbons and Gielen show that the particle will now come back inward, and continue forever by oscillating back and forth between two radii at which the cosine vanishes.

Closed Timelike Curves

This oscillation still doesn't sound like the motion of a particle in a uniform field, but another strange thing happens, as we can see by taking another look at the values of r at which the cosine vanishes. At such a value of r, construct a curve of the form (t = constant, r = constant, ϕ, z = constant). This is a closed curve, and its proper length is zero, i.e., it is lightlike. This violates causality.

A photon could travel around this path and arrive at its starting point at the same time when it was emitted. Something similarly weird hapens to the test particle described above: whereas it seems to fall sometimes up and sometimes down, in fact it is always falling down — but sometimes it achieves this by falling up while moving backward in time!

Although the Petov metric violates causality, Gibbons and Gielen have shown that it satisfies the chronology protection conjecture: "In the context of causality violation we have shown that one cannot create CTCs [closed timelike curves] by spinning up a cylinder beyond its critical angular velocity by shooting in particles on timelike or null curves."

We have an exact vacuum solution to the Einstein field equations that violates causality. This raises troublesome questions about the logical self-consistency of general relativity. A very readable and entertaining overview of these issues is given in the final chapter of Kip Thorne's *Black Holes and Time Warps: Einstein's Outrageous Legacy*. In a toy model constructed by Thorne's students, involving a billiard ball and a wormhole, it turned out that there always seemed to be self-consistent solutions to the ball's equations of motion, but they were not unique, and they often involved disquieting possibilities in which the ball went back in time and collided with its earlier self. Among other things, this seems to lead to a violation of conservation of mass-energy, since no mass was put into the system to create extra copies of the ball. This would then be an example of the fact that, as discussed in section 4.5, general relativity does not admit global conservation laws. However, there is also an argument that the mouths of the wormhole change in mass in such a way as to preserve conservation of energy.[17]

References

[14] Max Born, Einstein's *Theory of Relativity*, 1920. In the 1962 Dover edition, the relevant passage is on p. 320

[16] Petrov, in Recent Developments in General Relativity, 1962, Pergamon, p. 383. For a presentation that is freely accessible online, see Gibbons and Gielen, "The Petrov and Kaigorodov-Ozsv´ath Solutions: Spacetime as a Group Manifold," arxiv.org/abs/0802.4082.

[17] http://golem.ph.utexas.edu/string/ar...es/000550.html

7.E: Symmetries (Exercises)

1. Example 3 gave the Killing vectors ∂_z and ∂_ϕ of a cylinder. If we express these instead as two linearly independent Killing vectors that are linear combinations of these two, what is the geometrical interpretation?

2. Section 7.4 told the story of Alice trying to find evidence that her spacetime is not stationary, and also listed the following examples of spacetimes that were not stationary: (a) the solar system, (b) cosmological models, (c) gravitational waves propagating at the speed of light, and (d) a cloud of matter undergoing gravitational collapse. For each of these, show that it is possible for Alice to accomplish her mission.

3. If a spacetime has a certain symmetry, then we expect that symmetry to be detectable in the behavior of curvature scalars such as the scalar curvature $R = R^a_a$ and the Kretchmann invariant $k = R^{abcd}R_{abcd}$.

   a. Show that the metric

   $$ds^2 = e^{2gz}dt^2 - dx^2 - dy^2 - dz^2 \qquad (7.\text{E}.1)$$

   from section 7.5 has constant values of $R = 1/2$ and $k = 1/4$. Note that Maxima's ctensor package has built-in functions for these; you have to call the `lriemann` and `uriemann` before calling them.

   b. Similarly, show that the Petrov metric

   $$ds^2 = -dr^2 - e^{-2r}dz^2 + e^r[2\sin\sqrt{3}rd\phi dt - \cos\sqrt{3}r(d\phi^2 - dt^2)] \qquad (7.\text{E}.2)$$

   has $R = 0$ and $k = 0$.

4. Section 7.5 presented the Petrov metric. The purpose of this problem is to verify that the gravitational field it represents does not fall off with distance. For simplicity, let's restrict our attention to a particle released at an r such that $\cos\sqrt{3}r = 1$, so that t is the timelike coordinate. Let the particle be released at rest in the sense that initially it has $\dot{z} = \dot{r} = \dot{\phi} = 0$, where dots represent differentiation with respect to the particle's proper time. Show that the magnitude of the proper acceleration is independent of r.

5. The idea that a frame is "rotating" in general relativity can be formalized by saying that the frame is stationary but not static. Suppose someone says that any rotation must have a center. Give a counterexample.

CHAPTER OVERVIEW

8: Sources

8.1: Sources in General Relativity (Part 1)

Point Sources in a Background-independent Theory

The Schrödinger equation and Maxwell's equations treat spacetime as a stage on which particles and fields act out their roles. General relativity, however, is essentially a theory of spacetime itself. The role played by atoms or rays of light is so peripheral that by the time Einstein had derived an approximate version of the Schwarzschild metric, and used it to find the precession of Mercury's perihelion, he still had only vague ideas of how light and matter would fit into the picture. In his calculation, Mercury played the role of a test particle: a lump of mass so tiny that it can be tossed into spacetime in order to measure spacetime's curvature, without worrying about its effect on the spacetime, which is assumed to be negligible. Likewise the sun was treated as in one of those orchestral pieces in which some of the brass play from off-stage, so as to produce the effect of a second band heard from a distance. Its mass appears simply as an adjustable parameter m in the metric, and if we had never heard of the Newtonian theory we would have had no way of knowing how to interpret m.

When Schwarzschild published his exact solution to the vacuum field equations, Einstein suffered from philosophical indigestion. His strong belief in Mach's principle led him to believe that there was a paradox implicit in an exact spacetime with only one mass in it. If Einstein's field equations were to mean anything, he believed that they had to be interpreted in terms of the motion of one body relative to another. In a universe with only one massive particle, there would be no relative motion, and so, it seemed to him, no motion of any kind, and no meaningful interpretation for the surrounding spacetime.

Not only that, but Schwarzschild's solution had a singularity at its center. When a classical field theory contains singularities, Einstein believed, it contains the seeds of its own destruction. As we've seen in section 6.3, this issue is still far from being resolved, a century later.

However much he might have liked to disown it, Einstein was now in possession of a solution to his field equations for a point source. In a linear, background-dependent theory like electromagnetism, knowledge of such a solution leads directly to the ability to write down the field equations with sources included. If Coulomb's law tells us the $\frac{1}{r^2}$ variation of the electric field of a point charge, then we can infer Gauss's law. The situation in general relativity is not this simple. The field equations of general relativity, unlike the Gauss's law, are nonlinear, so we can't simply say that a planet or a star is a solution to be found by adding up a large number of point-source solutions. It's also not clear how one could represent a moving source, since the singularity is a point that is not even part of the continuous structure of spacetime (and its location is also hidden behind an event horizon, so it **cannot** be observed from the outside).

The Einstein Field Equation

The Einstein Tensor

Given these difficulties, it's not surprising that Einstein's first attempt at incorporating sources into his field equation was a dead end. He postulated that the field equation would have the Ricci tensor on one side, and the stress-energy tensor T^{ab} (Section 5.2) on the other,

$$R_{ab} = 8\pi T_{ab}, \tag{8.1.1}$$

where a factor of $\frac{G}{c^4}$ on the right is suppressed by our choice of units, and the 8π is determined on the basis of consistency with Newtonian gravity in the limit of weak fields and low velocities. The problem with this version of the field equations can be demonstrated by counting variables. R and T are symmetric tensors, so the field equation contains 10 constraints on the metric: 4 from the diagonal elements and 6 from the off-diagonal ones.

In addition, local conservation of mass-energy requires the divergence-free property $\nabla_b T^{ab} = 0$. In order to construct an example, we recall that the only component of T for which we have so far introduced any physical interpretation is T^{tt}, which gives the density of mass-energy. Suppose we had a stress-energy tensor whose components were all zero, except for a time-time component varying as $T^{tt} = kt$. This would describe a region of space in which mass-energy was uniformly appearing or disappearing everywhere at a constant rate. To forbid such examples, we need the divergence-free property to hold. This is exactly analogous to the continuity equation in fluid mechanics or electromagnetism,

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J} = 0 \tag{8.1.2}$$

(or $\nabla_a J^a = 0$), which states that the quantity of fluid or charge is conserved.

But imposing the divergence-free condition adds four more constraints on the metric, for a total of 14. The metric, however, is a symmetric rank-2 tensor itself, so it only has 10 independent components. This overdetermination of the metric suggests that the proposed field equation will not in general allow a solution to be evolved forward in time from a set of initial conditions given on a spacelike surface, and this turns out to be true. It can in fact be shown that the only possible solutions are those in which the traces $R = R^a{}_a$ and $T = T^a_a$ are constant throughout spacetime.

The solution is to replace R_{ab} in the field equations with a different tensor G_{ab}, called the *Einstein tensor*, defined by

$$G_{ab} = R_{ab} - \left(\frac{1}{2}\right) R g_{ab} \tag{8.1.3}$$

or

$$G_{ab} = 8\pi T_{ab} \tag{8.1.4}$$

The Einstein tensor is constructed exactly so that it is divergence-free, $\nabla_b G^{ab} = 0$. (This is not obvious, but can be proved by direct computation.) Therefore any stress-energy tensor that satisfies the field equation is automatically divergence-less, and thus no additional constraints need to be applied in order to guarantee conservation of mass-energy.

> ### Exercise 8.1.1:
>
> Self-check: Does replacing R_{ab} with G_{ab} invalidate the Schwarzschild metric?

This procedure of making local conservation of mass-energy "baked in" to the field equations is analogous to the way conservation of charge is treated in electricity and magnetism, where it follows from Maxwell's equations rather than having to be added as a separate constraint.

Interpretation of the Stress-energy Tensor

The stress-energy tensor was briefly introduced in section 5.2. By applying the Newtonian limit of the field equation to the Schwarzschild metric, we find that T^{tt} is to be identified as the mass density ρ. The Schwarzschild metric describes a spacetime using coordinates in which the mass is at rest. In the cosmological applications we'll be considering shortly, it also makes sense to adopt a frame of reference in which the local mass-energy is, on average, at rest, so we can continue to think of T^{tt} as the (average) mass density. By symmetry, T must be diagonal in such a frame. For example, if we had $T^{tx} \neq 0$, then the positive x direction would be distinguished from the negative x direction, but there is nothing that would allow such a distinction.

> ### Example 1: Dust in a different frame
>
> As discussed in example 14, it is convenient in cosmology to distinguish between radiation and "dust," meaning noninteracting, nonrelativistic materials such as hydrogen gas or galaxies. Here "nonrelativistic" means that in the comoving frame, in which the average flow of dust vanishes, the dust particles all have $|v| \ll 1$. What is the stress-energy tensor associated with dust?
>
> Since the dust is nonrelativistic, we can obtain the Newtonian limit by using units in which $c \neq 1$, and letting c approach infinity. In Cartesian coordinates, the components of the stress-energy have units that cause them to scale like
>
> $$T^{\mu\nu} \propto \begin{pmatrix} 1 & \frac{1}{c} & \frac{1}{c} & \frac{1}{c} \\ \frac{1}{c} & \frac{1}{c^2} & \frac{1}{c^2} & \frac{1}{c^2} \\ \frac{1}{c} & \frac{1}{c^2} & \frac{1}{c^2} & \frac{1}{c^2} \\ \frac{1}{c} & \frac{1}{c^2} & \frac{1}{c^2} & \frac{1}{c^2} \end{pmatrix} \tag{8.1.5}$$
>
> In the limit of $c \to \infty$, we can therefore take the only source of gravitational fields to be T^{tt}, which in Newtonian gravity must be the mass density ρ, so

$$T^{\mu\nu} = \begin{pmatrix} \rho & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{8.1.6}$$

Under a Lorentz boost by v in the x direction, the tensor transformation law gives

$$T^{\mu'\nu'} = \begin{pmatrix} \gamma^2\rho & \gamma^2 v\rho & 0 & 0 \\ \gamma^2 v\rho & \gamma^2 v^2\rho & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{8.1.7}$$

The over-all factor of γ^2 arises because of the combination of two effects: each dust particle's mass-energy is increased by a factor of γ, and length contraction also multiplies the density of dust particles by a factor of γ. In the limit of small boosts, the stress-energy tensor becomes

$$T^{\mu'\nu'} \approx \begin{pmatrix} \rho & v\rho & 0 & 0 \\ v\rho & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{8.1.8}$$

This motivates the interpretation of the time-space components of T as the flux of mass-energy along each axis. In the primed frame, mass-energy with density ρ flows in the x direction at velocity v, so that the rate at which mass-energy passes through a window of area A in the y − z plane is given by $\rho v A$.

This is also consistent with our imposition of the divergence-free property, by which we were essentially stating T tx to be the rate of flow of T^{tt}.

Example 2: The center of mass-energy

In Newtonian mechanics, for motion in one dimension, the total momentum of a system of particles is given by $p_{tot} = Mv_{cm}$, where M is the total mass and v_{cm} the velocity of the center of mass. Is there such a relation in relativity?

Since mass and energy are equivalent, we expect that the relativistic equivalent of the center of mass would have to be a center of mass-energy.

It should also be clear that a center of mass-energy can only be well defined for a region of spacetime that is small enough so that effects due to curvature are negligible. For example, we can have cosmological models in which space is finite, and expands like the surface of a balloon being blown up. If the model is homogeneous (there are no "special points" on the surface of the balloon), then there is no point in space that could be a center. (A real balloon has a center, but in our metaphor only the balloon's spherical surface correponds to physical space.) The fundamental issue here is the same geometrical one that caused us to conclude that there is no global conservation of mass-energy in general relativity (see section 4.5). In a curved spacetime, parallel transport is pathdependent, so we can't unambiguously define a way of adding vectors that occur in different places. The center of mass is defined by a sum of position vectors. From these considerations we conclude that the center of mass-energy is only well defined in special relativity, not general relativity.

For simplicity, let's restrict ourselves to 1+1 dimensions, and adopt a frame of reference in which the center of mass is at rest at x = 0.

Since T^{tt} is interpreted as the density of mass-energy, the position of the center of mass must be given by

$$0 = \int x T^{tt} dx. \tag{8.1.9}$$

By analogy with the Newtonian relation $p_{tot} = Mv_{cm}$, let's see what happens when we differentiate with respect to time. The velocity of the center of mass is then $0 = \frac{dx_{cm}}{dt} = \int \partial_t T^{tt} x\, dx$. Applying the divergence-free property $\partial_t T^{tt} + \partial_x T^{tx} = 0$, this becomes $0 = -\int \partial_x T^{tx} x\, dx$. Integration by parts gives us finally

$$0 = \int T^{tx} dx. \qquad (8.1.10)$$

We've already interpreted T^{tx} as the rate of flow of mass-energy, which is another way of describing momentum. We can therefore interpret T^{tx} as the density of momentum, and the right-hand side of this equation as the total momentum. The interpretation is that a system's center of mass-energy is at rest if and only if it has zero total momentum.

Suppose, for example, that we prepare a uniform metal rod so that one end is hot and the other cold. We then deposit it in outer space, initially motionless relative to some observer. Although the rod itself is uniform, its mass-energy is very slightly nonuniform, so its center of mass-energy must be displaced a tiny bit away from the center, toward the hot end. As the rod approaches thermal equilibrium, the observer sees it accelerate very slightly and then come to rest again, so that its center of mass-energy remains fixed! An even stranger case is described in example 9.

Since the Einstein tensor is symmetric, the Einstein field equation requires that the stress-energy tensor be symmetric as well. It is reassuring that according to example 1 the tensor is symmetric for dust, and that symmetry is preserved by changes of coordinates and by superpositions of sources. Besides dust, the other cosmologically significant sources of gravity are electromagnetic radiation and the cosmological constant, and one can also check that these give symmetry. Belinfante noted in 1939 that symmetry seemed to fail in the case of fields with intrinsic spin, but he found that this problem could be avoided by modifying the previously assumed way of connecting T to the properties of the field. This shows that it can be rather subtle to interpret the stress-energy tensor and connect it to experimental observables. For more on this connection, and the case of electromagnetic fields, see examples 7 and 8.

In example 1, we found that T^{xt} had to be interpreted as the flux of T^{tt} (i.e., the flux of mass-energy) across the x axis. Lorentz invariance requires that we treat t, x, y, and z symmetrically, and this forces us to adopt the following interpretation: T$^{\mu\nu}$, where μ is spacelike, is the flux of the density of the mass-energy four-vector in the μ direction. In the comoving frame, in Cartesian coordinates, this means that T^{xx}, T^{yy}, and T^{zz} should be interpreted as pressures. For example, T^{xx} is the flux in the x direction of x-momentum. This is simply the pressure, P, that would be exerted on a surface with its normal in the x direction, so in the comoving frame we have T$^{\mu\nu}$ = diag(ρ, P, P, P). For a fluid that is not in equilibrium, the pressure need not be isotropic, and the stress exerted by the fluid need not be perpendicular to the surface on which it acts. The space-space components of T would then be the classical stress tensor, whose diagonal elements are the anisotropic pressure, and whose off-diagonal elements are the shear stress. This is the reason for calling T the stress-energy tensor.

The prediction of general relativity is then that pressure acts as a gravitational source with exactly the same strength as mass-energy density. This has important implications for cosmology, since the early universe was dominated by radiation, and a photon gas has P = $\frac{\rho}{3}$ (example 14).

Experimental Tests

But how do we know that this prediction is even correct? Can it be verified in the laboratory? The classic laboratory test of the strength of a gravitational source is the 1797 Cavendish experiment, in which a torsion balance was used to measure the very weak gravitational attractions between metal spheres. We could test this aspect of general relativity by doing a Cavendish experiment with boxes full of photons, so that the pressure is of the same order of magnitude as the mass-energy. This is unfortunately utterly impractical, since both P and ρ for a well-lit box are ridiculously small compared to ρ for a metal ball.
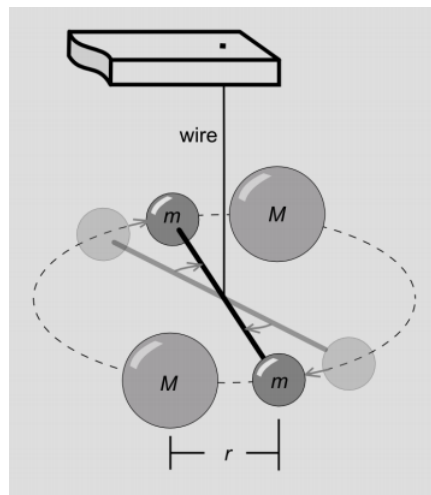
Figure_8.1.1b.png" />

Figure 8.1.1 - *A Cavendish balance, used to determine the gravitational constant.*

However, the repulsive electromagnetic pressure inside an atomic nucleus is quite large by ordinary standards — about 10^{33} Pa! To see how big this is compared to the nuclear mass density of $\rho \sim 10^{18}$ kg/m^3, we need to take into account the factor of $c^2 \neq 1$ in SI units, the result being that $\frac{P}{\rho}$ is about 10^{-2}, which is not too small. Thus if we measure gravitational interactions of nuclei with different values of $\frac{P}{\rho}$, we should be able to test this prediction of general relativity. This was done in a Princeton PhD-thesis experiment by Kreuzer[1] in 1966.

Before we can properly describe and interpret the Kreuzer experiment, we need to distinguish the several different types of mass that could in principle be different from one another in a theory of gravity. We've already encountered the distinction between inertial and gravitational mass, which Eötvös experiments show to be equivalent to about one part in 10^{12}. But there is also a distinction between an object's *active* gravitational mass m_a, which measures its ability to create gravitational fields, and its *passive* gravitational mass m_p, which measures the force it feels when placed in an externally generated field. For experiments using laboratoryscale material objects at nonrelativistic velocities, the Newtonian limit applies, and we can think of active gravitational mass as a scalar, with a density $T^{tt} = \rho$.

To understand how this relates to pressure as a source of gravitational fields, it is helpful to consider a case where P is about the same as ρ, which occurs for light. Light is inherently relativistic, so the Newtonian concept of a scalar gravitational mass breaks down, but we can still use "mass" in quotes to talk qualitatively about an electromagnetic wave's active and passive participation in gravitational effects. Experiments show that general relativity correctly predicts the deflection of light by the sun to about one part in 10^5 (section 6.2). This is the electromagnetic equivalent of an Eötvös experiment; it shows that general relativity predicts the right thing about the proportion between a light wave's inertial and passive gravitational "masses." Now suppose that general relativity was wrong, and pressure was not a source of gravitational fields. This would cause a drastic decrease in the active gravitational "mass" of an electromagnetic wave.

The Kreuzer experiment actually dealt with static electric fields inside nuclei, not electromagnetic waves, but it is still clear what we should expect in general: if pressure does not act as a gravitational source, then the ratio $\frac{m_a}{m_p}$ should be different for different nuclei. Specifically, it should be lower for a nucleus with a higher atomic number Z, in which the electrostatic pressures are higher.

Kreuzer did a Cavendish experiment, Figure 8.1.2, using masses made of two different substances. The first substance was teflon. The second substance was a mixture of the liquids trichloroethylene and dibromoethane, with the proportions chosen so as to give a density as close as possible to that of teflon. Teflon is 76% fluorine by weight, and the liquid is 74% bromine. Fluorine has atomic number Z = 9, bromine Z = 35, and since the electromagnetic force has a long range, the pressure within a nucleus scales upward roughly like $Z^{1/3}$ (because any given proton is acted on by Z − 1 other protons, and the size of a nucleus scales like Z 1/3 , so $P \propto \frac{Z}{(Z^{1/3})^2}$. The solid mass was immersed in the liquid, and the combined gravitational field of the solid and the liquid was detected by a Cavendish balance.
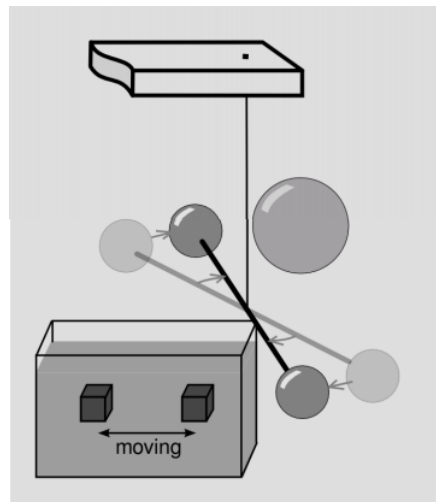
Figure 8.1.2 - A simplified diagram of Kreuzer's modification. The moving teflon mass is submerged in a liquid with nearly the same density.

Ideally, one would formulate the liquid mixture so that its passivemass density was exactly equal to that of teflon, as determined by buoyancy. Any oscillation in the torque measured by the Cavendish balance would then indicate an inequivalence between active and passive gravitational mass.
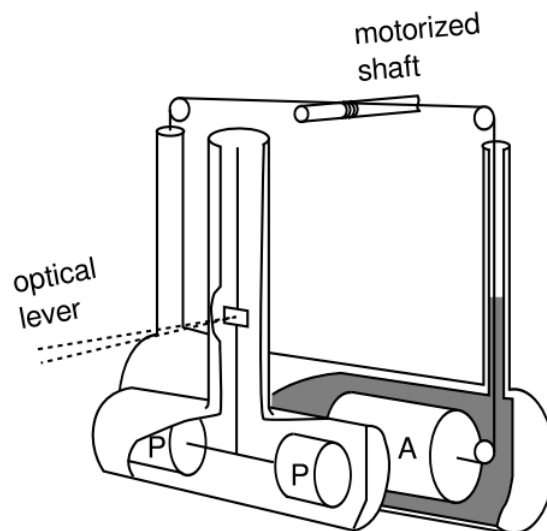
Figure_8.1.3b.png" />
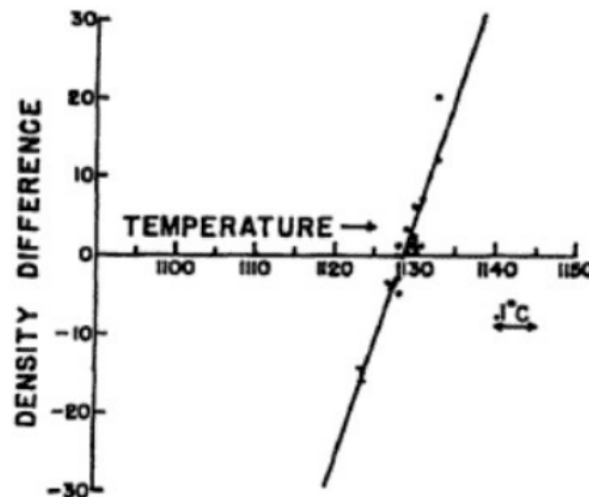
Figure 8.1.3 - The Kreuzer experiment. 1. There are two passive masses, P, and an active mass A consisting of a single 23-cm diameter teflon cylinder immersed in a fluid. The teflon cylinder is driven back and forth with a period of 400 s. The resulting deflection of the torsion beam is monitored by an optical lever and canceled actively by electrostatic forces from capacitor plates (not shown). The voltage required for this active cancellation is a measure of the torque exerted by A on the torsion beam. 2. Active mass as a function of temperature. 3. Passive mass as a function of temperature. In both 2 and 3, temperature is measured in units of ohms, i.e., the uncalibrated units of a thermistor that was immersed in the liquid.

In reality, the two substances involved had different coefficients of thermal expansion, so slight variations in temperature made their passive-mass densities unequal. Kreuzer therefore measured both the buoyant force and the gravitational torque as functions of temperature. He determined that these became zero at the same temperature, to within experimental errors, which verified the equivalence of active and passive gravitational mass to within a certain precision,

$$m_p \propto m_a \tag{8.1.11}$$

to within 5×10^{-5}.

Kreuzer intended this exeriment only as a test of $m_p \propto m_a$, but it was reinterpreted in 1976 by Will[2] as a test of the coupling of sources to gravitational fields as predicted by general relativity and other theories of gravity. Crudely, we've already argued that mp ∝ ma would be substance-dependent if pressure did not couple to gravitational fields. Will actually carried out a more careful calculation, of which I present a simplified summary. Suppose that pressure does not contribute as much to gravitational fields as is claimed by general relativity; its coupling is reduced by a factor $1 - x$, where x = 0 in general relativity.[3] Will considers a model consisting of pointlike particles interacting through static electrical forces, and shows that for such a system,

$$m_a = m_p + \frac{1}{2} x U_e, \tag{8.1.12}$$

where U_e is the electrical energy. The Kreuzer experiment then requires $|x| < 6 \times 10^{-2}$, meaning that pressure does contribute to gravitational fields as predicted by general relativity, to within a precision of 6%.

> **Note**
>
> In Will's notation, ζ_4 measures nonstandard coupling to pressure, ζ_3 to internal energy, and ζ_1 to kinetic energy. By requiring that point-particle models agree with perfect-fluid models, one obtains $(-\frac{2}{3})\zeta_1 = \zeta_3 = -\zeta_4 = x$.

One of the important ways in which Will's calculation goes beyond my previous crude argument is that it shows that when x = 0, as it does for general relativity, the correction term $\frac{x U_e}{2}$ vanishes, and $m_a = m_p$ exactly. This is interpreted as follows. Let a bromine nucleus be referred to with a capital M, fluorine with the lowercase m. Then when a bromine nucleus and a fluorine nucleus interact gravitationally at a distance r, the Newtonian approximation applies, and the total internal force acting on the pair of nuclei taken as a whole equals $\frac{m_p M_a - M_p m_a}{r^2}$ (in units where the Newtonian gravitational constant G equals 1). This vanishes only if $m_p M_a - M_p m_a = 0$, which is equivalent to $\frac{m_p}{M_p} = \frac{m_a}{M_a}$. If this proportionality fails, then the system violates Newton's third law and

conservation of momentum; its center of mass will accelerate along the line connecting the two nuclei, either in the direction of M or in the direction of m, depending on the sign of x.

Thus the vanishing of the correction term $\frac{xU_e}{2}$ tells us that general relativity predicts exact conservation of momentum in this interaction. This is comforting, but a little surprising on the face of it. Newtonian gravity treats active and passive massive perfectly symmetrically, so that there is a perfect guarantee of conservation of momentum. But relativity incorporates them in a completely asymmetric manner, so there is no obvious reason that we should have perfect conservation of momentum. In fact we don't have any general guarantee of conservation of momentum, since, as discussed in section 4.5, the language of general relativity doesn't even give us the symbols we would need in order to state a global conservation law for a vector. General relativity does, however, allow *local* conservation laws. We will have local conservation of mass-energy and momentum provided that the stress-energy tensor's divergence $\nabla_b T^{ab}$ vanishes.

Bartlett and van Buren[4] used this connection to conservation of momentum in 1986 to derive a tighter limit on x. Since the moon has an asymmetrical distribution of iron and aluminum, a nonzero x would cause it to have an anomalous acceleration along a certain line. Because lunar laser ranging gives extremely accurate data on the moon's orbit, the constraint is tightened to $|x| < 1 \times 10^{-8}$.
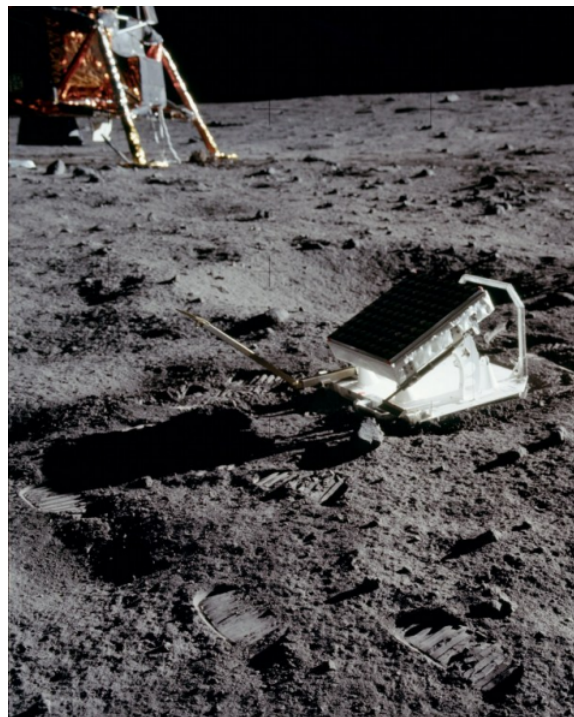
Figure 8.1.4 - The Apollo 11 mission left behind this mirror, which in this photo shows the reflection of the black sky. The mirror is used for lunar laser ranging measurements, which have an accuracy of about a centimeter.

These are tests of general relativity's predictions about the gravitational fields generated by the pressure of a static electric field. In addition, there is indirect confirmation (section 8.2) that general relativity is correct when it comes to electromagnetic waves.

References

[1] Kreuzer, Phys. Rev. 169 (1968) 1007

[2] Will, "Active mass in relativistic gravity: Theoretical interpretation of the Kreuzer experiment," Ap. J. 204 (1976) 234, available online at adsabs. harvard.edu. A broader review of experimental tests of general relativity is given in Will, "The Confrontation between General Relativity and Experiment," relativity.livingreviews.org/Articles/lrr-2006-3/. The Kreuzer experiment is discussed in section 3.7.

[4] Phys. Rev. Lett. 57 (1986) 21. The result is summarized in section 3.7 of the review by Will.

8.2: Sources in General Relativity (Part 2)

Energy of Gravitational Fields Not Included in the Stress-energy Tensor

Summarizing the story of the Kreuzer and Bartlett-van Buren results, we find that observations verify to high precision one of the defining properties of general relativity, which is that *all* forms of energy are equivalent to mass. That is, Einstein's famous E = mc^2 can be extended to gravitational effects, with the proviso that the source of gravitational fields is not really a scalar m but the stressenergy tensor T.

But there is an exception to this even-handed treatment of all types of energy, which is that the energy of the gravitational field itself is not included in T, and is not even generally a well-defined concept locally. In Newtonian gravity, we can have conservation of energy if we attribute to the gravitational field a negative potential energy density $-\frac{\mathbf{g}^2}{8\pi}$. But the equivalence principle tells us that **g** is not a tensor, for we can always make **g** vanish locally by going into the frame of a free-falling observer, and yet the tensor transformation laws will never change a nonzero tensor to a zero tensor under a change of coordinates. Since the gravitational field is not a tensor, there is no way to add a term for it into the definition of the stress-energy, which is a tensor. The grammar and vocabulary of the tensor notation are specifically designed to prevent writing down such a thing, so that the language of general relativity is not even capable of expressing the idea that gravitational fields would themselves contribute to T.

> **exercise 8.2.2**
>
> Self-check: (1) Convince yourself that the negative sign in the expression $-\frac{\mathbf{g}^2}{8\pi}$ makes sense, by considering the case where two equal masses start out far apart and then fall together and combine to make a single body with twice the mass. (2) The Newtonian gravitational field is the gradient of the gravitational potential ϕ, which corresponds in the Newtonian limit to the time-time component of the metric. With this motivation, suppose someone proposes generalizing the Newtonian energy density $-\frac{(\nabla\phi)^2}{8\pi}$ to a similar expression such as $-(\nabla_a g_b^a)(\nabla_c g_c^b)$, where ∇ is now the covariant derivative, and g is the metric, not the Newtonian field strength. What goes wrong?

As a concrete example, we observe that the Hulse-Taylor binary pulsar system (section 6.2) is gradually losing orbital energy, and that the rate of energy loss exactly matches general relativity's prediction of the rate of gravitational radiation. There is a net decrease in the forms of energy, such as rest mass and kinetic energy, that are accounted for in the stress energy tensor T. We can account for the missing energy by attributing it to the outgoing gravitational waves, but that energy is not included in T, and we have to develop special techniques for evaluating that energy. Those techniques only turn out to apply to certain special types of spacetimes, such as asymptotically flat ones, and they do not allow a uniquely defined energy density to be attributed to a particular small region of space (for if they did, that would violate the equivalence principle).

> **Example 3: Gravitational energy is locally unmeasurable**
>
> When a new form of energy is discovered, the way we establish that it is a form of energy is that it can be transformed to or from other forms of energy. For example, Becquerel discovered radioactivity by noticing that photographic plates left in a desk drawer along with radium salts became clouded: some new form of energy had been converted into previously known forms such as chemical energy. It is only in this limited sense that energy is ever locally observable, and this limitation prevents us from meaningfully defining certain measures of energy. For example we can never measure the local electrical potential in the same sense that we can measure the local barometric pressure; a potential of 137 volts only has meaning relative to some other region of space taken to be at ground. Let's use the acronym MELT to refer to measurement of energy by the local transformation of that energy from one form into another.
>
> The reason MELT works is that energy (or actually the momentum four-vector) is locally conserved, as expressed by the zerodivergence property of the stress-energy tensor. Without conservation, there is no notion of transformation. The Einstein field equations imply this zero-divergence property, and the field equations have been well verified by a variety of observations, including many observations (such as solar system tests and observation of the Hulse-Taylor system) that in Newtonian terms would be described as involving (non-local) transformations between kinetic energy and the energy of the gravitational field. This agreement with observation is achieved by taking T = 0 in vacuum, regardless of the gravitational field. Therefore any local transformation of gravitational field energy into another form of energy would be inconsistent with previous observation. This implies that MELT is impossible for gravitational field energy.

In particular, suppose that observer A carries out a local MELT of gravitational field energy, and that A sees this as a process in which the gravitational field is reduced in intensity, causing the release of some other form of energy such as heat. Now consider the situation as seen by observer B, who is free-falling in the same local region. B says that there was never any gravitational field in the first place, and therefore sees heat as violating local conservation of energy. In B's frame, this is a nonzero divergence of the stress-energy tensor, which falsifies the Einstein field equations.

Some Examples

We conclude this introduction to the stress-energy tensor with some illustrative examples.

Example 4: A perfect fluid

For a perfect fluid, we have $$T_{ab} = (\rho + P) v_a v_b - sPg_{ab}, \tag{8.1.11}$$

where s = 1 for our + − −− signature or −1 for the signature − + ++, and v represents the coordinate velocity of the fluid's rest frame.

Suppose that the metric is diagonal, but its components are varying, $g_{\alpha\beta}$ = diag(A^2, −B^2, . . .). The properly normalized velocity vector of an observer at (coordinate-)rest is v^α = (A^{-1}, 0, 0, 0). Lowering the index gives v_α = (sA, 0, 0, 0). The various forms of the stress-energy tensor then look like the following:

$$
\begin{aligned}
T_{00} &= A^2 \rho & T_{11} &= B^2 P \\
T^0_0 &= s\rho & T^1_1 &= -sP \\
T^{00} &= A^{-2} \rho & T^{11} &= B^{-2} P.
\end{aligned}
$$

Example 5: A rope dangling in a Schwarzschild spacetime

Suppose we want to lower a bucket on a rope toward the event horizon of a black hole. We have already made some qualitative remarks about this idea in example 14 on p. 64. This seemingly whimsical example turns out to be a good demonstration of some techniques, and can also be used in thought experiments that illustrate the definition of mass in general relativity and that probe some ideas about quantum gravity.[5]

The Schwarzschild metric (section 6.2) is

$$ds^2 = t^2 dt^2 - t^{-2} dr^2 + \ldots, \tag{8.1.12}$$

where f = (1 − \f(\frac{2m}{r}\))$^{1/2}$, and . . . represents angular terms. We will end up needing the following Christoffel symbols:

$$
\Gamma^t_{tr} = \frac{f'}{f}
$$
$$
\Gamma^\theta_{\theta r} = \Gamma^\phi_{\phi r} = r^{-1}
$$

Since the spacetime has spherical symmetry, it ends up being more convenient to consider a rope whose shape, rather than being cylindrical, is a cone defined by some set of (θ, ϕ). For convenience we take this set to cover unit solid angle. The final results obtained in this way can be readily converted into statements about a cylindrical rope. We let μ be the mass per unit length of the rope, and T the tension. Both of these may depend on r. The corresponding energy density and tensile stress are $\rho = \frac{\mu}{A} = \frac{\mu}{r^2}$ and S = $\frac{T}{A}$. To connect this to the stress-energy tensor, we start by comparing to the case of a perfect fluid from example 4. Because the rope is made of fibers that have stength only in the radial direction, we will have $T^{\theta\theta} = T^{\phi\phi} = 0$. Furthermore, the stress is tensile rather than compressional, corresponding to a negative pressure. The Schwarzschild coordinates are orthogonal but not orthonormal, so the properly normalized velocity of a static observer has a factor of f in it: v^α = (f^{-1}, 0, 0, 0), or, lowering an index, v_α = (f, 0, 0, 0). The results of example 4 show that the mixed-index form of T will be the most convenient, since it can be expressed without messy factors of f. We have

$$T^\kappa_\nu = diag(\rho, S, 0, 0) = r^{-2} diag(\mu, T, 0, 0). \tag{8.1.13}$$

By writing the stress-energy tensor in this form, which is independent of t, we have assumed static equilibrium outside the event horizon. Inside the horizon, the r coordinate is the timelike one, the spacetime itself is not static, and we do not expect to find static solutions, for the reasons given in example 14.

Conservation of energy is automatically satisfied, since there is no time dependence. Conservation of radial momentum is expressed by

$$\nabla_\kappa T_r{}^\kappa = 0, \tag{8.1.14}$$

or

$$0 = \nabla_r T_r{}^r + \nabla_t T_r{}^t + \nabla_\theta T_r{}^\theta + \nabla_\phi T_r{}^\phi. \tag{8.1.15}$$

It would be tempting to throw away all but the first term, since T is diagonal, and therefore $T_r{}^t = T_r{}^\theta = T_r{}^\phi = 0$. However, a covariant derivative can be nonzero even when the symbol being differentiated vanishes identically. Writing out these four terms, we have

$$
\begin{aligned}
0 \ &= \partial_r T_r{}^r + \Gamma^r_{rr} T_r{}^r - \Gamma^r_{rr} T_r{}^r \\
&+ \Gamma^t_{tr} T_r{}^r - \Gamma^t_{tr} T_t{}^t \\
&+ \Gamma^\theta_{\theta r} T_r{}^r \\
&+ \Gamma^\phi_{\phi r} T_r{}^r,
\end{aligned}
$$

where each line corresponds to one covariant derivative. Evaluating this, we have

$$0 = T' + \frac{f'}{f} T - \frac{f'}{f}\mu, \tag{8.1.16}$$

where primes denote differentiation with respect to r. Note that since no terms of the form $\partial_r T_t{}^t$ occur, this expression is valid regardless of whether we take μ to be constant or varying. Thus we are free to take $\rho \propto r^{-2}$, so that μ is constant, and this means that our result is equally applicable to a uniform cylindrical rope. This result is checked using computer software in example 6.

This is a differential equation that tells us how the tensile stress in the rope varies along its length. The coefficient $\frac{f'}{f} = \frac{m}{r(r-2m)}$ blows up at the event horizon, which is as expected, since we do not expect to be able to lower the rope to or below the horizon.

Let's check the Newtonian limit, where the gravitational field is g and the potential is Φ. In this limit, we have f ≈ 1 − Φ, $\frac{f'}{f} \approx$ g (with g > 0), and $\mu \gg$ T, resulting in

$$0 = T' - g\mu. \tag{8.1.17}$$

which is the expected Newtonian relation.

Returning to the full general-relativistic result, it can be shown that for a loaded rope with no mass of its own, we have a finite result for $lim_{r\to\infty}$ T, even when the bucket is brought arbitrarily close to the horizon. (The solution in this case is just T = $\frac{T_\infty}{f}$, where T_∞ is the tension at r = ∞.) However, this is misleading without the caveat that for $\mu <$ T, the speed of transverse waves in the rope is greater than c, which is not possible for any known form of matter — it would violate the null energy condition, discussed in the following section.

---

[5] Brown, "Tensile Strength and the Mining of Black Holes," arxiv.org/abs/ 1207.3342

Example 6: The rope, using computer algebra

The result of example 5 can be checked with the following Maxima code:

```
1   load(ctensor);
2   ct_coords:[t,r,theta,phi];
3   depends(f,r);
4   depends(ten,r); /* tension depends on r */
5   depends(mu,r); /* mass/length depends on r */
6   lg:matrix([f^2,0,0,0],
7              [0,-f^-2,0,0],
8              [0,0,-r^2,0],
9              [0,0,0,-r^2*sin(theta)^2]);
10  cmetric();
11  christof(mcs);
12  /* stress-energy tensor, T^mu_nu */
13  t:r^-2*matrix(
14     [mu,0,0,0],
15     [0,ten,0,0],
16     [0,0,0,0],
17     [0,0,0,0]
18  );
19  /*
20     Compute covariant derivative of the stress-energy
21     tensor with respect to its first index. The
22     function checkdiv is defined so that the first
23     index has to be covariant (lower); the T I'm
24     putting in is T^mu_nu, and since it's symmetric,
25     that's the same as T_mu^nu.
26  */
27  checkdiv(t);
```

Energy Conditions

Physical theories are supposed to answer questions. For example:

1. Does a small enough physical object always have a world-line that is approximately a geodesic?
2. Do massive stars collapse to form black-hole singularities?
3. Did our universe originate in a Big Bang singularity?
4. If our universe doesn't currently have violations of causality such as the closed timelike curves exhibited by the Petrov metric (section 7.5), can we be assured that it will never develop causality violation in the future?

We would like to "prove" whether the answers to questions like these are yes or no, but physical theories are not formal mathematical systems in which results can be "proved" absolutely. For example, the basic structure of general relativity isn't a set of axioms but a list of ingredients like the equivalence principle, which has evaded formal definition.[6]

Even the Einstein field equations, which appear to be completely well defined, are not mathematically formal predictions of the behavior of a physical system. The field equations are agnostic on the question of what kinds of matter fields contribute to the stress-energy tensor. In fact, any spacetime at all is a solution to the Einstein field equations, provided we're willing to admit the corresponding stress-energy tensor. We can never answer questions like the ones above without assuming something about the stress-energy tensor.

In example 14, we saw that radiation has P = $\frac{\rho}{3}$ and dust has P = 0. Both have $\rho \geq 0$. If the universe is made out of nothing but dust and radiation, then we can obtain the following four constraints on the energy-momentum tensor:

trace energy condition	$\rho - 3P \geq 0$
strong energy condition	$\rho + 3P \geq 0 \ and \ \rho + P \geq 0$
dominant energy condition	$\rho \geq 0 \ and \ |P| \leq \rho$
weak energy condition	$\rho \geq 0 \ and \ \rho + P \geq 0$
null energy condition	$\rho + P \geq 0$

These are arranged roughly in order from strongest to weakest. They all have to do with the idea that negative mass-energy doesn't seem to exist in our universe, i.e., that gravity is always attractive rather than repulsive. With this motivation, it would seem that there should only be one way to state an energy condition: $\rho > 0$. But the symbols ρ and P refer to the form of the stress-energy tensor in a special frame of reference, interpreted as the one that is at rest relative to the average motion of the ambient matter. (Such a frame is not even guaranteed to exist unless the matter acts as a perfect fluid.) In this frame, the tensor is diagonal. Switching to some other frame of reference, the ρ and P parts of the tensor would mix, and it might be possible to end up with a negative energy density. The weak energy condition is the constraint we need in order to make sure that the energy density is never negative in any frame.

The dominant energy condition is like the weak energy condition, but it also guarantees that no observer will see a flux of energy flowing at speeds greater than c.

The strong energy condition essentially states that gravity is never repulsive; it is violated by the cosmological constant (see section 8.2).

Example 7: An electromagnetic wave

In example 1, we saw that dust boosted along the x axis gave a stress-energy tensor

$$T_{\mu\nu} = \gamma^2 \rho \begin{pmatrix} 1 & v \\ v & v^2 \end{pmatrix}, \tag{8.1.18}$$

where we now suppress the y and z parts, which vanish. For v $\rightarrow$ 1, this becomes

$$T_{\mu\nu} = \rho' \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \tag{8.1.19}$$

where ρ' is the energy density as measured in the new frame. As a source of gravitational fields, this ultrarelativistic dust is indistinguishable from any other form of matter with v = 1 along the x axis, so this is also the stress-energy tensor of an electromagnetic wave with local energy-density ρ 0 , propagating along the x axis. (For the full expression for the stress-energy tensor of an arbitrary electromagnetic field, see the Wikipedia article "Electromagnetic stress-energy tensor.")

This is a stress-energy tensor that represents a flux of energy at a speed equal to c, so we expect it to lie at exactly the limit imposed by the dominant energy condition (DEC). Our statement of the DEC, however, was made for a diagonal stress-energy tensor, which is what is seen by an observer at rest relative to the matter. But we know that it's impossible to have an observer who, as the teenage Einstein imagined, rides alongside an electromagnetic wave on a motorcycle. One way to handle this is to generalize our definition of the energy condition. For the DEC, it turns out that this can be done by requiring that the matrix T, when multiplied by a vector on or inside the future light-cone, gives another vector on or inside the cone.

A less elegant but more concrete workaround is as follows. Returning to the original expression for the T of boosted dust at velocity v, we let v = 1 + ϵ, where $|\epsilon| << 1$. This gives a stress-energy tensor that (ignoring multiplicative constants) looks like:

$$\begin{pmatrix} 1 & 1+\epsilon \\ 1+\epsilon & 1+2\epsilon \end{pmatrix}. \tag{8.1.20}$$

If ϵ is negative, we have ultrarelativistic dust, and we can verify that it satisfies the DEC by un-boosting back to the rest frame. To do this explicitly, we can find the matrix's eigenvectors, which (ignoring terms of order ϵ^2) are (1, 1 + ϵ) and (1, 1 − ϵ), with eigenvalues 2 + 2ϵ and 0, respectively. For ϵ < 0, the first of these is timelike, the second spacelike. We interpret them simply as the t and x basis vectors of the rest frame in which we originally described the dust. Using them as a basis, the stress-energy tensor takes on the form diag(2 + 2ϵ, 0). Except for a constant factor that we didn't bother to keep track of, this is the original form of the T in the dust's rest frame, and it clearly satisfies the DEC, since P = 0.

For ϵ > 0, v = 1 + ϵ is a velocity greater than the speed of light, and there is no way to construct a boost corresponding to −v. We can nevertheless find a frame of reference in which the stressenergy tensor is diagonal, allowing us to check the DEC. The expressions found above for the eigenvectors and eigenvalues are still valid, but now the timelike and spacelike characters of the two basis vectors have been interchanged. The stress-energy tensor has the form diag(0, 2 + 2ϵ), with ρ = 0 and P > 0, which violates the DEC. As in this example, any flux of mass-energy at speeds greater than c will violate the DEC.

The DEC is obeyed for ϵ < 0 and violated for ϵ > 0, and since ϵ = 0 gives a stress-energy tensor equal to that of an electromagnetic wave, we can tell that light is exactly on the border between forms of matter that fulfill the DEC and those that

don't. Since the DEC is formulated as a non-strict inequality, it follows that light obeys the DEC.

Example 8: No "speed of flux"

The foregoing discussion may have encouraged the reader to believe that it is possible in general to read off a "speed of energy flux" from the value of T at a point. This is not true.

The difficulty lies in the distinction between flow with and without accumulation, which is sometimes valid and sometimes not. In springtime in the Sierra Nevada, snowmelt adds water to alpine lakes more rapidly than it can flow out, and the water level rises. This is flow with accumulation. In the fall, the reverse happens, and we have flow with depletion (negative accumulation).

Figure 8.1.5 (1) shows a second example in which the distinction seems valid. Charge is flowing through the lightbulb, but because there is no accumulation of charge in the DC circuit, we can't detect the flow by an electrostatic measurement; the wire does not attract the tiny bits of paper below it on the table.

But we know that with different measurements, we could detect the flow of charge in Figure 8.1.5 (1). For example, the magnetic field from the wire would deflect a nearby magnetic compass. This shows that the distinction between flow with and without accumulation may be sometimes valid and sometimes invalid. Flow without accumulation may or may not be detectable; it depends on the physical context.

In Figure 8.1.5 (2), an electric charge and a magnetic dipole are super-imposed at a point. The Poynting vector **P** defined as **E** × **B** is used in electromagnetism as a measure of the flux of energy, and it tells the truth, for example, when the sun warms your sun on a hot day. In Figure 8.1.5 (2), however, all the fields are static. It seems as though there can be no flux of energy. But that doesn't mean that the Poynting vector is lying to us. It tells us that there is a pattern of flow, but it's flow without accumulation; the Poynting vector forms circular loops that close upon themselves, and electromagnetic energy is transported in and out of any volume at the same rate. We would perhaps prefer to have a mathematical rule that gave zero for the flux in this situation, but it's acceptable that our rule **P** = **E** × **B** gives a nonzero result, since it doesn't incorrectly predict an accumulation, which is what would be detectable.
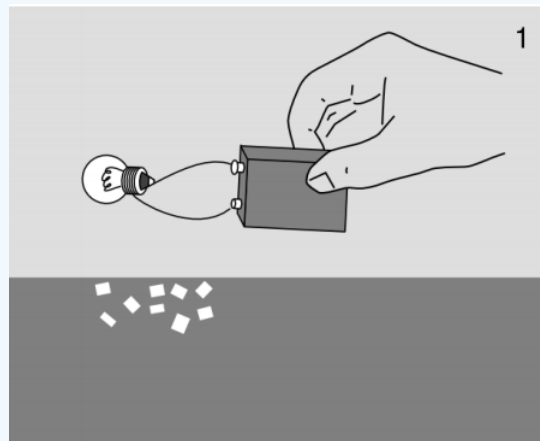
Figure_8.1.5b.png" />

***Figure* 8.2.5**

Now suppose we're presented with this stress-energy tensor, measured at a single point and expressed in some units:

$$T^{\mu\nu} = \begin{pmatrix} 4.037 \pm 0.002 & 4.038 \pm 0.002 \\ 4.036 \pm 0.002 & 4.036 \pm 0.002 \end{pmatrix}. \tag{8.1.21}$$

To within the experimental error bars, it has the right form to be many different things: (1) We could have a universe filled with perfectly uniform dust, moving along the x axis at some ultrarelativistic speed v so great that the ϵ in v = 1 − ϵ, as in example 7, is not detectably different from zero. (2) This could be a point sampled from an electromagnetic wave traveling along the x axis. (3) It could be a point taken from Figure 8.1.5 (2). (In cases 2 and 3, the off-diagonal elements are simply the Poynting vector.)

In cases 1 and 2, we would be inclined to interpret this stress-energy tensor by saying that its off-diagonal part measures the flux of mass-energy along the x axis, while in case 3 we would reject such an interpretation. The trouble here is not so much in our interpretation of T as in our Newtonian expectations about what is or isn't observable about fluxes that flow without accumulation. In Newtonian mechanics, a flow of mass is observable, regardless of whether there is accumulation, because it carries momentum with it; a flow of energy, however, is undetectable if there is no accumulation. The trouble here is that relativistically, we can't maintain this distinction between mass and energy. The Einstein field equations tell us that a flow of either will contribute equally to the stress-energy, and therefore to the surrounding gravitational field.

The flow of energy in Figure 8.1.5 (2) contributes to the gravitational field, and its contribution is changed, for example, if the magnetic field is reversed. The figure is in fact not a bad qualitative representation of the spacetime around a rotating, charged black hole. At large distances, however, the gravitational effect of the off-diagonal terms in T becomes small, because they average to nearly zero over a sufficiently large spherical region. The distant gravitational field approaches that of a point mass with the same mass-energy

Example 9: Momentum in static fields

Continuing the train of thought described in example 8, we can come up with situations that seem even more paradoxical. In Figure 8.1.5 (2), the total momentum of the fields vanishes by symmetry. This symmetry can, however, be broken by displacing the electric charge by $\Delta \mathbf{R}$ perpendicular to the magnetic dipole vector $\mathbf{D}$. The total momentum no longer vanishes, and now lies in the direction of $\mathbf{D} \times \Delta \mathbf{R}$. But we have proved in example 2 that a system's center of mass-energy is at rest if and only if its total momentum is zero. Since this system's center of mass-energy is certainly at rest, where is the other momentum that cancels that of the electric and magnetic fields?

Suppose, for example, that the magnetic dipole consists of a loop of copper wire with a current running around it. If we open a switch and extinguish the dipole, it appears that the system must recoil! This seems impossible, since the fields are static, and an electric charge does not interact with a magnetic dipole.

Babson et al.[7] have analyzed a number of examples of this type. In the present one, the mysterious "other momentum" can be attributed to a relativistic imbalance between the momenta of the electrons in the different parts of the wire. A subtle point about these examples is that even in the case of an idealized dipole of vanishingly small size, it makes a difference what structure we assume for the dipole. In particular, the field's momentum is nonzero for a dipole made from a current loop of infinitesimal size, but zero for a dipole made out of two magnetic monopoles.[8]

[7] Am. J. Phys. 77 (2009) 826

[8] Milton and Meille, arxiv.org/abs/1208.4826

References

[6] "Theory of gravitation theories: a no-progress report," Sotiriou, Faraoni, and Liberati, http://arxiv.org/abs/0707.2748

8.3: Cosmological Solutions (Part I)

We are thus led to pose two interrelated questions. First, what can empirical observations about the universe tell us about the laws of physics, such as the zero or nonzero value of the cosmological constant? Second, what can the laws of physics, combined with observation, tell us about the large-scale structure of the universe, its origin, and its fate?

Evidence for the Finite Age of the Universe

We have a variety of evidence that the universe's existence does not stretch for an unlimited time into the past.

When astronomers view light from the deep sky that has been traveling through space for billions of years, they observe a universe that looks different from today's. For example, quasars were common in the early universe but are uncommon today.

In the present-day universe, stars use up deuterium nuclei, but there are no known processes that could replenish their supply. We therefore expect that the abundance of deuterium in the universe should decrease over time. If the universe had existed for an infinite time, we would expect that all its deuterium would have been lost, and yet we observe that deuterium does exist in stars and in the interstellar medium.

The second law of thermodynamics predicts that any system should approach a state of thermodynamic equilibrium, and yet our universe is very far from thermal equilibrium, as evidenced by the fact that our sun is hotter than interstellar space, or by the existence of functioning heat engines such as your body or an automobile engine.

With hindsight, these observations suggest that we should not look for cosmological models that persist for an infinite time into the past.

Evidence for Expansion of the Universe

We don't only see time-variation in locally observable quantities such as quasar abundance, deuterium abundance, and entropy. In addition, we find empirical evidence for global changes in the universe. By 1929, Edwin Hubble at Mount Wilson had determined that the universe was expanding, and historically this was the first convincing evidence that Einstein's original goal of modeling a static cosmology had been a mistake. Einstein later referred to the cosmological constant as the "greatest blunder of my life," and for the next 70 years it was commonly assumed that Λ was exactly zero.

Since we observe that the universe is expanding, the laws of thermodynamics require that it also be cooling, just as the exploding air-gas mixture in a car engine's cylinder cools as it expands. If the universe is currently expanding and cooling, it is natural to imagine that in the past it might have been very dense and very hot. This is confirmed directly by looking up in the sky and seeing radiation from the hot early universe. In 1964, Penzias and Wilson at Bell Laboratories in New Jersey detected a mysterious background of microwave radiation using a directional horn antenna. As with many accidental discoveries in science, the important thing was to pay attention to the surprising observation rather than giving up and moving on when it confounded attempts to understand it. They pointed the antenna at New York City, but the signal didn't increase. The radiation didn't show a 24-hour periodicity, so it couldn't be from a source in a certain direction in the sky. They even went so far as to sweep out the pigeon droppings inside. It was eventually established that the radiation was coming uniformly from all directions in the sky and had a black-body spectrum with a temperature of about 3 K.
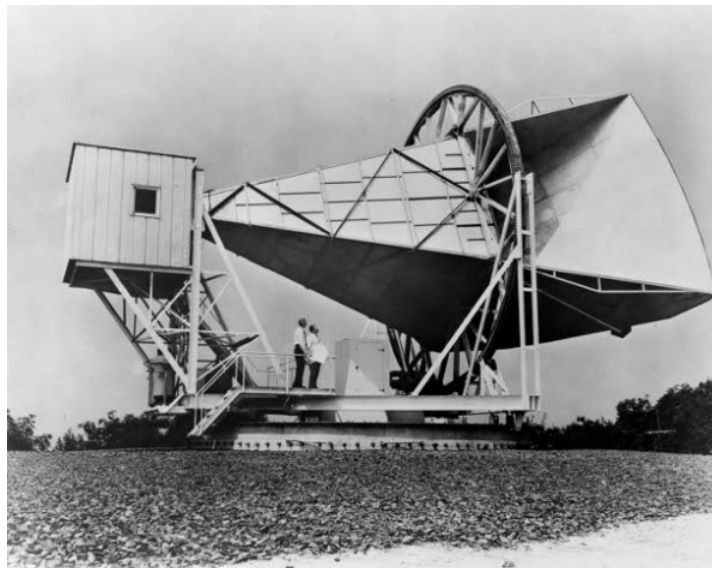
Figure 8.3.1 - The horn antenna used by Penzias and Wilson.

This is now interpreted as follows. At one time, the universe was hot enough to ionize matter. An ionized gas is opaque to light, since the oscillating fields of an electromagnetic wave accelerate the charged particles, depositing kinetic energy into them. Once the universe became cool enough, however, matter became electrically neutral, and the universe became transparent. Light from this time is the most long-traveling light that we can detect now. The latest data show that transparency set in when the temperature was about 3000 K. The surface we see, dating back to this time, is known as the surface of last scattering. Since then, the universe has expanded by about a factor of 1000, causing the wavelengths of photons to be stretched by the same amount due to the expansion of the underlying space. This is equivalent to a Doppler shift due to the source's motion away from us; the two explanations are equivalent. We therefore see the 3000 K optical black-body radiation red-shifted to 3 K, in the microwave region.

It is logically possible to have a universe that is expanding but whose local properties are nevertheless static, as in the steady-state model of Fred Hoyle, in which some novel physical process spontaneously creates new hydrogen atoms, preventing the infinite dilution of matter over the universe's history, which in this model extends infinitely far into the past. But we have already seen strong empirical evidence that the universe's local properties (quasar abundance, etc.) are changing over time. The CMB is an even more extreme and direct example of this; the universe full of hot, dense gas that emitted the CMB is clearly nothing like today's universe. A brief discussion of the steady-state model is given in section 8.4.

Evidence for Homogeneity and Isotropy

These observations demonstrate that the universe is not homogeneous in time, i.e., that one can observe the present conditions of the universe (such as its temperature and density), and infer what epoch of the universe's evolution we inhabit. A different question is the Copernican one of whether the universe is homogeneous in space. Surveys of distant quasars show that the universe has very little structure at scales greater than a few times 10^{25} m. (This can be seen on a remarkable logarithmic map constructed by Gott et al., astro.princeton.edu/universe.) This suggests that we can, to a good approximation, model the universe as being isotropic (the same in all spatial directions) and homogeneous (the same at all locations in space). (Isotropy does not follow from homogeneity. Examples of homogeneous but anisotropic cosmologies include rotating cosmologies and the Kantowsky-Sachs metric, problem 13.)

Further evidence comes from the extreme uniformity of the cosmic microwave background radiation, once one subtracts out the dipole anisotropy due to the Doppler shift arising from our galaxy's motion relative to the CMB. When the CMB was first discovered, there was doubt about whether it was cosmological in origin (rather than, say, being associated with our galaxy), and it was expected that its isotropy would be as large as 10%. As physicists began to be convinced that it really was a relic of the early universe, interest focused on measuring this anisotropy, and a series of measurements put tighter and tighter upper bounds on it. Other than the dipole term, there are two ways in which one might naturally expect anisotropy to occur. There might have been some lumpiness in the early universe, which might have served as seeds for the condensation of galaxy clusters out of the cosmic medium. Furthermore, we might wonder whether the universe as a whole is rotating. The general-relativistic notion of rotation is very different from the Newtonian one, and in particular, it is possible to have a cosmology that is rotating without having any

center of rotation (see problem 5). In fact one of the first exact solutions discovered for the Einstein field equations was the Gödel metric, which described a bizarre rotating universe with closed timelike curves, i.e., one in which causality was violated. In a rotating universe, one expects that radiation received from great cosmological distances will have a transverse Doppler shift, i.e., a shift originating from the time dilation due to the motion of the distant matter across the sky. This shift would be greatest for sources lying in the plane of rotation relative to us, and would vanish for sources lying along the axis of rotation. The CMB would therefore show variation with the form of a quadrupole term, $3\cos^2\theta - 1$. In 1977 a U-2 spyplane (the same type involved in the 1960 U.S.-Soviet incident) was used by Smoot et al.[14] to search for anisotropies in the CMB. This experiment was the first to definitively succeed in detecting the dipole anisotropy. After subtraction of the dipole component, the CMB was found to be uniform at the level of $\sim 3 \times 10^{-4}$. This provided strong support for homogeneous cosmological models, and ruled out rotation of the universe with $\omega \gtrsim 10^{-22}$ Hz.

> **Note**
>
> G. F. Smoot, M. V. Gorenstein, and R. A. Muller, "Detection of Anisotropy in the Cosmic Blackbody Radiation," Phys. Rev. Lett. 39 (1977) 898. The interpretation of the CMB measurements is somewhat model-dependent; in the early years of observational cosmology, it was not even universally accepted that the CMB had a cosmological origin. The best model-independent limit on the rotation of the universe comes from observations of the solar system, Clemence, "Astronomical Time," Rev. Mod. Phys. 29 (1957) 2.

The FRW Cosmologies

The FRW Metric and the Standard Coordinates

Motivated by Hubble's observation that the universe is expanding, we hypothesize the existence of solutions of the field equation in which the properties of space are homogeneous and isotropic, but the over-all scale of space is increasing as described by some scale function a(t). Because of coordinate invariance, the metric can still be written in a variety of forms. One such form is

$$ds^2 = dt^2 - a(t)^2 d\ell^2, \tag{8.3.1}$$

where the spatial part is

$$d\ell^2 = f(r)dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2. \tag{8.3.2}$$

To interpret the coordinates, we note that if an observer is able to determine the functions a and f for her universe, then she can always measure some scalar curvature such as the Ricci scalar or the Kretschmann invariant, and since these are proportional to a raised to some power, she can determine a and t. This shows that t is a "look-out-the-window" time, i.e., a time coordinate that we can determine by looking out the window and observing the present conditions in the universe. Because the quantity being measured directly is a scalar, the result is independent of the observer's state of motion. (In practice, these scalar curvatures are difficult to measure directly, so we measure something else, like the sky-wide average temperature of the cosmic microwave background.) Simultaneity is supposed to be ill-defined in relativity, but the look-out-the-window time defines a notion of simultaneity that is the most naturally interesting one in this spacetime. With this particular definition of simultaneity, we can also define a preferred state of rest at any location in spacetime, which is the one in which t changes as slowly as possible relative to one's own clock. This local rest frame, which is more easily determined in practice as the one in which the microwave background is most uniform across the sky, can also be interpreted as the one that is moving along with the Hubble flow, i.e., the average motion of the galaxies, photons, or whatever else inhabits the spacetime. The time t is interpreted as the proper time of a particle that has always been locally at rest. The spatial distance measured by L = $\int$ a dℓ is called the proper distance. It is the distance that would be measured by a chain of rulers, each of them "at rest" in the above sense.

These coordinates are referred as the "standard" cosmological coordinates; one will also encounter other choices, such as the comoving and conformal coordinates, which are more convenient for certain purposes. Historically, the solution for the functions a and f was found by de Sitter in 1917.

The spatial metric

The unknown function f(r) has to give a 3-space metric $d\ell^2$ with a constant Einstein curvature tensor. The following Maxima program computes the curvature.

```
1    load(ctensor);
2    dim:3;
3    ct_coords:[r,theta,phi];
4    depends(f,t);
5    lg:matrix([f,0,0],
6              [0,r^2,0],
7              [0,0,r^2*sin(theta)^2]);
8    cmetric();
9    einstein(true);
```

Line 2 tells Maxima that we're working in a space with three dimensions rather than its default of four. Line 4 tells it that f is a function of time. Line 9 uses its built-in function for computing the Einstein tensor G^a_b. The result has only one nonvanishing component, $G^t_t = \frac{1-\frac{1}{f}}{r^2}$. This has to be constant, and since scaling can be absorbed in the factor a(t) in the 3+1-dimensional metric, we can just set the value of G_{tt} more or less arbitrarily, except for its sign. The result is f = $\frac{1}{1-kr^2}$, where k = −1, 0, or 1.

The resulting metric, called the Robertson-Walker metric, is

$$ds^2 = dt^2 - a^2 \left(\frac{dr^2}{1 - kr^2} + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \right).$$ (8.3.3)

The form of $d\ell^2$ shows us that k can be interpreted in terms of the sign of the spatial curvature. We recognize the k = 0 metric as a flat spacetime described in spherical coordinates. To interpret the k ≠ 0 cases, we note that a circle at coordinate r has proper circumference C = 2πar and proper radius R = $a \int_0^r \sqrt{f(r')}$ dr'. For k < 0, we have f < 1 and C > 2πR, indicating negative spatial curvature. For k > 0 there is positive curvature.

Let's examine the positive-curvature case more closely. Suppose we select a particular plane of simultaneity defined by t = constant and $\phi = \frac{\pi}{2}$, and we start doing geometry in this plane. In two spatial dimensions, the Riemann tensor only has a single independent component, which can be identified with the Gaussian curvature (sec. 5.4), and when this Gaussian curvature is positive and constant, it can be interpreted as the angular defect of a triangle per unit area (sec. 5.3). Since the sum of the interior angles of a triangle can never be greater than 3π, we have an upper limit on the area of any triangle. This happens because the positive-curvature Robertson-Walker metric represents a cosmology that is spatially finite. At a given t, it is the three-dimensional analogue of a two-sphere. On a two-sphere, if we set up polar coordinates with a given point arbitrarily chosen as the origin, then we know that the r coordinate must "wrap around" when we get to the antipodes. That is, there is a coordinate singularity there. (We know it can only be a coordinate singularity, because if it wasn't, then the antipodes would have special physical characteristics, but the FRW model was constructed to be spatially homogeneous.) This "wrap-around" behavior is described by saying that the model is *closed*.
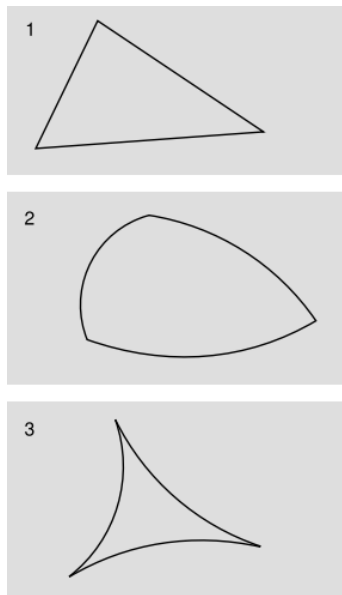
Figure 8.3.2 - 1. In the Euclidean plane, this triangle can be scaled by any factor while remaining similar to itself. 2. In a plane with positive curvature, geometrical figures have a maximum area and maximum linear dimensions. This triangle has almost the maximum area, because the sum of its angles is nearly 3π. 3. In a plane with negative curvature, figures have a maximum area but no maximum linear dimensions. This triangle has almost the maximum area, because the sum of its angles is nearly zero. Its vertices, however, can still be separated from one another without limit.

In the negative-curvature case, there is no limit on distances, Figure 8.2.2 (3). Such a universe is called *open*. In the case of an open universe, it is particularly easy to demonstrate a fact that bothers many students, which is that proper distances can grow at rates exceeding c. Let particles A and B both be at rest relative to the Hubble flow. The proper distance between them is then given by L $= a\ell$, where $\ell = \int_A^B d\ell$ is constant. Then differentiating L with respect to the look-out-the-window time t gives $\frac{dL}{dt} = \dot{a}\ell$. In an open universe, there is no limit on the size of ℓ, so at any given time, we can make $\frac{dL}{dt}$ as large as we like. This does not violate special relativity, since it is only locally that special relativity is a valid approximation to general relativity. Because GR only supplies us with frames of reference that are local, the velocity of two objects relative to one another is not even uniquely defined; our choice of $\frac{dL}{dt}$ was just one of infinitely many possible definitions.

The distinction between closed and open universes is not just a matter of geometry, it's a matter of topology as well. Just as a two-sphere cannot be made into a Euclidean plane without cutting or tearing, a closed universe is not topologically equivalent to an open one. The correlation between local properties (curvature) and global ones (topology) is a general theme in differential geometry. A universe that is open is open forever, and similarly for a closed one.

The Friedmann Equations

Having fixed f(r), we can now see what the field equation tells us about a(t). The next program computes the Einstein tensor for the full four-dimensional spacetime:

```
1    load(ctensor);
2    ct_coords:[t,r,theta,phi];
3    depends(a,t);
4    lg:matrix([1,0,0,0],
5              [0,-a^2/(1-k*r^2),0,0],
6              [0,0,-a^2*r^2,0],
7              [0,0,0,-a^2*r^2*sin(theta)^2]);
8    cmetric();
9    einstein(true);
```

The result is

$$G^t_t = 3\left(\frac{\dot{a}}{a}\right)^2 + 3ka^{-2}$$

$$G^r_r = G^\theta_\theta = G^\phi_\phi = 2\frac{\ddot{a}}{a} + \left(\frac{\dot{a}}{a}\right)^2 + ka^{-2},$$

where dots indicate differentiation with respect to time.

Since we have G^a_b with mixed upper and lower indices, we either have to convert it into $G_{ab,}$ or write out the field equations in this mixed form. The latter turns out to be simpler. In terms of mixed indices, g^a_b is always simply diag(1, 1, 1, 1). Arbitrarily singling out r = 0 for simplicity, we have g = diag(1, −a², 0, 0). The stress-energy tensor is T^μ_ν = diag(ρ, −P, −P, −P). (See example 4 for the signs.) Substituting into $G^a_b = 8\pi T^a_b + \Lambda g^a_b$, we find

$$3\left(\frac{\dot{a}}{a}\right)^2 + 3ka^{-2} - \Lambda = 8\pi\rho$$

$$2\frac{\ddot{a}}{a} + \left(\frac{\dot{a}}{a}\right)^2 + ka^{-2} - \Lambda = -8\pi P.$$

Rearranging a little, we have a set of differential equations known as the Friedmann equations,

$$\frac{\ddot{a}}{a} = \frac{1}{3}\Lambda - \frac{4\pi}{3}\Lambda - \frac{4\pi}{3}(\rho + 3P)$$

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{1}{3}\Lambda + \frac{8\pi}{3}\rho - ka^{-2}.$$

The cosmology that results from a solution of these differential equations is known as the Friedmann-Robertson-Walker (FRW) or Friedmann-Lemaître-Robertson-Walker (FLRW) cosmology.
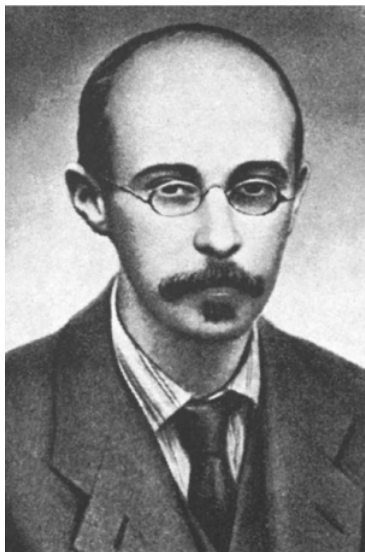
Figure 8.3.3 - Alexander Friedmann (1888- 1925)

The first Friedmann equation describes the rate at which cosmological expansion accelerates or decelerates. Let's refer to it as the acceleration equation. It expresses the basic idea of the field equations, which is that non-tidal curvature (left-hand side) is caused by the matter that is present locally (right-hand side). Example 15 illustrates this in a simple case.

The second Friedmann equation tells us the magnitude of the rate of expansion or contraction. Call it the velocity equation. The quantity $\frac{\dot{a}}{a}$, evaluated at the present cosmological time, is the Hubble constant H_o (which is constant only in the sense that at a fixed time, it is a constant of proportionality between distance and recession velocity).

To the practiced eye, it seems odd to have two dynamical laws, one predicting velocity and one acceleration. The analogous laws in freshman mechanics would be Newton's second law, which predicts acceleration, and conservation of energy, which predicts velocity. Newton's laws and conservation of energy are not independent, and for mechanical systems either can be derived from the other. The Friedmann equations, however, are not overdetermined or redundant. They are underdetermined, because we want to

predict three unknown functions of time: a, ρ, and P. Since there are only two equations, they are not sufficient to uniquely determine a solution for all three functions. The third constraint comes in the form of some type of equation of state for the matter described by ρ and P, which in simple models can often be written in the form P = wρ. For example, dust has w = 0.

Unlike a, ρ, and P, the cosmological constant Λ is not free to vary with time; if it did, then the stress-energy tensor would have a nonvanishing divergence, which is not consistent with the Einstein field equations (see section 8.1).

Although general relativity does not provide any scalar, globally conserved measure of mass-energy that is conserved in all spacetimes, the Friedmann velocity equation can be loosely interpreted as a statement of conservation of mass-energy in an FRW spacetime. The left-hand side acts like kinetic energy. In a cosmology that expands and then recontracts in a Big Crunch, the turnaround point is defined by the time at which the right-hand side equals zero. The origin of the velocity equation is in fact the time-time part of the field equations, whose source term is the mass-energy component of the stress-energy tensor.

Example 15: Scooping out a hole
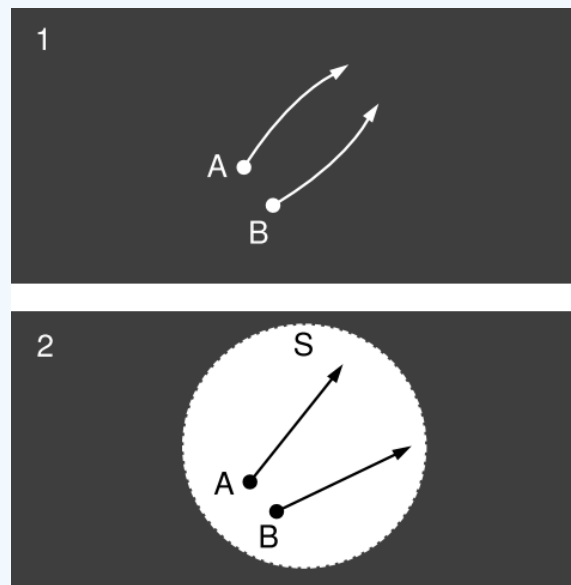
Figure 8.3.4

This example illustrates the connection between cosmological acceleration and local density of matter given by the Friedmann acceleration equation. Consider two cosmologies, each with $\Lambda = 0$. Cosmology 1 is an FRW spacetime in which all matter is in the form of nonrelativistic particles such as atoms or galaxies. 2 is identical to 1, except that all the matter has been scooped out of a small spherical region S, leaving a vacuum. ("Small" means small compared to the Hubble scale $\frac{1}{H_o}$.) Within S, we introduce test particles A and B. Because an FRW spacetime is homogeneous and isotropic, cosmology 2 retains spherical symmetry about the center of S. Since $\Lambda = 0$, Birkhoff's theorem applies to 2, so 2 is flat inside S. Therefore in 2, the relative acceleration **a** of the test particles equals zero.

Because S is small compared to cosmological distances, and because the dust is nonrelativistic, local observers can accurately attibute the difference in behavior between 1 and 2 to the Newtonian gravitational force from the dust that was present in 1 but not in 2. For convenience, let A and B both be initially at rest relative to the local dust (i.e., having $\dot{\theta} = \dot{\phi} = 0$). By the definition of the scale factor (i.e., by inspection of the FRW metric), the distance between them varies as const × a(t). If one of these particles is an observer, she sees a "force" acting on the other particle that causes an acceleration (ä /a)r, where r is the displacement between the particles.

Since **a** = 0 in 2, it follows that the acceleration in 1 can be calculated accurately by finding the Newtonian gravitational force due to the added dust. This results in a connection between $\frac{\ddot{a}}{a}$, on the left-hand side of the Friedmann acceleration equation, and ρ, on the right side.

For consistency, we can verify that the Newtonian gravitational force exerted by a uniform sphere, at a point on its interior, is proportional to **r**. This is a classic result that is easily derived from Newton's shell theorem.

8.4: Cosmological Solutions (Part 2)

A Singularity at the Big Bang

The Friedmann equations only allow a constant a in the case where Λ is perfectly tuned relative to the other parameters, and even this artificially fine-tuned equilibrium turns out to be unstable. These considerations make a static cosmology implausible on theoretical grounds, and they are also consistent with the observed Hubble expansion (section 8.2).

Since the universe is not static, what happens if we use general relativity to extrapolate farther and farther back in time?

If we extrapolate the Friedmann equations backward in time, we find that they always have a = 0 at some point in the past, and this occurs regardless of the details of what we assume about the matter and radiation that fills the universe. To see this, note that, as discussed in example 14, radiation is expected to dominate the early universe, for generic reasons that are not sensitive to the (substantial) observational uncertainties about the universe's present-day mixture of ingredients. Under radiation-dominated conditions, we can approximate Λ = 0 and P = $\frac{\rho}{3}$ (example 14) in the first Friedmann equation, finding

$$\frac{\ddot{a}}{a} = -\frac{8\pi}{3}\rho \tag{8.2.4}$$

where ρ is the density of mass-energy due to radiation. Since $\frac{\ddot{a}}{a}$ is always negative, the graph of a(t) is always concave down, and since a is currently increasing, there must be some time in the past when a = 0. One can readily verify that this is not just a coordinate singularity; the Ricci scalar curvature $R^a{}_a$ diverges, and the singularity occurs at a finite proper time in the past.

In section 6.3, we saw that a black hole contains a singularity, but it appears that such singularities are always hidden behind event horizons, so that we can never observe them from the outside. The FRW singularity, however, is not hidden behind an event horizon. It lies in our past light-cone, and our own world-lines emerged from it. The universe, it seems, originated in a Big Bang, a concept that originated with the Belgian Roman Catholic priest Georges Lemaître.

Figure 8.4.5 - Georges Lemaître (1894- 1966) proposed in 1927 that our universe be modeled in general relativity as a spacetime in which space expanded over time. Lemaître's ideas were initially treated skeptically by Eddington and Einstein, who told him, "Your calculations are correct, but your physics is abominable." Later, as Hubble's observational evidence for cosmological expansion became widely accepted, both Einstein and Eddington became converts, helping to bring Lemaître's ideas to the attention of the community. In 1931, an emboldened Lemaître described the idea that the universe began from a "Primeval Atom" or "Cosmic Egg." The name that eventually stuck was "Big Bang," coined by Fred Hoyle as a derisive term.

> **exercise 8.4.1:**
>
> Self-check: Why is it not correct to think of the Big Bang as an explosion that occurred at a specific point in space?

Does the FRW singularity represent something real about our universe?

One thing to worry about is the accuracy of our physical modeling of the radiation-dominated universe. The presence of an initial singularity in the FRW solutions does not depend sensitively on on assumptions like $P = \frac{\rho}{3}$, but it is still disquieting that no laboratory experiment has ever come close to attaining the conditions under which we could test whether a gas of photons produces gravitational fields as predicted by general relativity. We saw in section 8.1 that static electric fields do produce gravitational fields as predicted, but this is not the same as an empirical confirmation that electromagnetic waves also act as gravitational sources in exactly the manner that general relativity claims. We do, however, have a consistency check in the form of the abundances of nuclei. Calculations of nuclear reactions in the early, radiation-dominated universe predict certain abundances of hydrogen, helium, and deuterium. In particular, the relative abundance of helium and deuterium is a sensitive test of the relationships among a, $\dot{a}$, and $\ddot{a}$ predicted by the FRW equations, and they confirm these relationships to a precision of about $5 \pm 4\%$.[15]

> **Note**
>
> Steigman, Ann. Rev. Nucl. Part. Sci. 57 (2007) 463. These tests are stated in terms of the Hubble "constant" $H = \frac{\dot{a}}{a}$, which is actually varying over cosmological time-scales. The nuclear helium-deuterium ratio is sensitive to $\frac{\dot{H}}{H}$.

An additional concern is whether the Big Bang singularity is just a product of the unrealistic assumption of perfect symmetry that went into the FRW cosmology. One of the Penrose-Hawking singularity theorems proves that it is not.[16] This particular singularity theorem requires three conditions: (1) the strong energy condition holds; (2) there are no closed timelike curves; and (3) a trapped surface exists in the past timelike geodesics originating at some point. The requirement of a trapped surface can fail if the universe is inhomogeneous to $\gtrsim 10^{-4}$, but observations of the cosmic microwave background rule out any inhomogeneity this large (see section 8.2). The other possible failure of the assumptions is that if the cosmological constant is large enough, it violates the strong energy equation, and we can have a Big Bounce rather than a Big Bang (see section 8.2).

An Exceptional Case: the Milne Universe

There is still a third loophole in our conclusion that the Big Bang singularity must have existed. Consider the special case of the FRW analysis, found by Milne in 1932 (long before FRW), in which the universe is completely empty, with $\rho = 0$ and $\Lambda = 0$. This is of course not consistent with the fact that the universe contains stars and galaxies, but we might wonder whether it could tell us anything interesting as a simplified approximation to a very dilute universe. The result is that the scale factor a varies linearly with time (problem 3). If a is not constant, then there exists a time at which a = 0, but this doesn't turn out to be a real singularity (which isn't surprising, since there is no matter to create gravitational fields). Let this universe have a scattering of test particles whose masses are too small to invalidate the approximation of $\rho = 0$, and let the test particles be at rest in the (r, θ, ϕ) coordinates. The linear dependence of a on t means that these particles simply move inertially and without any gravitational interactions, spreading apart from one another at a constant rate like the raisins in a rising loaf of raisin bread. The Friedmann equations require k = −1, so the spatial geometry is one of constant negative curvature.

The Milne universe is in fact flat spacetime described in tricky coordinates. The connection can be made as follows. Let a spherically symmetric cloud of test particles be emitted by an explosion that occurs at some arbitrarily chosen event in flat spacetime. Make the cloud's density be nonuniform in a certain specific way, so that every observer moving along with a test particle (called a comoving observer) sees the same local conditions in his own frame; due to Lorentz contraction by a factor γ, this requires that the density be proportional to γ as described by the observer O who remained at the origin. This scenario turns out to be identical to the Milne universe under the change of coordinates from spatially flat coordinates (T, R) to FRW coordinates (t, r), where $t = \frac{T}{\gamma}$ is the proper time and r = $v\gamma$. (Cf. problem 12)

The Milne universe may be useful as an innoculation against the common misconception that the Big Bang was an explosion of matter spreading out into a preexisting vacuum. Such a description seems obviously incompatible with homogeneity, since, for example, an observer at the edge of the cloud sees the cloud filling only half of the sky. But isn't this a logical contradiction, since the Milne universe *does* have an explosion into vacuum, and yet it was derived as a special case of the FRW analysis, which explicitly assumed homogeneity? It is not a contradiction, because a comoving observer never actually sees an edge. In the limit as we approach the edge, the density of the cloud (as seen by the observer who stayed at the origin) approaches infinity, and the

Lorentz contraction also approaches infinity, so that O considers them to be like Hamlet saying, "I could be bounded in a nutshell, and count myself a king of infinite space." This logic only works in the case of the Milne universe. The explosion-into-preexisting-vacuum interpretation fails in Big Bang cosmologies with $\rho \neq 0$.

Observability of Expansion

Brooklyn is Not Expanding!

The proper interpretation of the expansion of the universe, as described by the Friedmann equations, can be tricky. The example of the Milne universe encourages us to imagine that the expansion would be undetectable, since the Milne universe can be described as either expanding or not expanding, depending on the choice of coordinates. A more general consequence of coordinate-independence is that relativity does not pick out any preferred distance scale. That is, if all our meter-sticks expand, and the rest of the universe expands as well, we would have no way to detect the expansion. The flaw in this reasoning is that the Friedmann equations only describe the average behavior of spacetime. As dramatized in the classic Woody Allen movie "Annie Hall:" "Well, the universe is everything, and if it's expanding, someday it will break apart and that would be the end of everything!" "What has the universe got to do with it? You're here in Brooklyn! Brooklyn is not expanding!"

To organize our thoughts, let's consider the following hypotheses:

1. The distance between one galaxy and another increases at the rate given by a(t) (assuming the galaxies are sufficiently distant from one another that they are not gravitationally bound within the same galactic cluster, supercluster, etc.).
2. The wavelength of a photon increases according to a(t) as it travels cosmological distances.
3. The size of the solar system increases at this rate as well (i.e., gravitationally bound systems get bigger, including the earth and the Milky Way).
4. The size of Brooklyn increases at this rate (i.e., electromagnetically bound systems get bigger). 5. The size of a helium nucleus increases at this rate (i.e., systems bound by the strong nuclear force get bigger).

We can imagine that:

- All the above hypotheses are true.
- All the above hypotheses are false, and in fact none of these sizes increases at all.
- Some are true and some false

If all five hypotheses were true, the expansion would be undetectable, because all available meter-sticks would be expanding together. Likewise if no sizes were increasing, there would be nothing to detect. These two possibilities are really the same cosmology, described in two different coordinate systems. But the Ricci and Einstein tensors were carefully constructed so as to be intrinsic. The fact that the expansion affects the Einstein tensor shows that it cannot interpreted as a mere coordinate expansion. Specifically, suppose someone tells you that the FRW metric can be made into a flat metric by a change of coordinates. (I have come across this claim on internet forums.) The linear structure of the tensor transformation equations guarantees that a nonzero tensor can never be made into a zero tensor by a change of coordinates. Since the Einstein tensor is nonzero for an FRW metric, and zero for a flat metric, the claim is false.

> **exercise** 8.4.2
>
> Self-check: The reasoning above implicitly assumed a non-empty universe. Convince yourself that it fails in the special case of the Milne universe.

We can now see some of the limitations of a common metaphor used to explain cosmic expansion, in which the universe is visualized as the surface of an expanding balloon. The metaphor correctly gets across several ideas: that the Big Bang is not an explosion that occurred at a preexisting point in empty space; that hypothesis 1 above holds; and that the rate of recession of one galaxy relative to another is proportional to the distance between them. Nevertheless the metaphor may be misleading, because if we take a laundry marker and draw any structure on the balloon, that structure will expand at the same rate. But this implies that hypotheses 1-5 all hold, which cannot be true.

Since some of the five hypotheses must be true and some false, and we would like to sort out which are which. It should also be clear by now that these are not five independent hypotheses. For example, we can test empirically whether the ratio of Brooklyn's size to the distances between galaxies changes like a(t), remains constant, or changes with some other time dependence, but it is only the ratio that is actually observable.

Empirically, we find that hypotheses 1 and 2 are true (i.e., the photon's wavelength maintains a constant ratio with the intergalactic distance scale), while 3, 4, and 5 are false. For example, the orbits of the planets in our solar system have been measured extremely accurately by radar reflection and by signal propagation times to space probes, and no expanding trend is detected.

General-relativistic Predictions

Does general relativity correctly reproduce these observations? General relativity is mainly a theory of gravity, so it should be well within its domain to explain why the solar system does not expand detectably while intergalactic distances do. It is impractical to solve the Einstein field equations exactly so as to describe the internal structure of all the bodies that occupy the universe: galaxies, superclusters, etc. We can, however, handle simple cases, as in example 20, where we display an exact solution for the case of a universe containing only two things: an isolated black hole, and an energy density described by a cosmological constant. We find that the characteristic scale of the black hole, i.e., the radius of its event horizon, does not increase with time. A fuller treatment of these issues is given later, after some facts about realistic cosmologies have been established. The result is that although bound systems like the solar system are in some cases predicted to expand, the expansion is absurdly small, too small to measure, and much smaller than the rate of expansion of the universe in general as represented by the scale factor a(t). This agrees with observation.

It is easy to show that atoms and nuclei do not *steadily* expand over time. because such an expansion would violate either the equivalence principle or the basic properties of quantum mechanics. One way of stating the equivalence principle is that the local geometry of spacetime is always approximately Lorentzian, so that the laws of physics do not depend on one's position or state of motion. Among these laws of physics are the principles of quantum mechanics, which imply that an atom or a nucleus has a well-defined ground state, with a certain size that depends only on fundamental constants such as Planck's constant and the masses of the particles involved. Atoms and nuclei do experience deformation due to gravitational strains (examples 24-25), but these deformations do not increase with time, and would only be detectable if cosmological expansion were to accelerate radically (example 26).

This is different from the case of a photon traveling across the universe. The argument given above fails, because the photon does not have a ground state. The photon *does* expand, and this is required by the correspondence principle. If the photon did not expand, then its wavelength would remain constant, and this would be inconsistent with the classical theory of electromagnetism, which predicts a Doppler shift due to the relative motion of the source and the observer. One can choose to describe cosmological redshifts either as Doppler shifts or as expansions of wavelength due to cosmological expansion.

A nice way of discussing atoms, nuclei, photons, and solar systems all on the same footing is to note that in geometrized units, the units of mass and length are the same. Therefore the existence of any fundamental massive particle sets a universal length scale, one that will be known to any intelligent species anywhere in the universe. Since photons are massless, they can't be used to set a universal scale in this way; a photon has a certain mass-energy, but that mass-energy can take on any value. Similarly, a solar system sets a length scale, but not a universal one; the radius of a planet's orbit can take on any value. A universe without massive fundamental particles would be a universe without length measurement. It would obey the laws of conformal geometry, in which angles and light-cones were the only measures. This is the reason that atoms and nuclei, which are made of massive fundamental particles, do not expand.

More than One Dimension Required

Another good way of understanding why a photon expands, while an atom does not, is to recall that a one-dimensional space can never have any intrinsic curvature. If the expansion of atoms were to be detectable, we would need to detect it by comparing against some other meter-stick. Let's suppose that a hydrogen atom expands more, while a more tightly bound uranium atom expands less, so that over time, we can detect a change in the ratio of the two atoms' sizes. The world-lines of the two atoms are one-dimensional curves in spacetime. They are housed in a laboratory, and although the laboratory does have some spatial extent, the equivalence principle guarantees that to a good approximation, this small spatial extent doesn't matter. This implies an intrinsic curvature in a one-dimensional space, which is mathematically impossible, so we have a proof by contradiction that atoms do not expand steadily.

Now why does this one-dimensionality argument fail for photons and galaxies? For a pair of galaxies, it fails because the galaxies are not sufficiently close together to allow them both to be covered by a single Lorentz frame, and therefore the set of world-lines comprising the observation cannot be approximated well as lying within a one-dimensional space. Similar reasoning applies for cosmological redshifts of photons received from distant galaxies. One could instead propose flying along in a spaceship next to an electromagnetic wave, and monitoring the change in its wavelength while it is in flight. All the world-lines involved in such an

experiment would indeed be confined to a one-dimensional space. The experiment is impossible, however, because the measuring apparatus cannot be accelerated to the speed of light. In reality, the speed of the light wave relative to the measuring apparatus will always equal c, so the two world-lines involved in the experiment will diverge, and will not be confined to a one-dimensional region of spacetime.

> ### Example 16: A cosmic girdle
>
> Since cosmic expansion has no significant effect on Brooklyn, nuclei, and solar systems, we might be tempted to infer that its effect on any solid body would also be negligible. To see that this is not true, imagine that we live in a closed universe, and the universe has a leather belt wrapping around it on a closed spacelike geodesic. All parts of the belt are initially at rest relative to the local galaxies, and the tension is initially zero everywhere. The belt must stretch and eventually break: for if not, then it could not remain everywhere at rest with respect to the local galaxies, and this would violate the symmetry of the initial conditions, since there would be no way to pick the direction in which a certain part of the belt should begin accelerating.

> ### Example 17: Østvang's quasi-metric relativity
>
> Over the years, a variety of theories of gravity have been proposed as alternatives to general relativity. Some of these, such as the Brans-Dicke theory, remain viable, i.e., they are consistent with all the available experimental data that have been used to test general relativity. One of the most important reasons for trying to construct such theories is that it can be impossible to interpret tests of general relativity's predictions unless one also possesses a theory that predicts something different. This issue, for example, has made it impossible to test Einstein's century-old prediction that gravitational effects propagate at c, since there is no viable theory available that predicts any other speed for them (see section 9.1).
>
> Østvang (arxiv.org/abs/gr-qc/0112025v6) has proposed an alternative theory of gravity, called quasi-metric relativity, which, unlike general relativity, predicts a significant cosmological expansion of the solar system, and which is claimed to be able to explain the observation of small, unexplained accelerations of the Pioneer space probes that remain after all accelerations due to known effects have been subtracted (the "Pioneer anomaly"). We've seen above that there are a variety of arguments against such an expansion of the solar system, and that many of these arguments do not require detailed technical calculations but only knowledge of certain fundamental principles, such as the structure of differential geometry (no intrinsic curvature in one dimension), the equivalence principle, and the existence of ground states in quantum mechanics. We therefore expect that Østvang's theory, if it is logically self-consistent, will probably violate these assumptions, but that the violations must be relatively small if the theory is claimed to be consistent with existing observations. This is in fact the case. The theory violates the strictest form of the equivalence principle.
>
> Over the years, a variety of explanations have been proposed for the Pioneer anomaly, including both glamorous ones (a modification of the $\frac{1}{r^2}$ law of gravitational forces) and others more pedestrian (effects due to outgassing of fuel, radiation pressure from sunlight, or infrared radiation originating from the spacecrafts radioisotope thermoelectric generator). Calculations by Iorio[17] in 2006-2009 show that if the force law for gravity is modified in order to explain the Pioneer anomalies, and if gravity obeys the equivalence principle, then the results are inconsistent with the observed orbital motion of the satellites of Neptune. This makes gravitational explanations unlikely, but does not obviously rule out Østvang's theory, since the theory is not supposed to obey the equivalence principle. Østvang says[18] that his theory predicts an expansion of $\sim$ 1m/yr in the orbit of Triton's moon Nereid, which is consistent with observation. In December 2010, the original discoverers of the effect made a statement in the popular press that they had a new analysis, which they were preparing to publish in a scientific paper, in which the size of the anomaly would be drastically revised downward, with a far greater proportion of the acceleration being accounted for by thermal effects. In my opinion this revision, combined with the putative effect's violation of the equivalence principle, make it clear that the anomaly is not gravitational.
>
> ---
>
> [17] http://arxiv.org/abs/0912.2947v1
>
> [18] private communication, Jan. 4, 2010

Does Space Expand?

Finally, the balloon metaphor encourages us to interpret cosmological expansion as a phenomenon in which space itself expands, or perhaps one in which new space is produced. Does space really expand? Without posing the question in terms of more rigorously defined, empirically observable quantities, we can't say yes or no. It is merely a matter of which definitions one chooses and which

conceptual framework one finds easier and more natural to work within. Bunn and Hogg have stated the minority view against expansion of space[19], while the opposite opinion is given by Francis et al.[20]

As an example of a self-consistent set of definitions that lead to the conclusion that space does expand, Francis et al. give the following. Define eight observers positioned at the corners of a cube, at cosmological distances from one another. Let each observer be at rest relative to the local matter and radiation that were used as ingredients in the FRW cosmology. (For example, we know that our own solar system is *not* at rest in this sense, because we observe that the cosmic microwave background radiation is slightly Doppler shifted in our frame of reference.) Then these eight observers will observe that, over time, the volume of the cube grows as expected according to the cube of the function a(t) in the FRW model.

This establishes that expansion of space is a plausible interpretation. To see that it is not the only possible interpretation, consider the following example. A photon is observed after having traveled to earth from a distant galaxy G, and is found to be red-shifted. Alice, who likes expansion, will explain this by saying that while the photon was in flight, the space it occupied expanded, lengthening its wavelength. Betty, who dislikes expansion, wants to interpret it as a kinematic red shift, arising from the motion of galaxy G relative to the Milky Way Malaxy, M. If Alice and Betty's disagreement is to be decided as a matter of absolute truth, then we need some objective method for resolving an observed redshift into two terms, one kinematic and one gravitational. But we've seen in section 7.4 that this is only possible for a stationary spacetime, and cosmological spacetimes are not stationary: regardless of an observer's state of motion, he sees a change over time in observables such as density of matter and curvature of spacetime. As an extreme example, suppose that Betty, in galaxy M, receives a photon without realizing that she lives in a closed universe, and the photon has made a circuit of the cosmos, having been emitted from her own galaxy in the distant past. If she insists on interpreting this as a kinematic red shift, the she must conclude that her galaxy M is moving at some extremely high velocity relative to itself. This is in fact not an impossible interpretation, if we say that M's high velocity is relative to itself *in the past*. An observer who sets up a frame of reference with its origin fixed at galaxy G will happily confirm that M has been accelerating over the eons. What this demonstrates is that we can split up a cosmological red shift into kinematic and gravitational parts in any way we like, depending on our choice of coordinate system (see also section 7.5).

> ## Example 18: A cosmic whip
>
> The cosmic girdle of example 16 does not transmit any information from one part of the universe to another, for its state is the same everywhere by symmetry, and therefore an observer near one part of the belt gets no information that is any different from what would be available to an observer anywhere else.
>
> Now suppose that the universe is open rather than closed, but we have a rope that, just like the belt, stretches out over cosmic distances along a spacelike geodesic. If the rope is initially at rest with respect to a particular galaxy G (or, more strictly speaking, with respect to the locally averaged cosmic medium), then by symmetry the rope will always remain at rest with respect to G, since there is no way for the laws of physics to pick a direction in which it should accelerate. Now the residents of G cut the rope, release half of it, and tie the other half securely to one of G's spiral arms using a square knot. If they do this smoothly, without varying the rope's tension, then no vibrations will propagate, and everything will be as it was before on that half of the rope. (We assume that G is so massive relative to the rope that the rope does not cause it to accelerate significantly.)
>
> Can observers at distant points observe the tail of the rope whipping by at a certain speed, and thereby infer the velocity of G relative to them? This would produce all kinds of strange conclusions. For one thing, the Hubble law says that this velocity is directly proportional to the length of the rope, so by making the rope long enough we could make this velocity exceed the speed of light. We've also convinced ourselves that the relative velocity of cosmologically distant objects is not even well defined in general relativity, so it clearly can't make sense to interpret the rope-end's velocity in that way.
>
> The way out of the paradox is to recognize that disturbances can only propagate along the rope at a certain speed v. Let's say that the information is transmitted in the form of longitudinal vibrations, in which case it propagates at the speed of sound. For a rope made out of any known material, this is far less than the speed of light, and we've also seen in example 14 and in problem 4 that relativity places fundamental limits on the properties of all possible materials, guaranteeing $v < c$. We can now see that all we've accomplished with the rope is to recapitulate using slower sound waves the discussion that was carried out earlier using light waves. The sound waves may perhaps preserve some information about the state of motion of galaxy G long ago, but all the same ambiguities apply to its interpretation as in the case of light waves — and in addition, we suspect that the rope has long since parted somewhere along its length.

References

[16] Hawking and Ellis, "The Cosmic Black-Body Radiation and the Existence of Singularities in Our Universe," Astrophysical Journal, 152 (1968) 25. Available online at articles.adsabs.harvard.edu.

[19] http://arxiv.org/abs/0808.1081v2

[20] http://arxiv.org/abs/0707.0380v1

8.5: Cosmological Solutions (Part 3)

The Vacuum-dominated Solution

For 70 years after Hubble's discovery of cosmological expansion, the standard picture was one in which the universe expanded, but the expansion must be decelerating. The deceleration is predicted by the special cases of the FRW cosmology that were believed to be applicable, and even if we didn't know anything about general relativity, it would be reasonable to expect a deceleration due to the mutual Newtonian gravitational attraction of all the mass in the universe.

But observations of distant supernovae starting around 1998 introduced a further twist in the plot. In a binary star system consisting of a white dwarf and a non-degenerate star, as the nondegenerate star evolves into a red giant, its size increases, and it can begin dumping mass onto the white dwarf. This can cause the white dwarf to exceed the *Chandrasekhar limit* (Section 4.4), resulting in an explosion known as a type Ia supernova. Because the Chandrasekhar limit provides a uniform set of initial conditions, the behavior of type Ia supernovae is fairly predictable, and in particular their luminosities are approximately equal. They therefore provide a kind of standard candle: since the intrinsic brightness is known, the distance can be inferred from the apparent brightness. Given the distance, we can infer the time that was spent in transit by the light on its way to us, i.e. the look-back time. From measurements of Doppler shifts of spectral lines, we can also find the velocity at which the supernova was receding from us. The result is that we can measure the universe's rate of expansion as a function of time. Observations show that this rate of expansion has been accelerating. The Friedmann equations show that this can only occur for $\Lambda \gtrsim 4\rho$. This picture has been independently verified by measurements of the cosmic microwave background (CMB) radiation. A more detailed discussion of the supernova and CMB data is given in Section 8.2.

With hindsight, we can see that in a quantum-mechanical context, it is natural to expect that fluctuations of the vacuum, required by the Heisenberg uncertainty principle, would contribute to the cosmological constant, and in fact models tend to overpredict Λ by a factor of about 10^{120}! From this point of view, the mystery is why these effects cancel out so precisely. A correct understanding of the cosmological constant presumably requires a full theory of quantum gravity, which is presently far out of our reach.

The latest data show that our universe, in the present epoch, is dominated by the cosmological constant, so as an approximation we can write the Friedmann equations as

$$\frac{\ddot{a}}{a} = \frac{1}{3}\Lambda$$
$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{1}{3}\Lambda.$$

This is referred to as a *vacuum-dominated universe* or the *de Sitter spacetime*. The solution is

$$a = e^{\sqrt{\frac{\Lambda}{3}}t}, \tag{8.2.5}$$

where observations show that $\Lambda \sim 10^{-26}$ kg/m³, giving $\sqrt{\frac{3}{\Lambda}} \sim 10^{11}$ years.

The implications for the fate of the universe are depressing. All parts of the universe will accelerate away from one another faster and faster as time goes on. The relative separation between two objects, say galaxy A and galaxy B, will eventually be increasing faster than the speed of light. (The Lorentzian character of spacetime is local, so relative motion faster than c is only forbidden between objects that are passing right by one another.) At this point, an observer in either galaxy will say that the other one has passed behind an event horizon. If intelligent observers do actually exist in the far future, they may have no way to tell that the cosmos even exists. They will perceive themselves as living in island universes, such as we believed our own galaxy to be a hundred years ago.

When I introduced the standard cosmological coordinates earlier, I described them as coordinates in which events that are simultaneous according to this t are events at which the local properties of the universe are the same. In the case of a perfectly vacuum-dominated universe, however, this notion loses its meaning. The only observable local property of such a universe is the vacuum energy described by the cosmological constant, and its density is always the same, because it is built into the structure of the vacuum. Thus the vacuum-dominated cosmology is a special one that maximally symmetric, in the sense that it has not only the symmetries of homogeneity and isotropy that we've been assuming all along, but also a symmetry with respect to time: it is a cosmology without history, in which all times appear identical to a local observer. One way of checking this claim is by calculating

curvature scalars, and we find, for example, that the Ricci scalar is a constant R = −12Λ (with the sign depending on the + − −− signature, example 25).

In the special case of this cosmology, the time variation of the scaling factor a(t) is unobservable, and may be thought of as the unfortunate result of choosing an inappropriate set of coordinates, which obscure the underlying symmetry. When I argued in section 8.2 for the observability of the universe's expansion, note that all my arguments assumed the presence of matter or radiation. These are completely absent in a perfectly vacuum-dominated cosmology.

For these reasons de Sitter originally proposed this solution as a static universe in 1927. But by 1920 it was realized that this was an oversimplification. The argument above only shows that the time variation of a(t) does not allow us to distinguish one epoch of the universe from another. That is, we can't look out the window and infer the date (e.g., from the temperature of the cosmic microwave background radiation). It does not, however, imply that the universe is static in the sense that had been assumed until Hubble's observations. The r-t part of the metric is

$$ds^2 = dt^2 - a^2 dr^2, \tag{8.2.6}$$

where a blows up exponentially with time, and the k-dependence has been neglected, as it was in the approximation to the Friedmann equations used to derive $a(t)$.[21] Let a test particle travel in the radial direction, starting at event A = (0, 0) and ending at B = (t', r'). In flat space, a world-line of the linear form r = vt would be a geodesic connecting A and B; it would maximize the particle's proper time. But in the this metric, it cannot be a geodesic. The curvature of geodesics relative to a line on an r-t plot is most easily understood in the limit where t 0 p is fairly long compared to the time-scale T = $\sqrt{\frac{3}{\Lambda}}$ of the exponential, so that a(t') is huge. The particle's best strategy for maximizing its proper time is to make sure that its dr is extremely small when a is extremely large. The geodesic must therefore have nearly constant r at the end. This makes it sound as though the particle was decelerating, but in fact the opposite is true. If r is constant, then the particle's spacelike distance from the origin is just ra(t), which blows up exponentially. The near-constancy of the coordinate r at large t actually means that the particle's motion at large t isn't really due to the particle's inertial memory of its original motion, as in Newton's first law. What happens instead is that the particle's initial motion allows it to move some distance away from the origin during a time on the order of T, but after that, the expansion of the universe has become so rapid that the particle's motion simply streams outward because of the expansion of space itself. Its initial motion only mattered because it determined how far out the particle got before being swept away by the exponential expansion.

> **Note**
>
> A computation of the Einstein tensor with $ds^2 = d\bar{t}^2 - a^2 (1 - kr^2)^{-1} dr^2$ shows that k enters only via a factor the form $(\ldots)e^{(\ldots)t}$ + $(\ldots)$k. For large t, the k term becomes negligible, and the Einstein tensor becomes $G^a{}_b = g^a{}_b\Lambda$, This is consistent with the approximation we used in deriving the solution, which was to ignore both the source terms and the k term in the Friedmann equations. The exact solutions with Λ > 0 and k = −1, 0, and 1 turn out in fact to be equivalent except for a change of coordinates.

> **Example 19: Geodesics in a vacuum-dominated universe**
>
> In this example we confirm the above interpretation in the special case where the particle, rather than being released in motion at the origin, is released at some nonzero radius r, with $\frac{dr}{dt}$ = 0 initially. First we recall the geodesic equation
>
> $$\frac{d^2 x^i}{d\lambda^2} = \Gamma^i{}_{jk} \frac{dx^j}{d\lambda} \frac{dx^k}{d\lambda}. \tag{8.2.7}$$
>
> from section 5.7. The nonvanishing Christoffel symbols for the 1+1- dimensional metric $ds^2 = dt^2 - a^2 dr^2$ are $\Gamma^r{}_{tr} = \frac{\dot{a}}{a}$ and $\Gamma^t{}_{rr} = \dot{a}a$. Setting T = 1 for convenience, we have $\Gamma^r{}_{tr} = 1$ and $\Gamma^t{}_{rr} = e^{-2t}$.
>
> We conjecture that the particle remains at the same value of r. Given this conjecture, the particle's proper time $\int$ ds is simply the same as its time coordinate t, and we can therefore use t as an affine coordinate. Letting $\lambda = t$, we have
>
> $$\frac{d^2 t}{dt^2} - \Gamma^t{}_{rr} \left(\frac{dr}{dt}\right)^2 = 0 \tag{8.2.8}$$
>
> $$0 - \Gamma^t{}_{rr} \dot{r}^2 = 0 \tag{8.2.9}$$
>
> $$\dot{r} = 0 \tag{8.2.10}$$

$$r = constant \tag{8.2.11}$$

This confirms the self-consistency of the conjecture that r = constant is a geodesic.

Note that we never actually had to use the actual expressions for the Christoffel symbols; we only needed to know which of them vanished and which didn't. The conclusion depended only on the fact that the metric had the form ds^2 = dt^2 − a^{2}dr^2 for some function a(t). This provides a rigorous justification for the interpretation of the cosmological scale factor a as giving a universal time-variation on all distance scales.

The calculation also confirms that there is nothing special about r = 0. A particle released with r = 0 and $\dot{r}$ = 0 initially stays at r = 0, but a particle released at any other value of r also stays at that r. This cosmology is homogeneous, so any point could have been chosen as r = 0. If we sprinkle test particles, all at rest, across the surface of a sphere centered on this arbitrarily chosen point, then they will all accelerate outward relative to one another, and the volume of the sphere will increase. This is exactly what we expect. The Ricci curvature is interpreted as the second derivative of the volume of a region of space defined by test particles in this way. The fact that the second derivative is positive rather than negative tells us that we are observing the kind of repulsion provided by the cosmological constant, not the attraction that results from the existence of material sources.

Example 20: Schwarzschild-de Sitter space

The metric

$$ds^2 = \left(1 - \frac{2m}{r} - \frac{1}{3}\Lambda r^2\right)dt^2 - \frac{dr^2}{1 - \frac{2m}{r} - \frac{1}{3}\Lambda r^2} - r^2 d\theta^2 - r^2 \sin^2\theta d\phi^2 \tag{8.2.12}$$

is an exact solution to the Einstein field equations with cosmological constant Λ, and can be interpreted as a universe in which the only mass is a black hole of mass m located at r = 0. Near the black hole, the Λ terms become negligible, and this is simply the Schwarzschild metric. As argued in section 8.2, this is a simple example of how cosmological expansion does not cause all structures in the universe to grow at the same rate.

Example 21: Conservation of energy-momentum

Suppose that we assume the de Sitter geometry, and ask what type of matter fields are necessary to create it. We know that a cosmological constant will do the job, but could we have some other matter field that would also work? Suppose that the matter field is constrained to be a perfect fluid. The total stress-energy is then of the form T_ν^μ = diag(ρ, −P, −P, −P) in Cartesian coordinates. (See example 4 for the signs, some of which depend on our use of the + − −− signature.) The divergence $\nabla_\mu T_t^\mu$ measures the rate at which an observer says energy is being created, and we need this to be zero. This expression is one of those tricky examples where the covariant derivative can be nonzero even when the thing being differentiated vanishes identically. The divergence is $\nabla_t T_t^t + \nabla_x T_t^x$, and the term that doesn't vanish is the second one, even though T^x$_t$ = 0. Using the nonvanishing Christoffel symbols this becomes $\Gamma_{xt}^x T_t^t - \Gamma_{tx}^x T_x^x = \frac{\dot{a}}{a}(\rho + P)$, so that ρ + P = 0. This condition is satisfied by a cosmological constant. Our result is that the only way to get a de Sitter geometry is with matter fields that exactly mimic a cosmological constant. This is of some historical interest in the context of the steady-state cosmologies, section 8.4. It may seem mysterious that we have obtained this result by requiring conservation of energy-momentum, but we could also have done it using the Einstein field equations. In fact these are not two separate requirements, since the field equations require conservation of energy-momentum in order to be consistent.

The Big Bang Singularity in a Universe with a Cosmological Constant

Earlier we discussed the possibility that the Big Bang singularity was an artifact of the unrealistically perfect symmetry assumed by our cosmological models, and we found that this was not the case: the Penrose-Hawking singularity theorems demonstrate that the singularity is real, provided that the cosmological constant is zero. The cosmological constant is *not* zero, however. Models with a very large positive cosmological constant can also display a Big Bounce rather than a Big Bang. If we imagine using the Friedmann equations to evolve the universe backward in time from its present state, the scaling arguments of example 14 suggest that at early enough times, radiation and matter should dominate over the cosmological constant. For a large enough value of the cosmological constant, however, it can happen that this switch-over never happens. In such a model, the universe is and always has been dominated by the cosmological constant, and we get a Big Bounce in the past because of the cosmological constant's repulsion. In this book I will only develop simple cosmological models in which the universe is dominated by a single component; for a

discussion of bouncing models with both matter and a cosmological constant, see Carroll, "The Cosmological Constant," http://www.livingreviews.org/lrr-2001-1. By 2008, a variety of observational data had pinned down the cosmological constant well enough to rule out the possibility of a bounce caused by a very strong cosmological constant.

The Matter-dominated Solution

Our universe is not perfectly vacuum-dominated, and in the past it was even less so. Let us consider the matter-dominated epoch, in which the cosmological constant was negligible compared to the material sources. The equation of state for nonrelativistic matter (example 4) is

$$P = 0. \tag{8.2.13}$$

The dilution of the dust with cosmological expansion gives

$$\rho \propto a^{-3} \tag{8.2.14}$$

(see example 23). The Friedmann equations become

$$\frac{\ddot{a}}{a} = -\frac{4\pi}{3}\rho$$
$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi}{3}\rho - ka^{-2},$$

where for compactness ρ's dependence on a, with some constant of proportionality, is not shown explicitly. A static solution, with constant a, is impossible, and $\ddot{a}$ is negative, which we can interpret in Newtonian terms as the deceleration of the matter in the universe due to gravitational attraction. There are three cases to consider, according to the value of k.

The Closed Universe

We've seen that k = +1 describes a universe in which the spatial curvature is positive, i.e., the circumference of a circle is less than its Euclidean value. By analogy with a sphere, which is the twodimensional surface of constant positive curvature, we expect that the total volume of this universe is finite.

The second Friedmann equation also shows us that at some value of a, we will have $\dot{a} = 0$. The universe will expand, stop, and then recollapse, eventually coming back together in a "Big Crunch" which is the time-reversed version of the Big Bang.

Suppose we were to describe an initial-value problem in this cosmology, in which the initial conditions are given for all points in the universe on some spacelike surface, say t = constant. Since the universe is assumed to be homogeneous at all times, there are really only three numbers to specify, a, $\dot{a}$, and ρ: how big is the universe, how fast is it expanding, and how much matter is in it? But these three pieces of data may or may not be consistent with the second Friedmann equation. That is, the problem is overdetermined. In particular, we can see that for small enough values of ρ, we do not have a valid solution, since the square of $\frac{\dot{a}}{a}$ would have to be negative. Thus a closed universe requires a certain amount of matter in it. The present observational evidence (from supernovae and the cosmic microwave background, as described above) is sufficient to show that our universe does not contain this much matter.

The Flat Universe

The case of k = 0 describes a universe that is spatially flat. It represents a knife-edge case lying between the closed and open universes. In a Newtonian analogy, it represents the case in which the universe is moving exactly at escape velocity; as t approaches infinity, we have a → ∞, ρ → 0, and $\dot{a}$ → 0. This case, unlike the others, allows an easy closed-form solution to the motion. Let the constant of proportionality in the equation of state $\rho \propto a^{-3}$ be fixed by setting $\frac{-4\pi\rho}{3} = -ca^{-3}$. The Friedmann equations are

$$\ddot{a} = -ca^{-2}$$
$$\dot{a} = \sqrt{2c}a^{-1/2}.$$

Looking for a solution of the form $a \propto t^p$, we find that by choosing p = 2/3 we can simultaneously satisfy both equations. The constant c is also fixed, and we can investigate this most transparently by recognizing that $\frac{\dot{a}}{a}$ is interpreted as the Hubble constant, H, which is the constant of proportionality relating a far-off galaxy's velocity to its distance. Note that H is a "constant" in the sense that it is the same for all galaxies, in this particular model with a vanishing cosmological constant; it does not stay constant with the

passage of cosmological time. Plugging back into the original form of the Friedmann equations, we find that the flat universe can only exist if the density of matter satisfies $\rho = \rho_{crit} = \frac{3H^2}{8\pi} = \frac{3H^2}{8\pi G}$. The observed value of the Hubble constant is about 1/(14 × 10^9 years), which is roughly interpreted as the age of the universe, i.e., the proper time experienced by a test particle since the Big Bang. This gives $\rho_{crit} \sim 10^{-26}$ kg/m^3.

As discussed in section 8.2, our universe turns out to be almost exactly spatially flat. Although it is presently vacuum-dominated, the flat and matter-dominated FRW cosmology is a useful description of its matter-dominated era.

The Open Universe

The k = −1 case represents a universe that has negative spatial curvature, is spatially infinite, and is also infinite in time, i.e., even if the cosmological constant had been zero, the expansion of the universe would have had too little matter in it to cause it to recontract and end in a Big Crunch.

The time-reversal symmetry of general relativity was discussed in section 6.2 in connection with the Schwarzschild metric.[22] Because of this symmetry, we expect that solutions to the field equations will be symmetric under time reversal (unless asymmetric boundary conditions were imposed). The closed universe has exactly this type of time-reversal symmetry. But the open universe clearly breaks this symmetry, and this is why we speak of the Big Bang as lying in the past, not in the future. This is an example of spontaneous symmetry breaking. Spontaneous symmetry breaking happens when we try to balance a pencil on its tip, and it is also an important phenomenon in particle physics. The time-reversed version of the open universe is an equally valid solution of the field equations. Another example of spontaneous symmetry breaking in cosmological solutions is that the solutions have a preferred frame of reference, which is the one at rest relative to the cosmic microwave background and the average motion of the galaxies. This is referred to as the Hubble flow.

> **Note**
>
> Problem 5 shows that this symmetry is also exhibited by the Friedmann equations.

> **Example 22: Size and age of the observable universe**
>
> The observable universe is defined by the region from which light has had time to reach us since the Big Bang. Many people are inclined to assume that its radius in units of light-years must therefore be equal to the age of the universe expressed in years. This is not true. Cosmological distances like these are not even uniquely defined, because general relativity only has local frames of reference, not global ones.
>
> Suppose we adopt the proper distance L defined in section 8.2 as our measure of radius. By this measure, realistic cosmological models say that our 14-billion-year-old universe has a radius of 46 billion light years.
>
> For a flat universe, f = 1, so by inspecting the FRW metric we find that a photon moving radially with ds = 0 has $|\frac{dr}{dt}| = a^{-1}$, giving $r = \pm \int_{t_1}^{t_2} \frac{dt}{a}$. Suppressing signs, the proper distance the photon traverses starting soon after the Big Bang is L = a(t$_2$) $\int d\ell = a(t_2) \int dr = a(t_2)r = a(t_2) \int_{t_1}^{t_2} \frac{dt}{a}$. In the matter-dominated case, a $\propto$ t$^{2/3}$, so this results in L = 3t$_2$ in the limit where t1 is small. Our universe has spent most of its history being matter-dominated, so it's encouraging that the matter-dominated calculation seems to do a pretty good job of reproducing the actual ratio of $\frac{46}{14}$ = 3.3 between L and t$_2$.
>
> While we're at it, we can see what happens in the purely vacuumdominated case, which has a $\propto$ e$^{t/T}$, where T = $\sqrt{\frac{3}{\Lambda}}$. This cosmology doesn't have a Big Bang, but we can think of it as an approximation to the more recent history of the universe, glued on to an earlier matter-dominated solution. Here we find L = [e$^{(t_2-t_1)/T}$ − 1]T, where t$_1$ is the time when the switch to vacuum-domination happened. This function grows more quickly with t$_2$ than the one obtained in the matter-dominated case, so it makes sense that the real-world ratio of $\frac{L}{t_2}$ is somewhat greater than the matter-dominated value of 3.
>
> The radiation-dominated version is handled in problem 12.

> **Example 23: Local conservation of mass-energy**
>
> Any solution to the Friedmann equations is a solution of the field equations, and therefore locally conserves mass-energy. We saved work above by applying this condition in advance in the form $\rho \propto$ a^{-3} to make the dust dilute itself properly with cosmological expansion. In this example we prove the same proportionality by explicit calculation.

Local conservation of mass-energy is expressed by the zero divergence of the stress-energy tensor, $\nabla_j T^{jb} = 0$. The definition of the covariant derivative gives

$$\nabla_j T^{bc} = \partial_j T^{bc} + \Gamma^b_{jd} T^{dc} + \Gamma^c_{jd} T^{bd}.\tag{8.2.15}$$

For convenience, we carry out the calculation at r = 0; if conservation holds here, then it holds everywhere by homogeneity.

In a local Cartesian frame (t', x', y', z') at rest relative to the dust, the stress-energy tensor is diagonal with $T^{t't'} = \rho$. At r = 0, the transformation from FLRW coordinates into these coordinates doesn't mix t or t' with the other coordinates, so by the tensor transformation law we still have $T^{tt} = \rho$.

There are a number of Christoffel symbols involved, but the only three of relevance that don't vanish at r = 0 turn out to be $\Gamma^r_{rt} = \Gamma^\theta_{\theta t} = \Gamma^\phi_{\phi t} = \frac{\dot{a}}{a}$. The result is

$$\nabla_\mu T^{t\mu} = \partial_t T^{tt} + 3\frac{\dot{a}}{a} T^{tt},\tag{8.2.16}$$

or $\frac{\dot{\rho}}{\rho} = -3\frac{\dot{a}}{a}$, which can be rewritten as

$$\frac{d}{dt}\ln\rho = -3\frac{d}{dt}\ln a,\tag{8.2.17}$$

producing the proportionality originally claimed.

8.6: Sources in General Relativity (Part 3)

Geodesic motion of test particles

Question 1 was: "Does a small enough physical object always have a world-line that is approximately a geodesic?" In other words, do Eötvös experiments give null results when carried out in laboratories using real-world apparatus of small enough size? We would like something of this type to be true, since general relativity is based on the equivalence principle, and the equivalence principle is motivated by the null results of Eötvös experiments. Nevertheless, it is fairly easy to show that the answer to the question is no, unless we make some more specific assumption, such as an energy condition, about the system being modeled.

Before we worry about energy conditions, let's consider why the small size of the apparatus is relevant. Essentially this is because of gravitational radiation. In a gravitationally radiating system such as the Hulse-Taylor binary pulsar (section 6.2), the material bodies lose energy, and as with any radiation process, the radiated power depends on the square of the strength of the source. The world-line of a such a body therefore depends on its mass, and this shows that its world-line cannot be an exact geodesic, since the initially tangent world-lines of two different masses diverge from one another, and these two world-lines can't both be geodesics.

Let's proceed to give a rough argument for geodesic motion and then try to poke holes in it. When we test geodesic motion, we do an Eötvös experiment that is restricted to a certain small region of spacetime S. Our test-body's world-line enters S with a certain energy-momentum vector p and exits with p'. If spacetime was flat, then Gauss's theorem would hold exactly, and the vanishing divergence $\nabla_b T^{ab}$ of the stress-energy tensor would require that the incoming flux represented by p be exactly canceled by the outgoing flux due to p'. In reality spacetime isn't flat, and it isn't even possible to compare p and p' except by parallel-transporting one into the same location as the other. Parallel transport is path-dependent, but if we make the reasonable restriction to paths that stay within S, we expect the ambiguity due to path-dependence to be proportional to the area enclosed by any two paths, so that if S is small enough, the ambiguity can be made small. Ignoring this small ambiguity, we can see that one way for the fluxes to cancel would be for the particle to travel along a geodesic, since both p and p' are tangent to the test-body's world-line, and a geodesic is a curve that parallel-transports its own tangent vector. Geodesic motion is therefore one solution, and we expect the solution to be nearly unique when S is small.

Although this argument is almost right, it has some problems. First we have to ask whether "geodesic" means a geodesic of the full spacetime including the object's own fields, or of the background spacetime B that would have existed without the object. The latter is the more sensible interpretation, since the question is basically asking whether a spacetime can really be defined geometrically, as the equivalence principle claims, based on the motion of test particles inserted into it. We also have to define words like "small enough" and "approximately;" to do this, we imagine a sequence of objects On that get smaller and smaller as n increases. We then form the following conjecture, which is meant to formulate question 1 more exactly: Given a vacuum background spacetime B, and a timelike world-line ℓ in B, consider a sequence of spacetimes Sn, formed by inserting the On into B, such that: (i) the metric of S_n is defined on the same points as the metric of B; (ii) On moves along ℓ, and for any r > 0, there exists some n such that for m ≥ n, O_m is smaller than r;[9] (iii) the metric of Sn approaches the metric of B as n → ∞. Then ℓ is a geodesic of B.

> **Note**
>
> i.e., at any point P on ℓ, an observer moving along ℓ at P defines a surface of simultaneity K passing through P, and sees the stress-energy tensor of On as vanishing outside of a three-sphere of radius r within K and centered on P

This is almost right but not quite, as shown by the following counterexample. Papapetrou[10] has shown that a spinning body in a curved background spacetime deviates from a geodesic with an acceleration that is proportional to LR, where L is its angular momentum and R is the Riemann curvature. Let all the O_n have a fixed value of L, but let the spinning mass be concentrated into a smaller and smaller region as n increases, so as to satisfy (ii). As the radius r decreases, the motion of the particles composing an On eventually has to become ultrarelativistic, so that the main contribution to the gravitational field is from the particles' kinetic energy rather than their rest mass. We then have L ~ pr ~ Er, so that in order to keep L constant, we must have $E \propto \frac{1}{r}$. This causes two problems. First, it makes the gravitational field blow up at small distances, violating (iii). Also, we expect that for any known form of matter, there will come a point (probably the Tolman-Oppenheimer-Volkoff limit) at which we get a black hole; the singularity is then not part of the spacetime S_n, violating (i). But our failed counterexample can be patched up. We obtain a supply of exotic matter, whose gravitational mass is negative, and we mix enough of this mysterious stuff into each O_n so that the gravitational field shrinks rather than growing as n increases, and no black hole is ever formed.

Ehlers and Geroch[11] have proved that it suffices to require an additional condition: (iv) The O_n satisfy the dominant energy condition. This rules out our counterexample.

The Newtonian limit

In units with c ≠ 1, a quantity like ρ + P is expressed as $\frac{\rho+P}{c^2}$. The Newtonian limit is recovered as c → ∞, which makes the pressure term negligible, so that all the energy conditions reduce to $\rho \geq 0$. What would it mean if this was violated? Would $\rho < 0$ describe an object with negative inertial mass, which would accelerate east when you pushed it to the west? Or would it describe something with negative gravitational mass, which would repel ordinary matter? We can imagine various possiblities, as shown in Figure 8.1.5. Anything that didn't lie on the main diagonal would violate the equivalence principle, and would therefore be impossible to accomodate within general relativity's geometrical description of gravity. If we had "upsidasium" matter such as that described by the second quadrant of the figure (example 2), gravity would be like electricity, except that like masses would attract and opposites repel; we could have gravitational dielectrics and gravitational Faraday cages. The fourth quadrant leads to amusing possibilities like Figure 8.1.6.
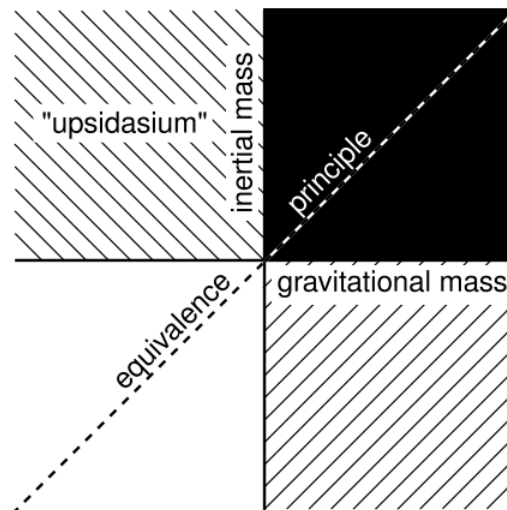
Figure 8.6.6 - Negative Mass.

Figure 8.6.7 - The black sphere is made of ordinary matter. The crosshatched sphere has positive gravitational mass and negative inertial mass. If the two of them are placed side by side in empty space, they will both accelerate steadily to the right, gradually approaching the speed of light. Conservation of momentum is preserved, because the exotic sphere has leftward momentum when it moves to the right, so the total momentum is always zero.

Example 10: no gravitational shielding

Electric fields can be completely excluded from a Faraday cage, and magnetic fields can be very strongly blocked with high-permeability materials such as mu-metal. It would be fun if we could do the same with gravitational fields, so that we could have zero-gravity or near-zero-gravity parties in a specially shielded room. It would be a form of antigravity, but a different one than the "upsidasium" type. Unfortunately this is difficult to do, and the reason it's difficult turns out to be related to the unavailability of materials that violate energy conditions.

First we need to define what we mean by shielding. We restrict ourselves to the Newtonian limit, and to one dimension, so that a gravitational field is specified by a function of one variable g(x). The best kind of shielding would be some substance that we could cut with shears and form into a box, and that would exclude gravitational fields from the interior of the box. This would be analogous to a Faraday cage; no matter what external field it was embedded in, it would spontaneously adjust itself so that the internal field was canceled out. A less desirable kind of shielding would be one that we could set up on an *ad hoc* basis to null out a specific, given, externally imposed field. Once we know what the external field is, we try to choose some

arrangement of masses such that the field is nulled out. We will show that even this kind of shielding is unachievable, if nulling out the field is interpreted to mean this: at some point, which for convenience we take to be the origin, we wish to have a gravitational field such that g(0) = 0, $\frac{dg}{dx(0)}$ = 0, . . . $\frac{d^n g}{dx^n(0)}$ = 0, where n is arbitrarily specified. For comparison, *magnetic* fields can be nulled out according to this definition by building an appropriately chosen configuration of coils such as a Helmholtz coil.

Since we're only doing the Newtonian limit, the gravitational field is the sum of the fields made by all the sources, and we can take this as a sum over point sources. For a point source m placed at x_0, the field g(x) is odd under reflection about x_0. The derivative of the field g'(x) is even. Since g' is even, we can't control its sign at x = 0 by choosing $x_0 > 0$ or $x_0 < 0$. The only way to control the sign of g' is by choosing the sign of m. Therefore if the sign of the externally imposed field's derivative is wrong, we can never null it out. Figure 8.1.8 shows a special case of this theorem. The theorem does not apply to three dimensions, and it does not prove that all fields are impossible to null out, only that some are. For example, the field inside a hemispherical shell can be nulled by adding another hemispherical shell to complete the sphere. I thank P. Allen for helpful discussion of this topic.
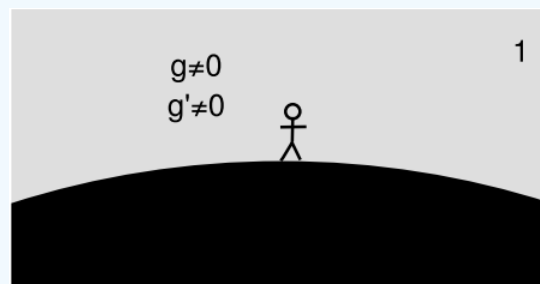
Figure_8.1.8b.png" />
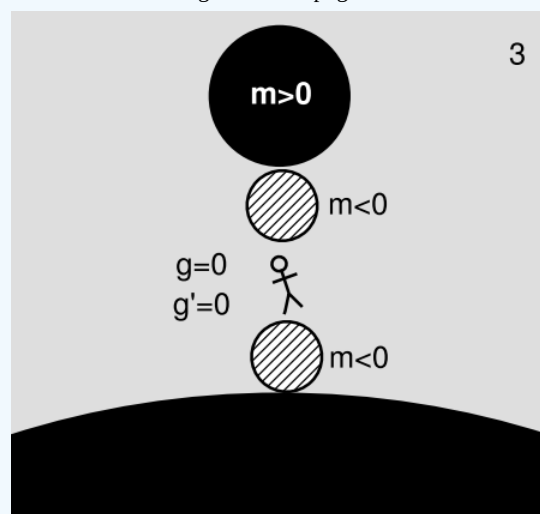
Figure 8.6.8 - Nulling out a gravitational field is impossible in one dimension without exotic matter. 1. The planet imposes a nonvanishing gravitational field with a nonvanishing gradient. 2. We can null the field at one point in space, by placing a sphere of very dense, but otherwise normal, matter overhead. The stick figure still experiences a tidal force, g' ≠ 0. 3. To change the field's derivative without changing the field, we can place two additional masses above and below the given point. But to change its derivative in the desired direction — toward zero — we would have to make these masses negative.

Singularity Theorems

An important example of the use of the energy conditions is that Hawking and Ellis have proved that under the assumption of the strong energy condition, any body that becomes sufficiently compact will end up forming a singularity. We might imagine that the formation of a black hole would be a delicate thing, requiring perfectly symmetric initial conditions in order to end up with the perfectly symmetric Schwarzschild metric. Many early relativists thought so, for good reasons. If we look around the universe at various scales, we find that collisions between astronomical bodies are extremely rare. This is partly because the distances are vast compared to the sizes of the objects, but also because conservation of angular momentum has a tendency to make objects swing past one another rather than colliding head-on. Starting with a cloud of objects, e.g., a globular cluster, Newton's laws make it

extremely difficult, regardless of the attractive nature of gravity, to pick initial conditions that will make them all collide the future. For one thing, they would have to have exactly zero total angular momentum.

Most relativists now believe that this is not the case. General relativity describes gravity in terms of the tipping of light cones. When the field is strong enough, there is a tendency for the light cones to tip over so far that the entire future light-cone points at the source of the field. If this occurs on an entire surface surrounding the source, it is referred to as a trapped surface.

To make this notion of light cones "pointing at the source" more rigorous, we need to define the volume expansion Θ. Let the set of all points in a spacetime (or some open subset of it) be expressed as the union of geodesics. This is referred to as a foliation in geodesics, or a congruence. Let the velocity vector tangent to such a curve be u^a. Then we define $\Theta = \nabla_a u^a$. This is exactly analogous to the classical notion of the divergence of the velocity field of a fluid, which is a measure of compression or expansion. Since Θ is a scalar, it is coordinate-independent. Negative values of Θ indicate that the geodesics are converging, so that volumes of space shrink. A trapped surface is one on which Θ is negative when we foliate with lightlike geodesics oriented outward along normals to the surface.

When a trapped surface forms, any lumpiness or rotation in the initial conditions becomes irrelevant, because every particle's entire future world-line lies inward rather than outward. A possible loophole in this argument is the question of whether the light cones will really tip over far enough. We could imagine that under extreme conditions of high density and temperature, matter might demonstrate unusual behavior, perhaps including a negative energy density, which would then give rise to a gravitational repulsion. Gravitational repulsion would tend to make the light cones tip outward rather than inward, possibly preventing the collapse to a singularity. We can close this loophole by assuming an appropriate energy condition. Penrose and Hawking have formalized the above argument in the form of a pair of theorems, known as the singularity theorems. One of these applies to the formation of black holes, and another one to cosmological singularities such as the Big Bang.

In a cosmological model, it is natural to foliate using worldlines that are at rest relative to the Hubble flow (or, equivalently, the world-lines of observers who see a vanishing dipole moment in the cosmic microwave background). The Θ we then obtain is positive, because the universe is expanding. The volume expansion is $\Theta = 3H_0$, where $H_0 \approx 2.3 \times 10^{-18}$ s^{-1} is the Hubble constant (the fractional rate of change of the scale factor of cosmological distances). The factor of three occurs because volume is proportional to the cube of the linear dimensions.

Current status

The current status of the energy conditions is shaky. Although it is clear that all of them hold in a variety of situations, there are strong reasons to believe that they are violated at both microscopic and cosmological scales, for reasons both classical and quantummechanical.[12] We will see such a violation in the following section.

The Cosmological Constant

Having included the source term in the Einstein field equations, our most important application will be to cosmology. Some of the relevant ideas originate long before Einstein. Once Newton had formulated a theory of gravity as a universal attractive force, he realized that there would be a tendency for the universe to collapse. He resolved this difficulty by assuming that the universe was infinite in spatial extent, so that it would have no center of symmetry, and therefore no preferred point to collapse toward. The trouble with this argument is that the equilibrium it describes is unstable. Any perturbation of the uniform density of matter breaks the symmetry, leading to the collapse of some pocket of the universe. If the radius of such a collapsing region is r, then its gravitational is proportional to r^3, and its gravitational field is proportional to $\frac{r^3}{r^2}$ = r. Since its acceleration is proportional to its own size, the time it takes to collapse is independent of its size. The prediction is that the universe will have a self-similar structure, in which the clumping on small scales behaves in the same way as clumping on large scales; zooming in or out in such a picture gives a landscape that appears the same. With modern hindsight, this is actually not in bad agreement with reality. We observe that the universe has a hierarchical structure consisting of solar systems, galaxies, clusters of galaxies, superclusters, and so on. Once such a structure starts to condense, the collapse tends to stop at some point because of conservation of angular momentum. This is what happened, for example, when our own solar system formed out of a cloud of gas and dust.

Einstein confronted similar issues, but in a more acute form. Newton's symmetry argument, which failed only because of its instability, fails even more badly in relativity: the entire spacetime can simply contract uniformly over time, without singling out any particular point as a center. Furthermore, it is not obvious that angular momentum prevents total collapse in relativity in the same way that it does classically, and even if it did, how would that apply to the universe as a whole? Einstein's Machian orientation would have led him to reject the idea that the universe as a whole could be in a state of rotation, and in any case it was

sensible to start the study of relativistic cosmology with the simplest and most symmetric possible models, which would have no preferred axis of rotation.

Because of these issues, Einstein decided to try to patch up his field equation so that it would allow a static universe. Looking back over the considerations that led us to this form of the equation, we see that it is very nearly uniquely determined by the following criteria:

1. It should be consistent with experimental evidence for local conservation of energy-momentum.
2. It should satisfy the equivalence principle.
3. It should be coordinate-independent.
4. It should be equivalent to Newtonian gravity or "plain" general relativity in the appropriate limit.
5. It should not be overdetermined.

This is not meant to be a rigorous proof, just a general observation that it's not easy to tinker with the theory without breaking it.

> **Example 11: A failed attempt at tinkering**
>
> As an example of the lack of "wiggle room" in the structure of the field equations, suppose we construct the scalar T^a_a, the trace of the stress-energy tensor, and try to insert it into the field equations as a further source term. The first problem is that the field equation involves rank-2 tensors, so we can't just add a scalar. To get around this, suppose we multiply by the metric. We then have something like $G_{ab} = c_1 T_{ab} + c_2 g_{ab} T^c_c$, where the two constants c1 and c2 would be constrained by the requirement that the theory agree with Newtonian gravity in the classical limit.
>
> To see why this attempt fails, note that the stress-energy tensor of an electromagnetic field is traceless, $T^c_c = 0$. Therefore the beam of light's coupling to gravity in the c_2 term is zero. As discussed earlier, empirical tests of conservation of momentum would therefore constrain c_2 to be $\lesssim 10^{-8}$.

One way in which we can change the field equation without violating any of these requirements is to add a term Λg_{ab}, giving

$$G_{ab} = 8\pi T_{ab} + \Lambda g_{ab}, \tag{8.1.22}$$

which is what we will refer to as the Einstein field equation.[13] As we'll see in example 12, this is consistent with conservation of energy-momentum (requirement 1 above) if and only if Λ is constant. In example 13 we find that its effects are only significant on the largest scales, which makes it undetectable, for example, in solar-system tests (criterion 4). For these reasons Λ is referred to as the cosmological constant. As we'll see below, Einstein introduced it in order to make a certain type of cosmology work.

> **Note**
>
> In books that use a $-+++$ metric rather then our $+---$, the sign of the cosmological constant term is reversed relative to ours.

We could also choose to absorb the Λg_{ab} term in the field equations into the $8\pi T_{ab}$, as if the cosmological constant term were due to some form of matter. It would then be a perfect fluid (example 4) with a negative pressure, and it would violate the strong energy condition (example 14). When we think of it this way, it's common these days to refer to it as dark energy. But even if we think of it as analogous to a matter field, its constancy means that it has none of its own independent degrees of freedom. It can't vibrate, rotate, flow, be compressed or rarefied, heated or cooled. It acts like a kind of energy that is automatically built in to every cubic centimeter of space. This is closely related to the fact that its contribution to the stress-energy tensor is proportional to the metric. One way of stating the equivalence principle (requirement 2 above) is that space itself does not come equipped with any other tensor besides the metric.

Einstein originally introduced a positive cosmological constant because he wanted relativity to be able to describe a static universe. To see why it would have this effect, compare its behavior with that of an ordinary fluid. When an ordinary fluid, such as the exploding air-gas mixture in a car's cylinder, expands, it does work on its environment, and therefore by conservation of energy its own internal energy is reduced. A positive cosmological constant, however, acts like a certain amount of mass-energy built into every cubic meter of vacuum. Thus when it expands, it *releases* energy. Its pressure is negative.

Now consider the following semi-relativistic argument. Although we've already seen (section 6.2) that there is no useful way to separate the roles of kinetic and potential energy in general relativity, suppose that there are some quantities analogous to them in the description of the universe as a whole. (We'll see below that the universe's contraction and expansion is indeed described by a

set of differential equations that can be interpreted in essentially this way.) If the universe contracts, a cubic meter of space becomes less than a cubic meter. The cosmological-constant energy associated with that volume is reduced, so some energy has been consumed. The kinetic energy of the collapsing matter goes down, and the collapse is decelerated.

Example 12: Cosmological constant must be constant

If Λ is thought of as a form of matter, then it becomes natural to ask whether it's spread more thickly in some places than others: is the cosmological "constant" really constant? The following argument shows that it cannot vary. The field equations are $G_{ab} = 8\pi T_{ab} + \Lambda g_{ab}$. Taking the divergence of both sides, we have $\nabla^a G_{ab} = 8\pi \nabla^a T_{ab} + \nabla^a(\Lambda g_{ab})$. The left-hand side vanishes (see earlier). Since laboratory experiments have verified conservation of mass-energy to high precision for all the forms of matter represented by T, we have $\nabla^a T_{ab} = 0$ as well. Applying the product rule to the term $\nabla^a(\Lambda g_{ab})$, we get $g_{ab}\nabla^a\Lambda + \Lambda\nabla^a g_{ab}$. But the covariant derivative of the metric vanishes, so the result is simply $\nabla_b\Lambda$. Thus any variation in the cosmological constant over space or time violates the field equations, and the violation is equivalent to the violation we would get from a form of matter than didn't conserve mass-energy locally.

Example 13: Cosmological constant is cosmological

The addition of the Λ term constitutes a change to the *vacuum* field equations, and the good agreement between theory and experiment in the case of, e.g., Mercury's orbit puts an upper limit on Λ then implies that Λ must be small. For an order-of-magnitude estimate, consider that Λ has units of mass density, and the only parameters with units that appear in the description of Mercury's orbit are the mass of the sun, m, and the radius of Mercury's orbit, r. The relativistic corrections to Mercury's orbit are on the order of v^2, or about 10^{-8}, and they come out right. Therefore we can estimate that the cosmological constant could not have been greater than about $(10^{-8})\frac{m}{r^3} \sim 10^{-10}$ kg/m^3, or it would have caused noticeable discrepancies. This is a very poor bound; if Λ was this big, we might even be able to detect its effects in laboratory experiments. Looking at the role played by r in the estimate, we see that the upper bound could have been made tighter by increasing r. Observations on galactic scales, for example, constrain it much more tightly. This justifies the description of Λ as cosmological: the larger the scale, the more significant the effect of a nonzero Λ would be.

Example 14: Energy conditions

Since the right-hand side of the field equation is $8\pi T_{ab} + \Lambda g_{ab}$, it is possible to consider the cosmological constant as a type of matter contributing to the stress-energy tensor. We then have $\rho = -P = \frac{\Lambda}{8\pi}$. As described in more detail in section 8.2, we now know that Λ is positive. With $\Lambda > 0$, the weak and dominant energy conditions are both satisfied, so that in every frame of reference, ρ is positive and there is no flux of energy flowing at speeds greater than c. The negative pressure does violate the strong energy condition, meaning that the constant acts as a form of gravitational repulsion. If the cosmological constant is a product of the quantum-mechanical structure of the vacuum, then this violation is not too surprising, because quantum fields are known to violate various energy conditions. For example, the energy density between two parallel conducting plates is negative due to the Casimir effect.

References

[10] Proc. Royal Soc. London A 209 (1951) 248. The relevant result is summarized in Misner, Thorne, and Wheeler, *Gravitation*, p. 1121.

[11] arxiv.org/abs/gr-qc/0309074v1

[12] Barcelo and Visser, "Twilight for the energy conditions?," http://arxiv.org/abs/gr-qc/0205066v1.

8.7: Cosmological Solutions (Part 4)

The Radiation-dominated Solution

For the reasons discussed in example 14, the early universe was dominated by radiation. The solution of the Friedmann equations for this case is taken up in problem 11.

Local Effects of Expansion

In this section we discuss the predictions of general relativity concerning the effect of cosmological expansion on small, gravitationally bound systems such as the solar system or clusters of galaxies. The short answer is that in most realistic cosmologies (but not necessarily in "Big Rip" scenarios) the effect of expansion is not zero, but is many orders of magnitude too small to measure. Many readers will probably be willing to accept these assertions while skipping the following demonstrations.

To begin with, we observe that there are two qualitatively distinct types of effects that could exist. Suppose that a loaf of raisin bread is rising. Let's say that the loaf's scale factor a doubles by the time the yeast's efforts are spent. By definition, this means that the raisins (galaxies, test particles) get farther apart by a factor of 2. We could imagine that in addition: (1) the strain of expansion could cause each raisin to puff up by, say, 1%, and to maintain this increased size over the entire course of expansion; or that (2) expansion could could cause each raisin to expand gradually, to 0.2% more than its original size, then 0.4% more than its original size, and so on, until, at the end of the process, each had grown beyond its original size by some amount such as 3.8%, which, while less than the 100% growth of the inter-raisin distances, was nevertheless nonzero. Astronomers refer to the second possibility as a "secular" trend. For example, simulations of solar systems often show that over billions of years, planets gradually migrate either inward or outward, under the influence of their gravitational interactions with other planets. As an example of an expansion without a secular trend, asteroids may experience a nonnegligible $\frac{1}{r^2}$ force due to radiation pressure from the sun. The effect is exactly as if the sun's mass or the gravitational constant had been slightly reduced. Kepler's elliptical orbit law holds, the law of periods is slightly off, and the orbital radius shows zero trend over time.

If either type of effect exists, an observer in some local inertial frame will interpret it as a "force." (The scare quotes are a reminder that general relativity doesn't describe gravity as Newton-style linearly additive, instantaneous action at a distance.) Such a force, if it exists, cannot simply be proportional to the rate of expansion $\frac{\dot{a}}{a}$. As a counterexample, the Milne universe is just flat spacetime described in silly coordinates, and it has $\dot{a} \neq 0$.

It would make more sense for the force to depend on the second derivative of the scale factor. To justify this more precisely, imagine releasing two test particles, initially separated by some distance that is much less than the Hubble scale. They are initially at rest relative to the Hubble flow, and no locally gravitating bodies are present. As discussed in example 15, the acceleration of one test particle relative to the other is given by $(\frac{\ddot{a}}{a})\mathbf{r}$, where $\mathbf{r}$ is their relative displacement.

Thus if we are to observe any nonzero effects of expansion on a local system, they are not really effects of expansion at all, but effects of the *acceleration* of expansion. The factor $\frac{\ddot{a}}{a}$ is on the order of the inverse square of the age of the universe, i.e., $H^2{}_o \sim 10^{-35} \text{ s}^{-2}$. The smallness of this factor is what makes the effect on a system such as the solar system so absurdly tiny.

> ### Example 24: A human body
>
> Let's estimate the effect of cosmological expansion on the length L of your thigh bone. The body is made of atoms, and for the reasons given earlier, there can be no steady trend in the sizes of these atoms or the lengths of the chemical bonds between them. The bone experiences a stress due to cosmological expansion, but it is in equilibrium, and the strain will disappear if the gravitational stress is removed (e.g., if other gravitational stresses are superimposed on top of the cosmological one in order to cancel it). The anomalous acceleration between the ends of the bone is $(\frac{\ddot{a}}{a})L$, which is observed as an anomalous stress. Taking $\frac{\ddot{a}}{a} \sim H_o^2$, the anomalous acceleration of one end of the bone relative to the other is $\sim LH^2$. The corresponding compression or tension is $\sim mLH^2$, where m is your body mass. The resulting strain is $\epsilon \sim \frac{mLH^2}{AE}$, where E is the Young's modulus of bone (about 10^{10} Pa) and A is the bone's cross-sectional area.
>
> Putting in numbers, the result for the strain is about 10^{-40}, which is much too small to be measurable by any imaginable technique, and would in reality be swamped by other effects. Since the sign of $\ddot{a}$ is currently positive, this strain is tensile, not compressive. In the earlier, matter-dominated era of the universe, it would have been compressive.

There is no "secular trend," i.e., your leg bone is not expanding over time. It's in equilibrium, and is simply elongated imperceptibly compared to the length if would have had without the effect of cosmological expanson.

Example 25: strain on an atomic nucleus

The estimate in example 24 can also be applied to an atomic nucleus, which has a nuclear "Young's modulus" on the order of 1 MeV/fm$^3 \sim 10^{32}$ Pa. The result is a strain $\epsilon \sim 10^{-52}$.

Example 26: a big rip

Known forms of matter are believed to have equations of state P = wρ with w ≥ −1. The value for a vacuum-dominated universe would be w = −1. Cosmological observations[23] show that empirically the present-day universe behaves as if it is made out of stuff with w = −1.03 ± .16, and this leaves open the possibility of w < −1. In this case, the solution to the Friedmann equations gives a scale factor a(t) that blows up to infinity at some finite t. In such a scenario, known as a "Big Rip," $\left(\frac{d}{dt}\right)\left(\frac{\ddot{a}}{a}\right)$ diverges, and any system, no matter how tightly bound, is ripped apart.[24]

---

[23] Carnero et al., arxiv.org/abs/1104.5426

[24] Caldwell et al., arxiv.org/abs/astro-ph/0302506

Examples 24-26 show that except under hypothetical extreme cosmological conditions, there is no hope of detecting any effect of cosmological expansion on systems made of condensed matter. We need to look at much larger systems to see any effect, and such systems are held together by gravity. For concreteness, let's keep talking about the earth-sun system. Not only is the anomalous force on the earth small, it is not guaranteed to produce any secular trend, which is what would be most likely to be detectable. The direction of the anomalous force on the earth is outward for an accelerating cosmological expansion, as we now know is the case for the present epoch. As an example in which no secular trend occurs, a vacuum-dominated cosmology gives a constant value for $\frac{\ddot{a}}{a}$, so the outward force is constant. As with the effect of radiation pressure, the existence of this constant, outward force is very nearly equivalent to rescaling the sun's gravitational force by a tiny amount, so the motion is still very nearly Keplerian, but with a slightly "wrong" constant of proportionality in Kepler's law of periods. The rate of change $\dot{r}$ in the radius of the circular orbit is therefore zero in this case.

But in most cosmologies $\frac{\ddot{a}}{a}$ is not exactly constant, and the anomalous force on the earth varies. In a matter-dominated cosmology with $\Lambda = 0$, in its expanding phase, the force is inward but decreasing over time, so the orbit expands over time. What really matters then, is $\left(\frac{d}{dt}\right)\left(\frac{\ddot{a}}{a}\right)$. If we were free to pick any function for a(t), we could make up examples in which $\dot{a} > 0$ but $\left(\frac{d}{dt}\right)\left(\frac{\ddot{a}}{a}\right) < 0$, so that the solar system would respond to cosmological expansion by shrinking!

The function a(t), however, has to satisfy the Friedmann equations, one of which is (in units with G ≠ 1)

$$\frac{\ddot{a}}{a} = G\left[\frac{1}{3}\Lambda - \frac{4\pi}{3}(\rho + 3P)\right]. \qquad (8.2.18)$$

The present epoch of the universe seems to be well modeled by dark energy described by a constant Λ plus dust with P << ρ. Differentiating both sides with respect to time gives

$$\frac{d}{dt}\left(\frac{\ddot{a}}{a}\right) \propto \dot{\rho}, \qquad (8.2.19)$$

with a negative constant of proportionality. This ensures that the sign of the effect is always as expected from the naive Manichean image of binding forces struggling against cosmological expansion (or perhaps cooperating during the contracting phase of a Big Crunch cosmology).

One way of understanding why this reduces so nicely to a dependence on $\dot{\rho}$ is the reasoning given in example 15, in which we found that the relative acceleration of two test particles A and B in a matter-dominated FRW cosmology could be calculated accurately by pretending that it was due to the presence of the dust in any given sphere S surrounding the two particles. We now let A be the sun, B the earth, and S a sphere centered on the sun whose radius equals the radius of the earth's circular orbit. Due to cosmological expansion, the dust inside S thins out with time, reducing its density ρ. Applying Newton's laws to the orbit of the earth gives $\omega^2 r = \frac{GM}{r^2}$, and conservation of angular momentum results in ωr^2 = const. A calculation gives

$$\frac{r}{r_o} = \frac{M + (\frac{4\pi}{3})\rho_o r_o^3}{M + (\frac{4\pi}{3})\rho r^3} \tag{8.2.20}$$

, which results in $\frac{\dot{r}}{r_o} \approx -(\frac{4\pi}{3})G\omega_o^{-2}\dot{\rho}$. Application of the Friedmann equations yields

$$\frac{\dot{r}}{r_o} = \omega_o^{-2}\left(\frac{d}{dt}\right)\left(\frac{\ddot{a}}{a}\right), \tag{8.2.21}$$

which is valid generally, not just for P = 0. The ω_o^{-2} factor shows that the effect is smaller for more tightly bound systems.

We know that the universe in the present era has $\left(\frac{d}{dt}\right)\left(\frac{\ddot{a}}{a}\right) > 0$ because $\dot{\rho} < 0$, and for purposes of an order-of-magnitude estimate we can take $\left(\frac{d}{dt}\right)\left(\frac{\ddot{a}}{a}\right) \sim H_o^3$. Plugging in numbers for the earth-sun system, we find that since the age of the dinosaurs, the radius of the earth's orbit has grown by less than the diameter of an atomic nucleus.[25]

> **Note**
>
> The picturesque image comes from Cooperstock et al., http://arxiv.org/abs/astro-ph/9803097v1, who give a different calculation leading to a result for $\dot{r}$ exactly equivalent to the one derived here.

Observations

Historically, it was believed that the cosmological constant was zero, that nearly all matter in the universe was in the form of atoms, and that there was therefore only one interesting cosmological parameter to measure, which was the average density of matter. This density was very difficult to determine, even to within an order of magnitude, because most of the matter in the universe probably doesn't emit light, making it difficult to detect. Astronomical distance scales were also very poorly calibrated against absolute units such as the SI. Starting around 1995, however, a new set of techniques led to an era of high-precision cosmology.

Spatial Curvature from CMB Fluctuations

A strong constraint on the models comes from accurate measurements of the cosmic microwave background, especially by the 1989-1993 COBE probe, and its 2001-2009 successor, the Wilkinson Microwave Anisotropy Probe, positioned at the L2 Lagrange point of the earth-sun system, beyond the Earth on the line connecting sun and earth.[26] The temperature of the cosmic microwave background radiation is not the same in all directions, and it can be measured at different angles. In a universe with negative spatial curvature, the sum of the interior angles of a triangle is less than the Euclidean value of 180 degrees. Therefore if we observe a variation in the CMB over some angle, the distance between two points on the surface of last scattering is actually greater than would have been inferred from Euclidean geometry. The distance scale of such variations is limited by the speed of sound in the early universe, so one can work backward and infer the universe's spatial curvature based on the angular scale of the anisotropies. The measurements of spatial curvature are usually stated in terms of the parameter Ω, defined as the total average density of all source terms in the Einstein field equations, divided by the critical density that results in a flat universe. Ω includes contributions from matter, Ω_M, the cosmological constant, Ω_Λ, and radiation (negligible in the present-day universe). The results from WMAP, combined with other data from other methods, gives $\Omega = 1.005 \pm .006$. In other words, the universe is very nearly spatially flat.
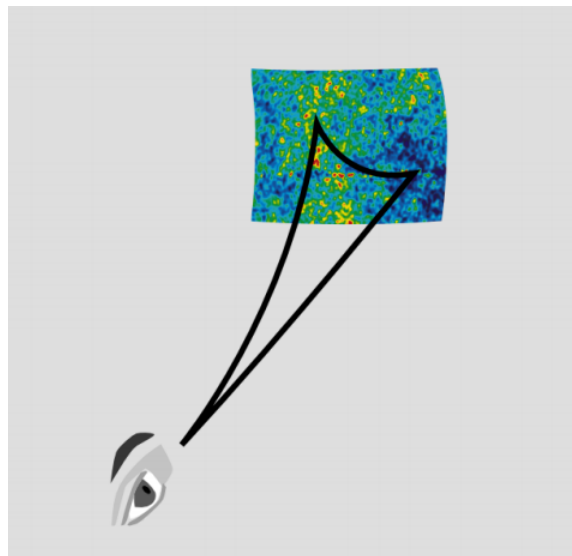
Figure 8.7.6 - The angular scale of fluctuations in the cosmic microwave background can be used to infer the curvature of the universe.

Accelerating Expansion from Supernova Data

The supernova data described in section 8.2 complement the CMB data because they are mainly sensitive to the difference $\Omega_\Lambda - \Omega_M$, rather than their sum $\Omega = \Omega_\Lambda + \Omega_M$. This is because these data measure the acceleration or deceleration of the universe's expansion. Matter produces deceleration, while the cosmological constant gives acceleration. Figure 8.2.7 shows some recent supernova data.[27] The horizontal axis gives the redshift factor z = $\frac{\lambda' - \lambda}{\lambda}$, where λ' is the wavelength observed on earth and λ the wavelength originally emitted. It measures how fast the supernova's galaxy is receding from us. The vertical axis is Δ(m−M) = (m−M)−(m−M)$_{empty}$, where m is the apparent magnitude, M is the absolute magnitude, and (m − M)$_{empty}$ is the value expected in a model of an empty universe, with Ω = 0. The difference m−M is a measure of distance, so essentially this is a graph of distance versus recessional velocity, of the same general type used by Hubble in his original discovery of the expansion of the universe. Subtracting (m−M)$_{empty}$ on the vertical axis makes it easier to see small differences. Since the WMAP data require Ω = 1, we need to fit the supernova data with values of Ω_M and Ω_Λ that add up to one. Attempting to do so with Ω_M = 1 and Ω_Λ = 0 is clearly inconsistent with the data, so we can conclude that the cosmological constant is definitely positive.
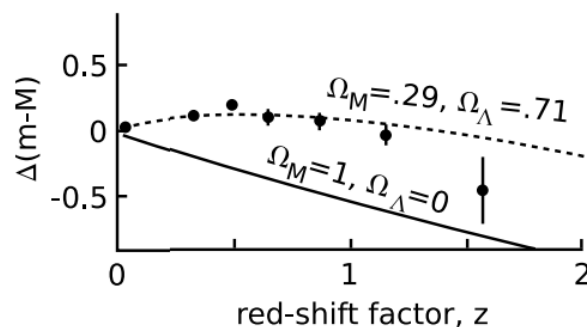
Figure 8.7.7 - A Hubble plot for distant supernovae. Each data point represents an average over several different supernovae with nearly the same z.

Density of Matter from Baryonic Acoustic Oscillations

Efforts such as the Sloan Digital Sky Survey have made threedimensional maps of the density of luminous matter in the universe.[28] The distribution is clumpy. Measuring the average correlation ξ between the density at points separated by some distance s (measured in the comoving frame), one would expect that the function ξ(s) would be largest when s was small and would simply taper off with increasing s. By analogy, we don't usually find a Manhattan-style landscape of skyscrapers side by side with an uninhabited mountainous wilderness. On the other hand, imagine constructing such a correlation function for houses in a subdivision in which the roads do not form any regular grid, but zoning regulations prohibit construction of houses on lots of less than a certain size. In this situation, there would be zero probability of finding houses separated by very small distances, and ξ(s)

would exhibit a peak at some larger scale set by the legal code. The actual results of the sky surveys do show such a peak, which is due to well known physics referred to as baryon acoustic oscillations (BAO).[29] In the early universe, any region of overdensity would tend to create a radiating sound wave like the bang of a firecracker. Such waves propagated at a known speed (about half the speed of light) for a known time (about 400,000 years, until matter became deionized and transparent to radiation, making it immune to the photon pressure that drove the oscillations). This leads to a known distance s, which forms a standard ruler at which the peak in $\xi(s)$ occurs. In cosmological models, these results strongly constrain Ω_M, while being relatively insensitive to Ω_Λ, and they are therefore complementary to both the supernova data and the CMB results.
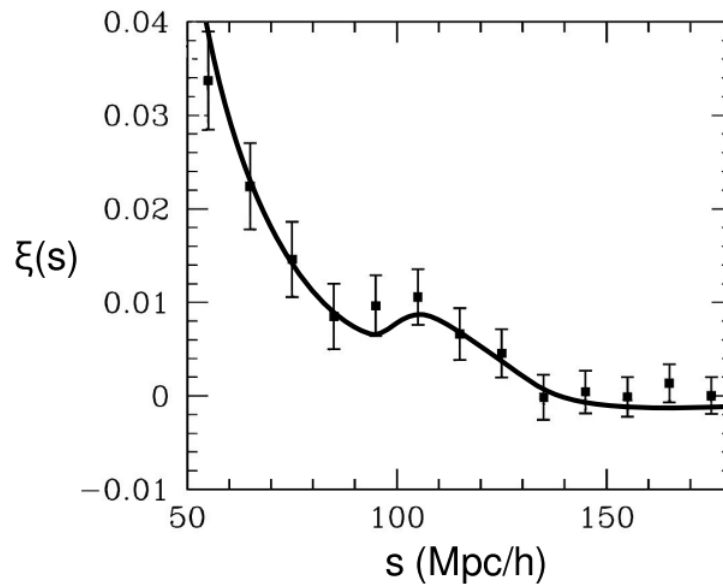
Figure 8.7.8 - The acoustic peak in the BAO correlation function. Redrawn from Eisenstein, New Astronomy Reviews. 49 (2005) 360, as reproduced in Bassett and Hlozek, 2009, arxiv.org/abs/0910.5224.

Conclusions about Cosmology

Figure 8.2.9 summarizes what we can conclude about our universe, parametrized in terms of a model with both Ω_M and Ω_Λ nonzero.[30] We can tell that it originated in a Big Bang singularity, that it will go on expanding forever, and that it is very nearly flat. Note that in a cosmology with nonzero values for both Ω_M and Ω_Λ, there is no strict linkage between the spatial curvature and the question of recollapse, as there is in a model with only matter and no cosmological constant; therefore even though we know that the universe will not recollapse, we do not know whether its spatial curvature is slightly positive (closed) or negative (open).
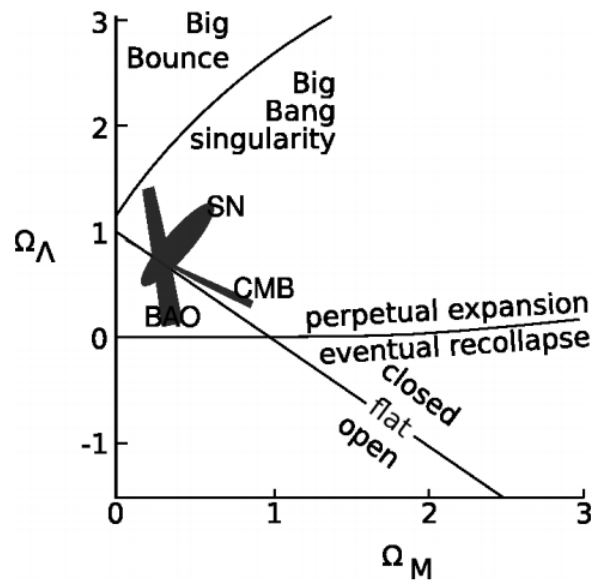
Figure 8.7.9 - The cosmological parameters of our universe, after Perlmutter, 1998, arxiv.org/abs/astro-ph/9812133 and Kowalski, 2008, arxiv.org/abs/0804.4142. The three shaded regions represent the 95% confidence regions for the three types of observations.

Consistency Checks

Astrophysical considerations provide further constraints and consistency checks. In the era before the advent of high-precision cosmology, estimates of the age of the universe ranged from 10 billion to 20 billion years, and the low end was inconsistent with the age of the oldest globular clusters. This was believed to be a problem either for observational cosmology or for the astrophysical models used to estimate the age of the clusters: "You can't be older than your ma." Current data have shown that the low estimates of the age were incorrect, so consistency is restored.

That only a small fraction of the universe's matter was luminous had been suspected by astronomers such as Zwicky as early as 1933, based on the inability to reconcile the observed kinematics with Newton's laws if all matter was assumed to be luminous.

Dark Matter

Another constraint comes from models of nucleosynthesis during the era shortly after the Big Bang (before the formation of the first stars). The observed relative abundances of hydrogen, helium, and deuterium cannot be reconciled with the density of "dust" (i.e., nonrelativistic matter) inferred from the observational data. If the inferred mass density were entirely due to normal "baryonic" matter (i.e., matter whose mass consisted mostly of protons and neutrons), then nuclear reactions in the dense early universe should have proceeded relatively efficiently, leading to a much higher ratio of helium to hydrogen, and a much lower abundance of deuterium. The conclusion is that most of the matter in the universe must be made of an unknown type of exotic non-baryonic matter, known generically as "dark matter."

The existence of nonbaryonic matter is also required in order to reconcile the observed density of galaxies with the observed strength of the CMB fluctuations, and in merging galaxy clusters it has been observed that the gravitational potential is offset from the radiating plasma. A 2012 review paper on dark matter is Roos, arxiv.org/abs/1208.3662.

A number of experiments are under way to detect dark matter directly. As of 2013, the most sensitive experiment has given null results: arxiv.org/abs/1310.8214.

At one time it was widely expected that dark matter would consist of the lightest supersymmetric particle, which might for example be the neutralino. However, results from the LHC seem to make it unlikely that our universe exhibits supersymmetry, assuming that the energy scale is the electroweak scale, which is the only scale that has strong motivation. It now appears more likely that dark matter consists of some other particle such as sterile neutrinos or axions. Even with the inclusion of dark matter, there is a problem with the abundance of lithium-7 relative to hydrogen, which models greatly overpredict.[31]

References

[26] Komatsu et al., 2010, arxiv.org/abs/1001.4538

[27] Riess et al., 2007, arxiv.org/abs/astro-ph/0611572. A larger data set is analyzed in Kowalski et al., 2008, arxiv.org/abs/0804.4142.

[28] Sanchez et al., 2012, arxiv.org/abs/1203.6616

[29] Bassett and Hlozek, 2009, arxiv.org/abs/0910.5224

[30] See Carroll, "The Cosmological Constant," http://www.livingreviews.org/lrr-2001-1 for a full mathematical treatment of such models.

[31] arxiv.org/abs/0808.2818, arxiv.org/abs/1107.1117

8.8: Mach's Principle Revisited

The Brans-Dicke Theory

Mach himself never succeeded in stating his ideas in the form of a precisely testable physical theory, and we've seen that to the extent that Einstein's hopes and intuition had been formed by Mach's ideas, he often felt that his own theory of gravity came up short. The reader has so far encountered Mach's principle in the context of certain thought experiments that are obviously impossible to realize, involving a hypothetical universe that is empty except for certain apparatus (e.g., section 3.6). It would be easy, then, to get an impression of Mach's principle as one of those theories that is "not even wrong," i.e., so ill-defined that it cannot even be falsified by experiment, any more than Christianity can be.

But in 1961, Robert Dicke and his student Carl Brans came up with a theory of gravity that made testable predictions, and that was specifically designed to be more Machian than general relativity. Their paper[32] is extremely readable, even for the non-specialist. In this theory, the seemingly foolproof operational definition of a Lorentz frame given in section 1.5 fails. On the first page, Brans and Dicke propose one of those seemingly foolish thought experiments about a nearly empty universe:

> *The imperfect expression of [Mach's ideas] in general relativity can be seen by considering the case of a space empty except for a lone experimenter in his laboratory. [...] The observer would, according to general relativity, observe normal behavior of his apparatus in accordance with the usual laws of physics. However, also according to general relativity, the experimenter could set his laboratory rotating by leaning out a window and firing his 22-caliber rifle tangentially. Thereafter the delicate gyroscope in the laboratory would continue to point in a direction nearly fixed relative to the direction of motion of the rapidly receding bullet. The gyroscope would rotate relative to the walls of the laboratory. Thus, from the point of view of Mach, the tiny, almost massless, very distant bullet seems to be more important than the massive, nearby walls of the laboratory in determining inertial coordinate frames and the orientation of the gyroscope.*

They then proceed to construct a mathematical and more Machian theory of gravity. From the Machian point of view, the correct local definition of an inertial frame must be determined relative to the bulk of the matter in the universe. We want to retain the Lorentzian local character of spacetime, so this influence can't be transmitted via instantaneous action at a distance. It must propagate via some physical field, at a speed less than or equal to c. It is implausible that this field would be the gravitational field as described by general relativity. Suppose we divide the cosmos up into a series of concentric spherical shells centered on our galaxy. In Newtonian mechanics, the gravitational field obeys Gauss's law, so the field of such a shell vanishes identically on the interior. In relativity, the corresponding statement is Birkhoff's theorem, which states that the Schwarzschild metric is the unique spherically symmetric solution to the vacuum field equations. Given this solution in the exterior universe, we can set a boundary condition at the outside surface of the shell, use the Einstein field equations to extend the solution through it, and find a unique solution on the interior, which is simply a flat space.

Since the Machian effect can't be carried by the gravitational field, Brans and Dicke took up an idea earlier proposed by Pascual Jordan[33] of hypothesizing an auxiliary field ϕ. The fact that such a field has never been detected directly suggests that it has no mass or charge. If it is massless, it must propagate at exactly c, and this also makes sense because if it were to propagate at speeds less than c, there would be no obvious physical parameter that would determine that speed. How many tensor indices should it have? Since Mach's principle tries to give an account of inertia, and inertial mass is a scalar,[34] ϕ should presumably be a scalar (quantized by a spin-zero particle). Theories of this type are called tensor-scalar theories, because they use a scalar field in addition to the metric tensor.

Notes

[33] Jordan was a member of the Nazi *Sturmabteilung* or "brown shirts" who nevertheless ran afoul of the Nazis for his close professional relationships with Jews.

[34] A limit of 5×10^{-23} has been placed on the anisotropy of the inertial mass of the proton: R.W.P. Drever, "A search for anisotropy of inertial mass using a free precession technique," Philosophical Magazine, 6:687 (1961) 683.

The wave equation for a massless scalar field, in the absence of sources, is simply $\nabla_i \nabla^i_{\phi} = 0$. The solutions of this wave equation fall off as $\phi \sim \frac{1}{r}$. This is gentler than the $\frac{1}{r^2}$ variation of the gravitational field, so results like Newton's shell theorem and Birkhoff's theorem no longer apply. If a spherical shell of mass acts as a source of ϕ, then ϕ can be nonzero and varying inside the shell. The ϕ that you experience right now as you read this book should be a sum of wavelets originating from all the masses whose world-lines intersected the surface of your past light-cone. In a static universe, this sum would diverge linearly, so a self-consistency requirement for Brans-Dicke gravity is that it should produce cosmological solutions that avoid such a divergence, e.g., ones that begin with Big Bangs.

Masses are the sources of the field ϕ. How should they couple to it? Since ϕ is a scalar, we need to construct a scalar as its source, and the only reasonable scalar that can play this role is the trace of the stress-energy tensor, T^i_i. As discussed in example 11, this vanishes for light, so the only sources of ϕ are material particles.[35] Even so, the Brans-Dicke theory retains a form of the equivalence principle. As discussed in section 1.5 and problem 7, the equivalence principle is a statement about the results of local experiments, and ϕ at any given location in the universe is dominated by contributions from matter lying at cosmological distances. Objects of different composition will have differing fractions of their mass that arise from internal electromagnetic fields. Two such objects will still follow identical geodesics, since their own effect on the local value of ϕ is negligible. This is unlike the behavior of electrically charged objects, which experience significant back-reaction effects in curved space (problem 7). However, the strongest form of the equivalence principle requires that all experiments in free-falling laboratories produce identical results, no matter where and when they are carried out. Brans-Dicke gravity violates this, because such experiments could detect differences between the value of ϕ at different locations — but of course this is part and parcel of the purpose of the theory.

> **Note**
>
> This leads to an exception to the statement above that all Brans-Dicke spacetimes are expected to look like Big Bang cosmologies. Any solution of the GR field equations containing nothing but vacuum and electromagnetic fields (known as an "elevtrovac" solution) is also a valid Brans-Dicke spacetime. In such a spacetime, a constant ϕ can be set arbitrarily. Such a spacetime is in some sense not generic for Brans-Dicke gravity.

We now need to see how to connect ϕ to the local notion of inertia so as to produce an effect of the kind that would tend to fulfill Mach's principle. In Mach's original formulation, this would entail some kind of local rescaling of all inertial masses, but Brans and Dicke point out that in a theory of gravity, this is equivalent to scaling the Newtonian gravitational constant G *down* by the same factor. The latter turns out to be a better approach. For one thing, it has a natural interpretation in terms of units. Since ϕ's amplitude falls off as $\frac{1}{r}$, we can write $\phi \sim \frac{\Sigma m_i}{r}$, where the sum is over the past light cone. If we then make the identification of ϕ with $\frac{1}{G}$ (or $\frac{c^2}{G}$ in a system where c $\neq$ 1), the units work out properly, and the coupling constant between matter and ϕ can be unitless. If this coupling constant, notated $\frac{1}{\omega}$, were not unitless, then the theory's predictive value would be weakened, because there would be no way to know what value to pick for it. For a unitless constant, however, there is a reasonable way to guess what it should be: "in any sensible theory," Brans and Dicke write, "ω must be of the general order of magnitude of unity." This is, of course, assuming that the Brans-Dicke theory was correct. In general, there are other reasonable values to pick for a unitless number, including zero and infinity. The limit of $\omega \to \infty$ recovers the special case of general relativity. Thus Mach's principle, which once seemed too vague to be empirically falsifiable, comes down to measuring a specific number, ω, which quantifies how non-Machian our universe is.[36]

> **Note**
>
> Another good technical reasons for thinking of ϕ as relating to the gravitational constant is that general relativity has a standard prescription for describing fields on a background of curved spacetime. The vacuum field equations of general relativity can be derived from the principle of least action, and although the details are beyond the scope of this book (see, e.g., Wald, *General Relativity*, appendix E), the general idea is that we define a Lagrangian density LG that depends on the Ricci scalar curvature, and then extremize its integral over all possible histories of the evolution of the gravitational field. If we want to describe some other field, such as matter, light, or ϕ, we simply take the special-relativistic Lagrangian $\mathcal{L}_M$ for that field, change all the derivatives to covariant derivatives, and form the sum $(\frac{1}{G})\mathcal{L}_G + \mathcal{L}_M$. In the Brans-Dicke theory, we have three pieces, $(\frac{1}{G})\mathcal{L}_G + \mathcal{L}_M + \mathcal{L}_\phi$, where $\mathcal{L}_M$ is for matter and $\mathcal{L}_\phi$ for ϕ. If we were to interpret ϕ as a rescaling of inertia, then we would have to have ϕ appearing as a fudge factor modifying all the inner workings of $\mathcal{L}_M$. If, on the other hand, we think of ϕ as changing the value of the gravitational constant G, then the necessary modification is extremely simple. Brans and Dicke

> introduce one further modification to $\mathcal{L}_\phi$ so that the coupling constant ω between matter and ϕ can be unitless. This modification has no effect on the wave equation of ϕ in flat spacetime.

Predictions of the Brans-Dicke Theory

Returning to the example of the spherical shell of mass, we can see based on considerations of units that the value of ϕ inside should be $\sim \frac{m}{r}$, where m is the total mass of the shell and r is its radius. There may be a unitless factor out in front, which will depend on ω, but for $\omega \sim 1$ we expect this constant to be of order 1. Solving the nasty set of field equations that result from their Lagrangian, Brans and Dicke indeed found $\phi \approx \left[\frac{2}{3+2\omega}\right]\left(\frac{m}{r}\right)$, where the constant in square brackets is of order unity if ω is of order unity. In the limit of $\omega \to \infty$, $\phi = 0$, and the shell has no physical effect on its interior, as predicted by general relativity.

Brans and Dicke were also able to calculate cosmological models, and in a typical model with a nearly spatially flat universe, they found ϕ would vary according to

$$\phi = 8\pi \frac{4+3\omega}{6+4\omega} \rho_o t_o^2 \left(\frac{t}{t_o}\right)^{\frac{2}{4+\omega}},\tag{8.8.1}$$

where ρ_o is the density of matter in the universe at time t = t_0. When the density of matter is small, G is large, which has the same observational consequences as the disappearance of inertia; this is exactly what one expects according to Mach's principle. For $\omega \to \infty$, the gravitational "constant" G = $\frac{1}{\phi}$ really is constant.

Returning to the thought experiment involving the 22-caliber rifle fired out the window, we find that in this imaginary universe, with a very small density of matter, G should be very large. This causes a frame-dragging effect from the laboratory on the gyroscope, one much stronger than we would see in our universe. Brans and Dicke calculated this effect for a laboratory consisting of a spherical shell, and although technical difficulties prevented the reliable extrapolation of their result to $\rho_o \to 0$, the trend was that as ρ_o became small, the frame-dragging effect would get stronger and stronger, presumably eventually forcing the gyroscope to precess in lock-step with the laboratory. There would thus be no way to determine, once the bullet was far away, that the laboratory was rotating at all — in perfect agreement with Mach's principle.

Hints of Empirical Support

Only six years after the publication of the Brans-Dicke theory, Dicke himself, along with H.M. Goldenberg[37] carried out a measurement that seemed to support the theory empirically. Fifty years before, one of the first empirical tests of general relativity, which it had seemed to pass with flying colors, was the anomalous perihelion precession of Mercury. The word "anomalous," which is often left out in descriptions of this test, is required because there are many nonrelativistic reasons why Mercury's orbit precesses, including interactions with the other planets and the sun's oblate shape. It is only when these other effects are subtracted out that one sees the general-relativistic effect calculated in section 6.2. The sun's oblateness is difficult to measure optically, so the original analysis of the data had proceeded by determining the sun's rotational period by observing sunspots, and then assuming that the sun's bulge was the one found for a rotating fluid in static equilibrium. The result was an assumed oblateness of about 1×10^{-5}. But we know that the sun's dynamics are more complicated than this, since it has convection currents and magnetic fields. Dicke, who was already a renowned experimentalist, set out to determine the oblateness by direct optical measurements, and the result was $(5.0 \pm 0.7) \times 10^{-5}$, which, although still very small, was enough to put the observed perihelion precession out of agreement with general relativity by about 8%. The perihelion precession predicted by Brans-Dicke gravity differs from the general relativistic result by a factor of $\frac{4+3\omega}{6+3\omega}$. The data therefore appeared to require $\omega \approx 6 \pm 1$, which would be inconsistent with general relativity.
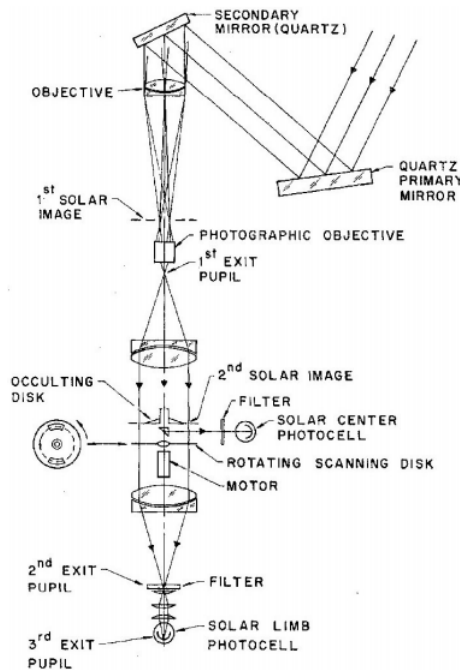
Figure 8.8.1 - The apparatus used by Dicke and Goldenberg to measure the oblateness of the sun was essentially a telescope with a disk inserted in order to black out most of the light from the sun.

Mach's Principle is False

The trouble with the solar oblateness measurements was that they were subject to a large number of possible systematic errors, and for this reason it was desirable to find a more reliable test of Brans-Dicke gravity. Not until about 1990 did a consensus arise, based on measurements of oscillations of the solar surface, that the pre-Dicke value was correct. In the interim, the confusion had the salutary effect of stimulating a renaissance of theoretical and experimental work in general relativity. Often if one doesn't have an alternative theory, one has no reasonable basis on which to design and interpret experiments to test the original theory.

Currently, the best bound on ω is based on measurements[38] of the propagation of radio signals between earth and the Cassini-Huygens space probe in 2003, which require $\omega > 4 \times 10^4$. This is so much greater than unity that it is reasonable to take Brans and Dicke at their word that "in any sensible theory, ω must be of the general order of magnitude of unity." Brans-Dicke fails this test, and is no longer a "sensible" candidate for a theory of gravity. We can now see that Mach's principle, far from being a fuzzy piece of philosophical navel-gazing, is a testable hypothesis. It has been tested and found to be false, in the following sense. Brans-Dicke gravity is about as natural a formal implementation of Mach's principle as could be hoped for, and it gives us a number ω that parametrizes how Machian the universe is. The empirical value of ω is so large that it shows our universe to be essentially as non-Machian as general relativity.

References

[32] C. Brans and R. H. Dicke, "Mach's Principle and a Relativistic Theory of Gravitation," Physical Review 124 (1961) 925

[37] Dicke and Goldenberg, "Solar Oblateness and General Relativity," Physical Review Letters 18 (1967) 313

[38] Bertotti, Iess, and Tortora, "A test of general relativity using radio links with the Cassini spacecraft," Nature 425 (2003) 374

8.9: Historical Note - The Steady-state Model

From 1948 until around the mid-1960s, the Big Bang theory had viable competition in the form of the steady-state model, originated by the British trio of Fred Hoyle, Hermann Bondi, and Thomas Gold. Legend has it that they came up with the idea after seeing a horror movie called Dead of Night, in which events from the beginning of the story repeat themselves later. This led them to imagine that the universe could, although expanding, remain locally in the same state at all times. If this were to happen, the empty space being opened up between the galaxies would have to be filled back in by the spontaneous creation of matter. The model holds a strong philosophical appeal because it generalizes the Copernican principle so that it applies not just to conditions everywhere in space but also at all times.

They published the idea in a pair of back-to-back papers, one by Bondi and Gold[39] and one by Hoyle,[40] with comments appended to the former on the differences between the two approaches. The Bondi-Gold paper is especially fun to read, because it is written in nontechnical language and shows a type of daring and creative science that is not often encountered today. Much of it reads like a catalog of cherished principles of physics that were to be given up, including Lorentz invariance, general relativity, the equivalence principle, and possibly the laws of conservation of charge and mass-energy. The following is a brief presentation (in slightly different notation) of the Hoyle's more mathematically detailed ideas, as sketched in his original paper. Although Hoyle eventually fleshed out the ideas more thoroughly, by the time he had done so the steady-state theory was already on its way to being crushed under the weight of contrary observations.

Since the model is always to be in the same state, the quantity $\frac{\dot{a}}{a}$ must always be the same, i.e., the Hubble constant really is a constant over time. This requires exponential growth, which means that the geometry is that of de Sitter space. In any model that assigns an infinite age to the universe, one must explain why the universe has not undergone heat death due to the second law of thermodynamics. The steady-state model successfully addresses this problem, because the exponential expansion is rapid enough to prevent thermal equilibrium from happening.

Hoyle sets out to preserve local conservation of energy-momentum, without which the Einstein field equations become inconsistent. (This was Hoyle's more conservative approach. Bondi and Gold advocated replacing general relativity completely rather than modifying it.) He postulates a massless, chargeless, scalar field C, called the "C field," the letter "C" standing for "creation." Suppose that the C field's contribution to the stress-energy tensor ends up being that of a perfect fluid with the same rest frame as the ordinary matter. The rate of creation of mass-energy is then given by the divergence $\nabla_\mu T_t^\mu$, and we need this to be zero. As shown in example 21, this requires that our total stress-energy mimic that of a cosmological constant, with ρ + P = 0. Since the ordinary matter has ρ > 0 and P > 0, the C field will either need to contribute negative energy density or negative pressure. We'll see below that Hoyle's model is constructed so that the C field has zero energy and negative pressure. One will often see the C field described incorrectly as having negative energy to cancel out the positive energy of the matter being created. That wouldn't have worked, because then the total energy density ρ would always be zero, which is not what we observe. (For example, the Friedmann equation for $\frac{\ddot{a}}{a}$ relates ρ to the square of the Hubble constant.)

To understand more about why the theory took the form it did, it is helpful to look at some general physical considerations about symmetry. As Bondi and Gold admit candidly, any theory of this type is likely to violate Lorentz invariance. We can observe an evacuated box and wait for hydrogen atoms to appear. When they appear, they're in some state of motion, at least on the average. This state of motion defines a preferred frame. In addition to breaking symmetry under Lorentz transformations, the theory lacks timereversal invariance (because matter appears but never disappears) and charge-conjugation symmetry (because matter appears but antimatter doesn't). All of these asymmetries arise because in this approach, we try to explain the observed asymmetries of the cosmological state of the universe as arising directly from asymmetries in the underlying local laws of physics. Such an approach is very different from the modern one, in which we expect the asymmetries to arise from either boundary conditions or instabilities (spontaneous symmetry breaking).

Because the C field is massless and chargeless, we would normally expect it to obey the wave equation $\nabla_a \nabla_C^a = 0$. Hoyle's field does not, however, evolve according to any Lorentz-invariant dynamical law. Instead it simply evolves as C = t, where t is a preferred time coordinate. In any cosmological model in which the matter fields are modeled as perfect fluids, we have a preferred time coordinate which is the proper time of an observer at rest with respect to the fluid, and in the Hoyle model we do assume that this is the time t we should use in defining C. However, Hoyle's theory is different because it gives this preferred time a role in the local laws of physics, thereby breaking Lorentz invariance.

The value of the scalar field C cannot have any directly observable effects, since then its time-evolution would distinguish one epoch of the universe from another. Instead we form the gradient $\nabla_a C$. This gives a vector, which can be interpreted as a velocity

vector defining the preferred frame of reference. An observer in this frame is at rest relative to the local cosmological fluid, observes the universe to be homogeneous, and also observes that when new atoms are created from the vacuum, they are on the average at rest. Thus $\nabla_a C$ is observable.

The contribution of the C field to the stress-energy must be a rank-two tensor, and if we want to construct such a tensor, the only good possibility that occurs to me[41] is $k\nabla_a\nabla_b C$, with k a positive constant. If the derivatives had been ordinary partial derivatives, the second derivative would have vanished because C is linear in time, but the covariant derivatives do not vanish, and in fact the second derivative is a tensor measuring the rate of cosmological expansion; the trace $\nabla^a\nabla_a C$ is the volume expansion Θ defined in section 8.1. For de Sitter space, Θ has a constant value equal to three times the Hubble constant H_0. We can now see why we could not take the C field to evolve according to the usual wave equation $\nabla_a\nabla^a C = 0$; if it did, then we would have $\Theta = 0$, and the universe would not be expanding.

> **Note**
>
> The only other obvious possibility would have been something like $-k\nabla_a C\nabla_b C$. This would be the stress-energy of a negative-mass dust, which would be unacceptable for the reasons discussed earlier.

When we evaluate the second derivative for the de Sitter metric, the only nonvanishing Christoffel symbols that occur are

$$\Gamma^t_{xx} = \Gamma^t_{yy} = \Gamma^t_{zz} = \dot{a}a. \tag{8.9.1}$$

We find

$$T^t_t = 0 \tag{8.9.2}$$

and

$$T^x_x = T^y_y = T^z_z = \frac{k\Theta}{3}. \tag{8.9.3}$$

Thus the C field's mass-energy density is zero, while for its pressure we have $P = -\frac{k\Theta}{3}$, which is negative.

For simplicity, we take the ordinary matter to be dust. The total stress-energy then consists of an energy density ρ that is due only to the dust, and a negative pressure P that comes only from the C field. If we require both of these to be constant, take the cosmological constant to be zero, set

$$a = e^{H_0 t} = e^{\Theta t/3} \tag{8.9.4}$$

and substitute into the Friedmann equations from Section 8.2, we find P = $-\rho$, or

$$k = \frac{3\rho}{\Theta} = \frac{\rho}{H_0}. \tag{8.9.5}$$

Like the cosmological constant, the C field is taken to be a universally prescribed property of the vacuum. There is a difference, however, because the cosmological constant's contribution to the stress-energy is proportional to the metric, which preserves the equivalence principle. As remarked at the end of the Bondi-Gold paper, the C field violates the equivalence principle. No calculation is spelled out, but they say based on a personal communication from Hoyle that the field exerts a force on matter which produces a significant acceleration in an atom, but a negligible one in a star.

A claimed selling point of the C field was that it would prevent the formation of singularities, including both a Big Bang singularity and black holes. This is reasonable, since the C field violates all of the energy conditions listed in section 8.1 except for the trace energy condition. The Penrose singularity theorem depends on the null energy condition, and the Hawking singularity theorem requires either the strong or the null energy condition.

Because we can't make the C field obey the proper wave equation for a massless, spin-zero particle, there is no obvious way to make up a dynamical law for its evolution in order to replace the fixed relation C = t, and we do not expect to have any classical field theory for the C field. Hoyle did attempt to add dynamics to the model by making it into what is known as a "direct field," which was a type of action-at-a-distance theory that in the 1960s was believed to be a good candidate for the fundamental description of the forces of physics. (Quantum field theory had not developed to the point where it could handle the strong or weak nuclear forces.) Such theories were shown to be nonviable as quantum theories in 1963 by Currie, Jordan, and Sudarshan.

The steady-state model began to succumb to contrary evidence when Ryle and coworkers counted radio sources and found that they did not show the statistical behavior predicted by the model. The coup de grace came with the discovery of the cosmic microwave background, which demonstrated directly that the universe had once been much hotter than it is now. Attempts have been made to produce variations on the model that are consistent with these observations, but they have not succeeded; for a detailed discussion see http://www.astro.ucla.edu/~wright/stdystat.htm.

References

39 "The Steady-State Theory of the Expanding Universe," MNRAS 108 (1948) 252; adsabs.harvard.edu/cgi-bin/nph-bib_query?bibcode=1948MNRAS.108.252B

40 "A New Model for the Expanding Universe," MNRAS 108 (1948) 372, adsabs.harvard.edu/cgi-bin/nph-bib_query?bibcode=1948MNRAS.108.372H

8.E: Sources in General Relativity (Exercise)

1. Verify, as claimed in section 8.1, that the electromagnetic pressure inside a medium-weight atomic nucleus is on the order of 10^{33} Pa.

2. Is the Big Bang singularity removable by the coordinate transformation $t \to \frac{1}{t}$?

3. Verify the claim made in section 8.2 that a is a linear function of time in the case of the Milne universe, and that k = −1.

4. Examples 16 and 18 discussed ropes with cosmological lengths. Reexamine these examples in the case of the Milne universe.

5. (a) Show that the Friedmann equations are symmetric under time reversal. (b) The spontaneous breaking of this symmetry in perpetually expanding solutions was discussed in section 8.2. Use the definition of a manifold to show that this symmetry cannot be restored by gluing together an expanding solution and a contracting one "back to back" to create a single solution on a single, connected manifold.

6. The Einstein field equations are

$$G_{ab} = 8\pi T_{ab} + \Lambda g_{ab}, \tag{8.E.1}$$

and when it is possible to adopt a frame of reference in which the local mass-energy is at rest on average, we can interpret the stress-energy tensor as

$$T_\nu^\mu = diag(-\rho, P, P, P), \tag{8.E.2}$$

where ρ is the mass-energy density and P is the pressure. Fix some point as the origin of a local Lorentzian coordinate system. Analyze the properties of these relations under a reflection such as x → −x or t → −t.

7. (a) Show that a positive cosmological constant violates the strong energy condition in a vacuum. In applying the definition of the strong energy condition, treat the cosmological constant as a form of matter, i.e., "roll in" the cosmological constant term to the stress-energy term in the field equations. (b) Comment on how this affects the results of the following paper: Hawking and Ellis, "The Cosmic Black-Body Radiation and the Existence of Singularities in Our Universe," Astrophysical Journal, 152 (1968) 25, articles.adsabs.harvard.edu/f...pJ...152...25H.

8. In problem 7 in chapter 5, we analyzed the properties of the metric

$$ds^2 = e^{2gz}dt^2 - dz^2. \tag{8.E.3}$$

   a. In that problem we found that this metric had the same properties at all points in space. Verify in particular that it has the same scalar curvature R at all points in space.
   b. Show that this is a vacuum solution in the two-dimensional (t, z) space.
   c. Suppose we try to generalize this metric to four dimensions as

$$ds^2 = e^{2gz}dt^2 - dx^2 - dy^2 - dz^2. \tag{8.E.4}$$

   Show that this requires an Einstein tensor with unphysical properties.

9. Consider the following proposal for defeating relativity's prohibition on velocities greater than c. Suppose we make a chain billions of light-years long and attach one end of the chain to a particular galaxy. At its other end, the chain is free, and it sweeps past the local galaxies at a very high speed. This speed is proportional to the length of the chain, so by making the chain long enough, we can make the speed exceed c.
   Debunk this proposal in the special case of the Milne universe.

10. Make a rigorous definition of the volume V of the *observable* universe. Suppose someone asks whether V depends on the observer's state of motion. Does this question have a well-defined answer? If so, what is it? Can we calculate V 's observer-dependence by applying a Lorentz contraction?

11. For a perfect fluid, we have P = wρ, where w is a constant. The cases w = 0 and w = $\frac{1}{3}$ correspond, respectively, to dust and radiation. Show that for a flat universe with Λ = 0 dominated by a single component that is a perfect fluid, the solution to the Friedmann equations is of the form a $\propto$ t$^\delta$, and determine the exponent δ. Check your result in the dust case against the one in section 8.2, then find the exponent in the radiation case. Although the w = −1 case corresponds to a cosmological constant, show that the solution is not of this form for w = −1.

12. Apply the result of problem 11 to generalize the result of example 22 for the size of the observable universe. What is the result in the case of the radiation-dominated universe?

13. The Kantowski-Sachs metric is

$$ds^2 = dt^2 - \Lambda^{-1}(d\theta^2 + \sin^2\theta d\phi^2) - e^{2\sqrt{\Lambda}t}dz^2. \tag{8.E.5}$$

It describes a universe with the spatial topology of a 3-cylinder. Use a computer algebra system such as Maxima to verify the following facts.

a. Any world-line of the form (t, θ, ϕ, z) = (λ, constants) is a geodesic parametrized by proper time. (If using Maxima, you will find that the function cgeodesic() saves time here.)

b. If two such geodesics are separated only in the z direction, the distance between them along a surface of fixed t increases exponentially with t, while geodesics separated only in θ and ϕ do not recede from one another.

c. There are no matter fields, only a cosmological constant Λ.

d. The Ricci scalar R = −4Λ (+ − −− signature) equals $\frac{1}{3}$ of the value for the de Sitter vacuum-dominated cosmology (sec. 8.2), the factor of 3 occurring because there is expansion along only one axis rather than three.

e. The vacuum-dominated cosmology found by de Sitter and presented in the text was supposed to be the unique cosmological solution of this type. Why is the Kantowski-Sachs metric not a counterexample?

Contributors and Attributions

9: Gravitational Waves

Thumbnail: Two-dimensional representation of gravitational waves generated by two neutron stars orbiting each other. (Public Domain; NASA).

9.1: The Speed of Gravity

In Newtonian gravity, gravitational effects are assumed to propagate at infinite speed, so that for example the lunar tides correspond at any time to the position of the moon at the same instant. This clearly can't be true in relativity, since simultaneity isn't something that different observers even agree on. Not only should the "speed of gravity" be finite, but it seems implausible that it would be greater than c; in section 2.2, we argued based on empirically well established principles that there must be a maximum speed of cause and effect. Although the argument was only applicable to special relativity, i.e., to a flat spacetime, it seems likely to apply to general relativity as well, at least for low-amplitude waves on a flat background. As early as 1913, before Einstein had even developed the full theory of general relativity, he had carried out calculations in the weak-field limit showing that gravitational effects should propagate at c. We will work out an argument to this effect (using a different technique than Einstein's) in section 9.2. This seems eminently reasonable, since (a) it is likely to be consistent with causality, and (b) G and c are the only constants with units that appear in the field equations (obscured by our choice of units, in which G = 1 and c = 1), and the only velocity-scale that can be constructed from these two constants is c itself.[1]

> **Note**
>
> High-amplitude waves need *not* propagate at c. For example, general relativity predicts that a gravitational-wave pulse propagating on a background of curved spacetime develops a trailing edge that propagates at less than c (Misner, Thorne, and Wheeler, p. 957). This effect is weak when the amplitude is small or the wavelength is short compared to the scale of the background curvature.

As shown by the following timeline, Einstein's prediction was surprisingly difficult to verify.

1913	Einstein predicts gravitational waves traveling at c.
1982	Hulse-Taylor pulsar (sections 6.2, 9.2) seen to lose energy at the rate predicted by general relativity's prediction of gravitational radiation.
2016-2017	Direct detection of gravitational waves and verification that they propagate at c.

Why did this process take over a century? Naive arguments suggest that it should have been much easier. Workers as early as Newton and Laplace had investigated the consequences of a gravitational force that propagated at some finite speed. It is easy to show that, if nonrelativistic ideas about spacetime are retained, the predicted results are dramatic and not consistent with observation. For example, the earth and moon orbit about their common center of mass, which is inside the earth but offset from the earth's center. Suppose that we retain Newton's ideas about spacetime, but modify Newton's law of gravity to incorporate a time delay, with changes in the gravitational field propagating at some speed u. The force acting on the moon would then point toward the earth's location at a slightly earlier time, and this force would therefore have a component parallel to the moon's direction of motion. The force would do positive work on the moon and also exert a positive torque, the result being that the moon would spiral away. This is not consistent with the fact that the earth-moon system has remained fairly stable for billions of years, unless we take u to be very large. From the stability of orbits in the solar system, Laplace estimated $u \gtrsim 10^{15}$ m/s, many orders of magnitude greater than c. This seemed to support the Newtonian picture, in which gravity acts instantaneously at a distance. A time delay in Newtonian spacetime would also have been easily detected by twentieth-century measurements using space probes and radio astronomy.[2]

The trouble with such arguments is that when we substitute relativistic spacetime for Newtonian spacetime, it is no longer expected that a time-delayed field will point toward the retarded position of the source. For example, if an electric charge moves inertially, and is observed in a frame in which it is moving, then Lorentz invariance requires that its electric field lines be straight, and converge on the charge's present position in that frame.[3] The speed of gravity therefore turns out to be much harder to measure than Laplace had believed.

References

[2] For an example of an erroneous 2003 claim to have performed such a test, see Fomalont and Kopeikin, http://arxiv.org/abs/astro-ph/0302294. Their claims were debunked by Samuel, http://arxiv.org/abs/astro-ph/0304006, and Will, http://arxiv.org/abs/astro-

ph/0301145.

3 Crowell, Special Relativity, section 10.4

This page titled 9.1: The Speed of Gravity is shared under a CC BY-SA 3.0 license and was authored, remixed, and/or curated by Benjamin Crowell.

9.2: Gravitational Radiation (Part 1)

Empirical Evidence

The first strong empirical evidence of gravitational waves came in 1982. The Hulse-Taylor system (section 6.2) contains two neutron stars orbiting around their common center of mass, and the period of the orbit is observed to be decreasing gradually over time (Figure 9.2.1). This is interpreted as evidence that the stars are losing energy to radiation of gravitational waves.[4] As we'll see later, the rate of energy loss is in excellent agreement with the predictions of general relativity.
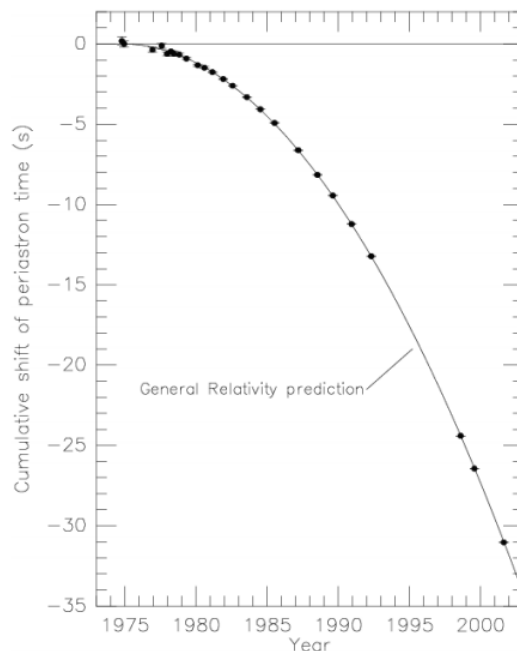
Figure 9.2.1 - The Hulse-Taylor pulsar's orbital motion is gradually losing energy due to the emission of gravitational waves. The linear decrease of the period is integrated on this plot, resulting in a parabola. From Weisberg and Taylor, http://arxiv.org/abs/astro-ph/0211217.

An even more dramatic, if less clearcut, piece of evidence is Komossa, Zhou, and Lu's observation[5] of a supermassive black hole that appears to be recoiling from its parent galaxy at a velocity of 2650 km/s (projected along the line of sight). They interpret this as evidence for the following scenario. In the early universe, galaxies form with supermassive black holes at their centers. When two such galaxies collide, the black holes can merge. The merger is a violent process in which intense gravitational waves are emitted, and these waves carry a large amount of momentum, causing the black holes to recoil at a velocity greater than the escape velocity of the merged galaxy.

Although the energy loss from systems such as the Hulse-Taylor binary provide strong evidence that gravitational waves exist and carry energy, physicists and astronomers still wanted to detect them directly, and serious attempts to design and build such systems began around 1962. The design that finally achieved success used interferometers which detect oscillations in the lengths of their own arms. The first gravitational-wave event was detected, by this method, in 2016, by the Advanced LIGO collaboration.[6] The event is believed to have been the result of the collision of two black holes.
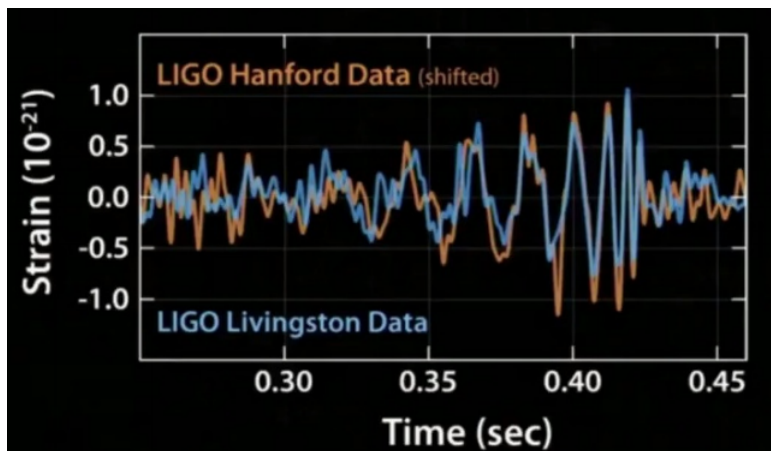
Figure **9.2.2** *- The gravitational waveform observed in 2016 by Advanced Ligo.*

In 2017, an event interpreted as the collision of two neutron stars was detected by both gravitational and electromagnetic radiation, verifying to high precision that gravitational waves propagate at c.

Although the earlier 2016 collision of black holes did not directly compare the propagation of light and gravity, it provided a different kind of check on the propagation of gravitational waves at c. The waveform detected in this event was a "chirp" that glided up in frequency as the black holes spiraled toward one another and sped up. Since the wave was in transit for over a billion years, and the waveform lasted a fraction of a second, it follows that gravitational waves within this frequency range all travel at very nearly the same velocity, i.e., there is a very tight upper limit on the dispersion of gravitational waves.

A complementary space-based system, LISA, has been proposed for launch in 2020, but its funding is uncertain. The two devices would operate in complementary frequency ranges (Figure 9.2.3). A selling point of LISA is that if it is launched, there are a number of sources in the sky, with known properties, that are known to be easily within its range of sensitivity.[7] One excellent candidate is HM Cancri, a pair of white dwarfs with an orbital period of 5.4 minutes, shorter than that of any other known binary star.[8]
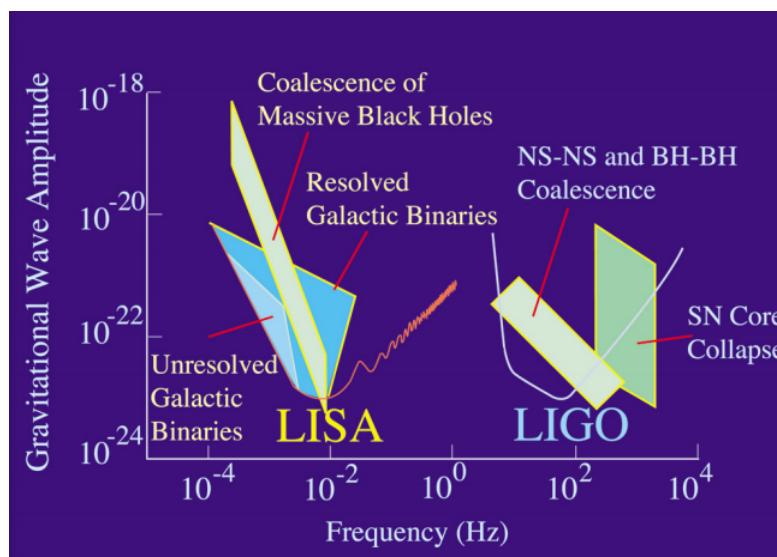
Figure 9.2.3 - Predicted sensitivities of LISA and LIGO to gravitational waves of various frequences.

Energy Content

Even without performing the calculations for a system like the Hulse-Taylor binary, it is easy to show that if such waves exist, they must be capable of carrying away energy. Consider two equal masses in highly elliptical orbits about their common center of mass, Figure 9.2.4. The motion is nearly one-dimensional. As the masses recede from one another, they feel a delayed version of the gravitational force originating from a time when they were closer together and the force was stronger. The result is that in the near-Newtonian limit, they lose more kinetic and gravitational energy than they would have lost in the purely Newtonian theory. Now

they come back inward in their orbits. As they approach one another, the time-delayed force is anomalously weak, so they gain less mechanical energy than expected. The result is that with each cycle, mechanical energy is lost. We expect that this energy is carried by the waves, in the same way that radio waves carry the energy lost by a transmitting antenna.[9]

Figure 9.2.4 - As the two planets recede from one another, each feels the gravitational attraction that the other one exerted in its previous position, delayed by the time it takes gravitational effects to propagate at c. At time t, the right-hand planet experiences the stronger deceleration corresponding to the left-hand planet's closer position at the earlier time t', not its current position at t. Mechanical energy is not conserved, and the orbits will decay

> **Note**
>
> One has to be careful with this type of argument. In particular, one can obtain incorrect correct results by attempting to generalize this one-dimensional argument to motion in more than one dimension, because the effective semi-Newtonian interaction is not just a time-delayed version of Newton's law; it also includes velocity-dependent forces. It is easy to see why such velocity-dependence must occur in the simpler case of electromagnetism. Suppose that charges A and B are not at rest relative to one another. In B's frame, the electric field from A must come from the direction of the position that an observer comoving with B would extrapolate linearly from A's last known position and velocity, as determined by light-speed calculation. This follows from Lorentz invariance, since this is the direction that will be seen by an observer comoving with A. A full discussion is given by Carlip, arxiv.org/abs/gr-qc/9909087v2.

Not only can these waves remove mechanical energy from a system, they can also deposit energy in a detector, as shown by the nonmathematical "sticky bead argument" (Figure 9.2.5), which was originated by Feynman in 1957 and later popularized by Bondi.
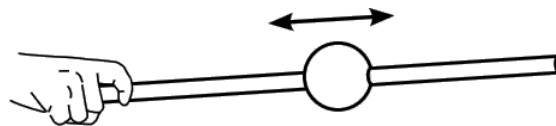
Figure 9.2.5 - T*he sticky bead argument for the reality of gravitational waves. As a gravitational wave with the appropriate polarization passes by, the bead vibrates back and forth on the rod. Friction creates heat. This demonstrates that gravitational waves carry energy, and are thus real, observable phenomena.*

Now strictly speaking, we have only shown that gravitational waves can extract or donate mechanical energy, but not that the waves themselves *transmit* this energy. The distinction isn't one that normally occurs to us, since we are trained to believe that energy is always conserved. But we know that, for fundamental reasons, general relativity doesn't have global conservation laws that apply to all spacetimes (section 4.5). Perhaps the energy lost by the Hulse-Taylor system is simply gone, never to reappear, and the energy imparted to the sticky bead is simply generated out of nowhere. On the other hand, general relativity does have global conservation laws for certain specific classes of spacetimes, including, for example, a conserved scalar mass-energy in the case of a stationary spacetime (section 7.1). Spacetimes containing gravitational waves are not stationary, but perhaps there is something similar we can do in some appropriate special case.

Suppose we want an expression for the energy of a gravitational wave in terms of its amplitude. This seems like it ought to be straightforward. We have such expressions in other classical field theories. In electromagnetism, we have energy densities $+(\frac{1}{8\pi k})|\mathbf{E}|^2$ and $+(\frac{1}{2\mu_o})|\mathbf{B}|^2$ associated with the electric and magnetic fields. In Newtonian gravity, we can assign an energy density $-(\frac{1}{8\pi G})|\mathbf{g}|^2$ to the gravitational field $\mathbf{g}$; the minus sign indicates that when masses glom onto each other, they produce a greater field, and energy is released.

In general relativity, however, the equivalence principle tells us that for any gravitational field measured by one observer, we can find another observer, one who is free-falling, who says that the local field is zero. It follows that we cannot associate an energy with the curvature of a particular region of spacetime in any exact way. The best we can do is to find expressions that give the energy density (1) in the limit of weak fields, and (2) when averaged over a region of space that is large compared to the wavelength. These expressions are not unique. There are a number of ways to write them in terms of the metric and its derivatives, and they all give the same result in the appropriate limit. The reader who is interested in seeing the subject developed in detail is referred to Carroll's *Lecture Notes on General Relativity*, arxiv.org/abs/gr-qc/?9712019. Although this sort of thing is technically

messy, we can accomplish quite a bit simply by knowing that such results do exist, and that although they are non-unique in general, they are uniquely well defined in certain cases. Specifically, when one wants to discuss gravitational waves, it is usually possible to assume an asymptotically flat spacetime. In an asymptotically flat spacetime, there is a scalar mass-energy, called the ADM mass, that is conserved. In this restricted sense, we are assured that the books balance, and that the emission and absorption of gravitational waves really does mean the *transmission* of a fixed amount of energy.

Expected Properties

To see what properties we should expect for gravitational radiation, first consider the reasoning that led to the construction of the Ricci and Einstein tensors. If a certain volume of space is filled with test particles, then the Ricci and Einstein tensors measure the tendency for this volume to "accelerate;" i.e., $-\frac{d^2 V}{dt^2}$ is a measure of the attraction of any mass lying inside the volume. A distant mass, however, will exert only tidal forces, which distort a region without changing its volume. This suggests that as a gravitational wave passes through a certain region of space, it should distort the shape of a given region, without changing its volume.

When the idea of gravitational waves was first discussed, there was some skepticism about whether they represented an effect that was observable, even in principle. The most naive such doubt is of the same flavor as the one discussed in section 8.2 about the observability of the universe's expansion: if everything distorts, then don't our meter-sticks distort as well, making it impossible to measure the effect? The answer is the same as before in section 8.2; systems that are gravitationally or electromagnetically bound do not have their scales distorted by an amount equal to the change in the elements of the metric.
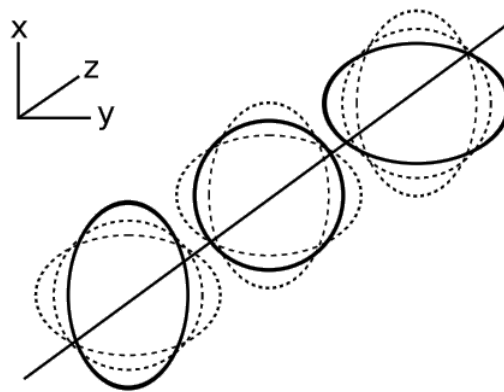
Figure 9.2.6 - As the gravitational wave propagates in the z direction, the metric oscillates in the x and y directions, preserving volume.

A less naive reason to be skeptical about gravitational waves is that just because a metric looks oscillatory, that doesn't mean its oscillatory behavior is observable. Consider the following example.

$$ds^2 = dt^2 - \left(1 + \frac{1}{10}\sin x\right) dx^2 - dy^2 - dz^2 \tag{9.2.1}$$

The Christoffel symbols depend on derivatives of the form $\partial_a g_{bc}$, so here the only nonvanishing Christoffel symbol is Γ^x_{xx}. It is then straightforward to check that the Riemann tensor

$$R^a_{bcd} = \partial_c \Gamma^a_{db} - \partial_d \Gamma^a_{cb} + \Gamma^a_{ce}\Gamma^e_{db} - \Gamma^a_{de}\Gamma^e_{cb} \tag{9.2.2}$$

vanishes by symmetry. Therefore this metric must really just be a flat-spacetime metric that has been subjected to a silly change of coordinates.

> **Exercise 9.2.1**
>
> Self-check: R vanishes, but Γ doesn't. Is there a reason for paying more attention to one or the other?

To keep the curvature from vanishing, it looks like we need a metric in which the oscillation is not restricted to a single variable.

For example, the metric

$$ds^2 = dt^2 - \left(1 + \frac{1}{10}\sin y\right) dx^2 - dy^2 - dz^2 \tag{9.2.3}$$

does have nonvanishing curvature. In other words, it seems like we should be looking for transverse waves rather than longitudinal ones.[10] On the other hand, this metric cannot be a solution to the vacuum field equations, since it doesn't preserve volume. It also stands still, whereas we expect that solutions to the field equations should propagate at the velocity of light, at least for small amplitudes. These conclusions are self-consistent, because a wave's polarization can only be constrained if it propagates at c (see section 4.2).

> **Note**
>
> 0A more careful treatment shows that longitudinal waves can always be interpreted as physically unobservable coordinate waves, in the limit of large distances from the source. On the other hand, it is clear that no such prohibition against longitudinal waves could apply universally, because such a constraint can only be Lorentz-invariant if the wave propagates at c (see section 4.2), whereas high-amplitude waves need not propagate at c. Longitudinal waves near the source are referred to as Type III solutions in a classification scheme due to Petrov. Transverse waves, which are what we could actually observe in practical experiments, are type N.

Based on what we've found out, the following seems like a metric that might have a fighting chance of representing a real gravitational wave:

$$ds^2 = dt^1 - (1 + A\sin(z-t))dx^2 - \frac{dy^2}{1 + A\sin(z-t)} - dz^2 \tag{9.2.4}$$

It is transverse, it propagates at c(= 1), and the fact that g_{xx} is the reciprocal of g_{yy} makes it volume-conserving. The following Maxima program calculates its Einstein tensor:

```
1    load(ctensor);
2    ct_coords:[t,x,y,z];
3    lg:matrix([1,0,0,0],
4              [0,-(1+A*sin(z-t)),0,0],
5              [0,0,-1/(1+A*sin(z-t)),0],
6              [0,0,0,-1]);
7    cmetric();
8    einstein(true);
```

For a representative component of the Einstein tensor, we find

$$G_{tt} = \frac{A^2 \cos^2(z-t)}{2 + 4A\sin(z-t) + 2A^2\sin^2(z-t)} \tag{9.2.5}$$

For small values of A, we have $|G_{tt}| \lesssim \frac{A^2}{2}$. The vacuum field equations require $G_{tt} = 0$, so this isn't an exact solution. But all the components of G, not just G_{tt}, are of order A^2, so this is an *approximate* solution to the equations.

It is also straightforward to check that propagation at approximately c was a necessary feature. For example, if we replace the factors of sin(z −t) in the metric with sin(z −2t), we get a G_{xx} that is of order unity, not of order A^2.

To prove that gravitational waves are an observable effect, we would like to be able to display a metric that (1) is an exact solution of the vacuum field equations; (2) is not merely a coordinate wave; and (3) carries momentum and energy. As late as 1936, Einstein and Rosen published a paper claiming that gravitational waves were a mathematical artifact, and did not actually exist.[11]

References

[4] Stairs, "Testing General Relativity with Pulsar Timing," relativity.livingreviews.org/...es/lrr-2003-5/

[5] http://arxiv.org/abs/0804.4585

[6] https://dcc.ligo.org/LIGO-P150914/public

[7] G. Nelemans, "The Galactic Gravitational wave foreground," arxiv.org/abs/0901.1778v1

8 Roelofs et al., "Spectroscopic Evidence for a 5.4-Minute Orbital Period in HM Cancri," arxiv.org/abs/1003.0658v1

11 Some of the history is related at http://en.Wikipedia.org/wiki/Sticky_bead_argument.

9.3: Gravitational Radiation (Part 2)

In this section we study several examples of exact solutions to the field equations. Each of these can readily be shown not to be a mere coordinate wave, since in each case the Riemann tensor has nonzero elements.

Example 1: An exact solution

We've already seen, e.g., in the derivation of the Schwarzschild metric in section 6.2, that once we have an approximate solution to the equations of general relativity, we may be able to find a series solution. Historically this approach was only used as a last resort, because the lack of computers made the calculations too complex to handle, and the tendency was to look for tricks that would make a closed-form solution possible. But today the series method has the advantage that any mere mortal can have some reasonable hope of success with it — and there is nothing more boring (or demoralizing) than laboriously learning someone else's special trick that only works for a specific problem. In this example, we'll see that such an approach comes tantalizingly close to providing an exact, oscillatory plane wave solution to the field equations.

Our best solution so far was of the form

$$ds^2 = dt^2 - (1+f)dx^2 - \frac{dy^2}{1+f} - dz^2, \tag{9.2.6}$$

where f = Asin(z − t). This doesn't seem likely to be an exact solution for large amplitudes, since the x and y coordinates are treated asymmetrically. In the extreme case of $|A| \geq 1$, there would be singularities in g_{yy}, but not in g_{xx}. Clearly the metric will have to have some kind of nonlinear dependence on f, but we just haven't found quite the right nonlinear dependence. Suppose we try something of this form:

$$ds^2 = dt^2 - (1+f+cf^2)dx^2 - (1-f+df^2)dy^2 - dz^2 \tag{9.2.7}$$

This approximately conserves volume, since (1+f +. . .)(1−f +. . .) equals unity, up to terms of order f^2. The following program tests this form.

```
1    load(ctensor);
2    ct_coords:[t,x,y,z];
3    f : A*exp(%i*k*(z-t));
4    lg:matrix([1,0,0,0],
5              [0,-(1+f+c*f^2),0,0],
6              [0,0,-(1-f+d*f^2),0],
7              [0,0,0,-1]);
8    cmetric();
9    einstein(true);
```

In line 3, the motivation for using the complex exponential rather than a sine wave in f is the usual one of obtaining simpler expressions; as we'll see, this ends up causing problems. In lines 5 and 6, the symbols c and d have not been defined, and have not been declared as depending on other variables, so Maxima treats them as unknown constants. The result is $G_{tt} \sim (4d + 4c - 3)A^2$ for small A, so we can make the A^2 term disappear by an appropriate choice of d and c. For symmetry, we choose c = d = $\frac{3}{8}$. With these values of the constants, the result for G_{tt} is of order A^4. This technique can be extended to higher and higher orders of approximation, resulting in an exact series solution to the field equations.

Unfortunately, the whole story ends up being too good to be true. The resulting metric has complex-valued elements. If general relativity were a linear field theory, then we could apply the usual technique of forming linear combinations of expressions of the form $e^{+i\cdots}$ and $e^{-i\cdots}$, so as to give a real result. But the field equations of general relativity are nonlinear, so the resulting linear combination is no longer a solution. The best we can do is to make a non-oscillatory real exponential solution (problem 2).

<div style="border-left: 4px solid #1a6fb5;">

Example 2: An exact, oscillatory, non-monochromatic solution

Assume a metric of the form

$$ds^2 = dt^2 - p(z-t)^2 dx^2 - q(z-t)^2 dy^2 - dz^2, \ldots \qquad (9.2.8)$$

where p and q are arbitrary functions. Such a metric would clearly represent some kind of transverse-polarized plane wave traveling at velocity c(= 1) in the z direction. The following Maxima code calculates its Einstein tensor.

```
1    load(ctensor);
2    ct_coords:[t,x,y,z];
3    depends(p,[z,t]);
4    depends(q,[z,t]);
5    lg:matrix([1,0,0,0],
6              [0,-p^2,0,0],
7              [0,0,-q^2,0],
8              [0,0,0,-1]);
9    cmetric();
10   einstein(true);
```

The result is proportional to $\frac{\ddot{q}}{q} + \frac{\ddot{p}}{p}$, so any functions p and q that satisfy the differential equation $\frac{\ddot{q}}{q} + \frac{\ddot{p}}{p} = 0$ will result in a solution to the field equations. Setting p(u) = 1 + Acos u, for example, we find that q is oscillatory, but with a period longer than 2π (problem 3).

</div>

<div style="border-left: 4px solid #1a6fb5;">

Example 3: An exact, plane, monochromatic wave

Any metric of the form

$$ds^2 = (1-h)dt^2 - dx^2 - dy^2 - (1+h)dz^2 + 2hdzdt, \qquad (9.2.9)$$

where h = f(z − t)xy, and f is any function, is an exact solution of the field equations (problem 4).

Because h is proportional to xy, this does not appear at first glance to be a uniform plane wave. One can verify, however, that all the components of the Riemann tensor depend only on z − t, not on x or y. Therefore there is no measurable property of this metric that varies with x and y.

</div>

Rate of Radiation

How can we find the rate of gravitational radiation from a system such as the Hulse-Taylor pulsar?

Let's proceed by analogy. The simplest source of sound waves is something like the cone of a stereo speaker. Since typical sound waves have wavelengths measured in meters, the entire speaker is generally small compared to the wavelength. The speaker cone is a surface of oscillating displacement $(x = x_o \sin \omega t)$. Idealizing such a source to a radially pulsating spherical surface, we have an oscillating monopole that radiates sound waves uniformly in all directions. To find the power radiated, we note that the velocity of the source-surface is proportional to $x_o\omega$, so the kinetic energy of the air immediately in contact with it is proportional to $\omega^2 x_o^2$. The power radiated is therefore proportional to $\omega^2 x_o^2$.
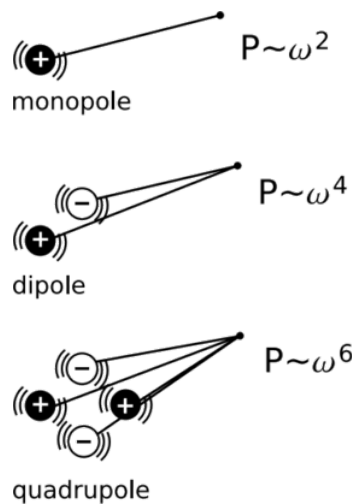
Figure 9.3.7: The power emitted by a multipole source of order m is proportional to $\omega^{2(m+1)}$, when the size of the source is small compared to the wavelength. The main reason for the ω dependence is that at low frequencies, the wavelength is long, so the number of wavelengths traveled to a particular point in space is nearly the same from any point in the source; we therefore get strong cancellation.

In electromagnetism, conservation of charge forbids the existence of an oscillating electric monopole. The simplest radiating source is therefore an oscillating electric dipole D = D_o sin ωt. If the dipole's physical size is small compared to a wavelength of the radiation, then the radiation is an inefficient process; at any point in space, there is only a small difference in path length between the positive and negative portions of the dipole, so there tends to be strong cancellation of their contributions, which were emitted with opposite phases. The result is that the wave's electromagnetic potential four-vector (Section 4.2) is proportional to $D_o\omega$, the fields to $D_o\omega^2$, and the radiated power to $D_o^2\omega^4$. The factor of ω^4 can be broken down into $(\omega^2)(\omega^2)$, where the first factor of ω^2 occurs for reasons similar to the ones that explain the ω^2 factor for the monopole radiation of sound, while the second ω^2 arises because the smaller ω is, the longer the wavelength, and the greater the inefficiency in radiation caused by the small size of the source compared to the wavelength.

Example 4: am radio

Commercial AM radio uses wavelengths of several hundred meters, so AM dipole antennas are usually orders of magnitude shorter than a wavelength. This causes severe attenuation in both transmission and reception. (There are theorems called reciprocity theorems that relate efficiency of transmission to efficiency of reception.) Receivers therefore need to use of a large amount of amplification. This doesn't cause problems, because the ambient sources of RF noise are attenuated by the short antenna just as severely as the signal.

Since our universe doesn't seem to have particles with negative mass, we can't form a gravitational dipole by putting positive and negative masses on opposite ends of a stick — and furthermore, such a stick will not spin freely about its center, because its center of mass does not lie at its center! In a more realistic system, such as the Hulse-Taylor pulsar, we have two unequal masses orbiting about their common center of mass. By conservation of momentum, the mass dipole moment of such a system is constant, so we cannot have an oscillating mass dipole. The simplest source of gravitational radiation is therefore an oscillating mass quadrupole, Q = Q_o sin ωt. As in the case of the oscillating electric dipole, the radiation is suppressed if, as is usually the case, the source is small compared to the wavelength. The suppression is even stronger in the case of a quadrupole, and the result is that the radiated power is proportional to $Q_o^2\omega^6$.

This result has the interesting property of being invariant under a rescaling of coordinates. In geometrized units, mass, distance, and time all have the same units, so that Q_o^2 has units of (length^3)^2 while ω^6 has units of (length)$^{-6}$. This is exactly what is required, because in geometrized units, power is unitless, energy/time = length/length = 1.

We can also tie the ω^6 dependence to our earlier argument for the dissipation of energy by gravitational waves. The argument was that gravitating bodies are subject to time-delayed gravitational forces, with the result that orbits tend to decay. This argument only works if the forces are time-varying; if the forces are constant over time, then the time delay has no effect. For example, in the semi-Newtonian limit the field of a sheet of mass is independent of distance from the sheet. (The electrical analog of this fact is easily proved using Gauss's law.) If two parallel sheets fall toward one another, then neither is subject to a time-varying force, so

there will be no radiation. In general, we expect that there will be no gravitational radiation from a particle unless the third derivative of its position d3 x/ dt 3 is nonzero. (The same is true for electric quadrupole radiation.) In the special case where the position oscillates sinusoidally, the chain rule tells us that taking the third derivative is equivalent to bringing out a factor of ω^3. Since the amplitude of gravitational waves is proportional to $\frac{d^3x}{dt^3}$, their energy varies as $(\frac{d^3x}{dt^3})^2$, or ω^6.

The general pattern we have observed is that for multipole radiation of order m (0=monopole, 1=dipole, 2=quadrupole), the radiated power depends on $\omega^{2(m+1)}$. Since gravitational radiation must always have m = 2 or higher, we have the very steep ω^6 dependence of power on frequency. This demonstrates that if we want to see strong gravitational radiation, we need to look at systems that are oscillating extremely rapidly. For a binary system with unequal masses of order m, with orbits having radii of order r, we have $Q_o \sim mr^2$. Newton's laws give $\omega \sim m^{1/2}r^{-3/2}$, which is essentially Kepler's law of periods. The result is that the radiated power should depend on $(\frac{m}{r})^5$. Reinserting the proper constants to give an equation that allows practical calculation in SI units, we have

$$P = k\frac{G^4}{c^5}\left(\frac{m}{r}\right)^5,\tag{9.2.10}$$

where k is a unitless constant of order unity.

For the Hulse-Taylor pulsar,[12] we have m $\sim$ 3 × 10^{30} kg (about one and a half solar masses) and r $\sim$ 10^9 m. The binary pulsar is made to order our purposes, since $\frac{m}{r}$ is extremely large compared to what one sees in almost any other astronomical system. The resulting estimate for the power is about 10^{24} watts. The pulsar's period is observed to be steadily lengthening at a rate of α = 2.418 × 10^{-12} seconds per second. To compare this with our crude theoretical estimate, we take the Newtonian energy of the system $\frac{Gm^2}{r}$ and multiply by $\omega\alpha$, giving 10^{25} W, which checks to within an order of magnitude. A full general-relativistic calculation reproduces the observed value of α to within the 0.1% error bars of the data.

References

[12] http://arxiv.org/abs/astro-ph/0407149

9.E: Gravitational Waves (Exercises)

1. (a) Starting in section 1.5, we have associated geodesics with the world-lines of low-mass objects (test particles). Use the HulseTaylor pulsar as an example to show that the assumption of low mass was a necessary one. How is this similar to the issues encountered in chapter 1 problems involving charged particles?

    (b) Show that if low-mass, uncharged particles did not follow geodesics (in a spacetime with no ambient electromagnetic fields), it would violate Lorentz invariance. Make sure that your argument explicitly invokes the low mass and the lack of charge, because otherwise your argument is wrong.

2. Show that the metric $ds^2 = dt^2 - A\,dx^2 - B\,dy^2 - dz^2$ with

$$A = 1 - f + \frac{3}{8}f^2 - \frac{25}{416}f^3 + \frac{15211}{10729472}f^5$$
$$B = 1 + f + \frac{3}{8}f^2 + \frac{25}{416}f^3 - \frac{15211}{10729472}f^5$$
$$f = Ae^{k(t-z)}$$

    is an approximate solution to the vacuum field equations, provided that k is real — which prevents this from being a physically realistic, oscillating wave. Find the next nonvanishing term in each series.

3. Verify the claims made in example 2. Characterize the (somewhat complex) behavior of the function q obtained when p(u) = 1 + A cos u.

4. Verify the claims made in example 3 using Maxima. Although the result holds for any function f, you may find it more convenient to use some specific form of f, such as a sine wave, so that Maxima will be able to simplify the result to zero at the end. Note that when the metric is expressed in terms of the line element, there is a factor of 2 in the 2h dz dt term, but when expressing it as a matrix, the 2 is not present in the matrix elements, because there are two elements in the matrix that each contribute an equal amount.

10: Appendices

Appendix A: Excerpts from three papers by Einstein

10.1: Appendix A (Part 1)

The following English translations of excerpts from three papers by Einstein were originally published in "The Principle of Relativity," Methuen and Co., 1923. The translation was by W. Perrett and G.B. Jeffery, and notes were provided by A. Sommerfeld. John Walker (www. fourmilab.ch) has provided machine-readable versions of the first two and placed them in the public domain. Some notation has been modernized, British spelling has been Americanized, etc. Footnotes by Sommerfeld, Walker, and B. Crowell are marked with initials. B. Crowell's modifications to the present version are also in the public domain.

- The paper "On the electrodynamics of moving bodies" contains two parts, the first dealing with kinematics and the second with electrodynamics. I've given only the first part here, since the second one is lengthy, and painful to read because of the cumbersome old-fashioned notation. The second section can be obtained from John Walker's web site.
- The paper "Does the inertia of a body depend upon its energy content?," which begins later, is very short and readable. A shorter and less general version of its main argument is given in section 4.2.
- "The foundation of the general theory of relativity" is a long review article in which Einstein systematically laid out the general theory, which he had previously published in a series of shorter papers. The first three sections of the paper give the general physical reasoning behind coordinate independence, referred to as general covariance. It begins later.

The reader who is interested in seeing these papers in their entirety can obtain them inexpensively in a Dover reprint of the original Methuen anthology.

"On the Electrodynamics of Moving Bodies"

A. Einstein, Annalen der Physik 17 (1905) 891. It is known that Maxwell's electrodynamics—as usually understood at the present time— when applied to moving bodies, leads to asymmetries which do not appear to be inherent in the phenomena.[13] Take, for example, the reciprocal electrodynamic action of a magnet and a conductor. The observable phenomenon here depends only on the relative motion of the conductor and the magnet, whereas the customary view draws a sharp distinction between the two cases in which either the one or the other of these bodies is in motion. For if the magnet is in motion and the conductor at rest, there arises in the neighbourhood of the magnet an electric field with a certain definite energy, producing a current at the places where parts of the conductor are situated. But if the magnet is stationary and the conductor in motion, no electric field arises in the neighbourhood of the magnet. In the conductor, however, we find an electromotive force, to which in itself there is no corresponding energy, but which gives rise— assuming equality of relative motion in the two cases discussed—to electric currents of the same path and intensity as those produced by the electric forces in the former case.

> **Note**
>
> Einstein begins by giving an example involving electromagnetic induction, considered in two different frames of reference. With modern hindsight, we would describe this by saying that a Lorentz boost mixes the electric and magnetic fields, as described in section 4.2. —BC

Examples of this sort, together with the unsuccessful attempts to discover any motion of the earth relative to the "light medium," suggest that the phenomena of electrodynamics as well as of mechanics possess no properties corresponding to the idea of absolute rest.[14] They suggest rather that, as has already been shown to the first order of small quantities,[15] the same laws of electrodynamics and optics will be valid for all frames of reference for which the equations of mechanics hold good.[16] We will raise this conjecture (the purport of which will hereafter be called the "Principle of Relativity") to the status of a postulate, and also introduce another postulate, which is only apparently irreconcilable with the former, namely, that light is always propagated in empty space with a definite velocity c which is independent of the state of motion of the emitting body.[17] These two postulates suffice for the attainment of a simple and consistent theory of the electrodynamics of moving bodies based on Maxwell's theory for stationary bodies. The introduction of a "luminiferous ether" will prove to be superfluous inasmuch as the view here to be developed will not require an "absolutely stationary space" provided with special properties, nor assign a velocity-vector to a point of the empty space in which electromagnetic processes take place.

> **Notes**
>
> [14] Einstein knew about the Michelson-Morley experiment by 1905 (J. van Dongen, arxiv.org/abs/0908.1545), but it isn't cited specifically here. The 1881 and 1887 Michelson-Morley papers are available online at en. wikisource.org. —BC

[15] I.e., to first order in $\frac{v}{c}$. Experimenters as early as Fresnel (1788-1827) had shown that there were no effects of order $\frac{v}{c}$ due to the earth's motion through the aether, but they were able to interpret this without jettisoning the aether, by contriving models in which solid substances dragged the aether along with them. The negative result of the Michelson-Morley experiment showed a lack of an effect of order $(\frac{v}{c})^2$. —BC

[16] The preceding memoir by Lorentz was not at this time known to the author. —AS

[17] The second postulate is redundant if we take the "laws of electrodynamics and optics" to refer to Maxwell's equations. Maxwell's equations require that light move at c in any frame of reference in which they are valid, and the first postulate has already claimed that they are valid in all inertial frames of reference. Einstein probably states constancy of c as a separate postulate because his audience is accustomed to thinking of Maxwell's equations as a partial mathematical representation of certain aspects of an underlying aether theory. Throughout part I of the paper, Einstein is able to derive all his results without assuming anything from Maxwell's equations other than the constancy of c. The use of the term "postulate" suggests the construction of a formal axiomatic system like Euclidean geometry, but Einstein's real intention here is to lay out a set of philosophical criteria for evaluating candidate theories; he freely brings in other, less central, assumptions later in the paper, as when he invokes homogeneity of spacetime later. —BC

[18] Essentially what Einstein means here is that you can't have Maxwell's equations without establishing position and time coordinates, and you can't have position and time coordinates without clocks and rulers. Therefore even the description of a purely electromagnetic phenomenon such as a light wave depends on the existence of material objects. He doesn't spell out exactly what he means by "rigid," and we now know that relativity doesn't actually allow the existence of perfectly rigid solids (see section 3.5). Essentially he wants to be able to talk about rulers that behave like solids rather than liquids, in the sense that if they are accelerated sufficiently gently from rest and later brought gently back to rest, their properties will be unchanged. When he derives the length contraction later, he wants it to be clear that this isn't a dynamical phenomenon caused by an effect such as the drag of the aether.—BC

The theory to be developed is based—like all electrodynamics—on the kinematics of the rigid body, since the assertions of any such theory have to do with the relationships between rigid bodies (systems of coordinates), clocks, and electromagnetic processes.[18] Insufficient consideration of this circumstance lies at the root of the difficulties which the electrodynamics of moving bodies at present encounters.

Kinematical Part

§1. Definition of Simultaneity

Let us take a system of coordinates in which the equations of Newtonian mechanics hold good.[19] In order to render our presentation more precise and to distinguish this system of coordinates verbally from others which will be introduced hereafter, we call it the "stationary system."

> **Note**
>
> i.e., to the first approximation.—AS

If a material point is at rest relative to this system of coordinates, its position can be defined relative thereto by the employment of rigid standards of measurement and the methods of Euclidean geometry, and can be expressed in Cartesian coordinates.

If we wish to describe the *motion* of a material point, we give the values of its coordinates as functions of the time. Now we must bear carefully in mind that a mathematical description of this kind has no physical meaning unless we are quite clear as to what we understand by "time." We have to take into account that all our judgments in which time plays a part are always judgments of *simultaneous events*. If, for instance, I say, "That train arrives here at 7 o'clock," I mean something like this: "The pointing of the small hand of my watch to 7 and the arrival of the train are simultaneous events."[20]

> **Note**
>
> We shall not here discuss the inexactitude which lurks in the concept of simultaneity of two events at approximately the same place, which can only be removed by an abstraction.—AS

It might appear possible to overcome all the difficulties attending the definition of "time" by substituting "the position of the small hand of my watch" for "time." And in fact such a definition is satisfactory when we are concerned with defining a time exclusively for the place where the watch is located; but it is no longer satisfactory when we have to connect in time series of events occurring at different places, or—what comes to the same thing—to evaluate the times of events occurring at places remote from the watch.

We might, of course, content ourselves with time values determined by an observer stationed together with the watch at the origin of the coordinates, and coordinating the corresponding positions of the hands with light signals, given out by every event to be timed, and reaching him through empty space. But this coordination has the disadvantage that it is not independent of the standpoint of the observer with the watch or clock, as we know from experience. We arrive at a much more practical determination along the following line of thought.

If at the point A of space there is a clock, an observer at A can determine the time values of events in the immediate proximity of A by finding the positions of the hands which are simultaneous with these events. If there is at the point B of space another clock in all respects resembling the one at A, it is possible for an observer at B to determine the time values of events in the immediate neighbourhood of B. But it is not possible without further assumption to compare, in respect of time, an event at A with an event at B. We have so far defined only an "A time" and a "B time." We have not defined a common "time" for A and B, for the latter cannot be defined at all unless we establish *by definition* that the "time" required by light to travel from A to B equals the "time" it requires to travel from B to A. Let a ray of light start at the "A time" t_A from A towards B, let it at the "B time" t_B be reflected at B in the direction of A, and arrive again at A at the "A time" t'_A.

In accordance with definition the two clocks synchronize[21] if

$$t_B - t_A = t'_A - t_B. \tag{10.1.1}$$

> **Note**
>
> The procedure described here is known as Einstein synchronization.—BC

We assume that this definition of synchronism is free from contradictions, and possible for any number of points; and that the following relations are universally valid:—

1. If the clock at B synchronizes with the clock at A, the clock at A synchronizes with the clock at B.
2. If the clock at A synchronizes with the clock at B and also with the clock at C, the clocks at B and C also synchronize with each other.[22]

> **Note**
>
> This assumption fails in a rotating frame (see section 3.5), but Einstein has restricted himself here to an approximately inertial frame of reference.—BC

Thus with the help of certain imaginary physical experiments we have settled what is to be understood by synchronous stationary clocks located at different places, and have evidently obtained a definition of "simultaneous," or "synchronous," and of "time." The "time" of an event is that which is given simultaneously with the event by a stationary clock located at the place of the event, this clock being synchronous, and indeed synchronous for all time determinations, with a specified stationary clock.

In agreement with experience we further assume the quantity

$$\frac{2AB}{t'_A - t_A} = c, \tag{10.1.2}$$

to be a universal constant—the velocity of light in empty space.

It is essential to have time defined by means of stationary clocks in the stationary system, and the time now defined being appropriate to the stationary system we call it "the time of the stationary system."

§ 2. On the Relativity of Lengths and Times

The following reflections are based on the principle of relativity and on the principle of the constancy of the velocity of light. These two principles we define as follows:—

1. The laws by which the states of physical systems undergo change are not affected, whether these changes of state be referred to the one or the other of two systems of coordinates in uniform translatory motion.
2. Any ray of light moves in the "stationary" system of coordinates with the determined velocity c, whether the ray be emitted by a stationary or by a moving body. Hence

$$velocity = \frac{light\ path}{time\ interval} \tag{10.1.3}$$

where time interval is to be taken in the sense of the definition in § 1.

Let there be given a stationary rigid rod; and let its length be l as measured by a measuringrod which is also stationary. We now imagine the axis of the rod lying along the axis of x of the stationary system of coordinates, and that a uniform motion of parallel translation with velocity v along the axis of x in the direction of increasing x is then imparted to the rod. We now inquire as to the length of the moving rod, and imagine its length to be ascertained by the following two operations:—

a. The observer moves together with the given measuring-rod and the rod to be measured, and measures the length of the rod directly by superposing the measuring-rod, in just the same way as if all three were at rest.
b. By means of stationary clocks set up in the stationary system and synchronizing in accordance with § 1, the observer ascertains at what points of the stationary system the two ends of the rod to be measured are located at a definite time. The distance between these two points, measured by the measuring-rod already employed, which in this case is at rest, is also a length which may be designated "the length of the rod."

In accordance with the principle of relativity the length to be discovered by the operation (a)—we will call it "the length of the rod in the moving system"—must be equal to the length l of the stationary rod.

The length to be discovered by the operation (b) we will call "the length of the (moving) rod in the stationary system." This we shall determine on the basis of our two principles, and we shall find that it differs from l.

Current kinematics tacitly assumes that the lengths determined by these two operations are precisely equal, or in other words, that a moving rigid body at the epoch t may in geometrical respects be perfectly represented by *the same* body *at rest* in a definite position.

We imagine further that at the two ends A and B of the rod, clocks are placed which synchronize with the clocks of the stationary system, that is to say that their indications correspond at any instant to the "time of the stationary system" at the places where they happen to be. These clocks are therefore "synchronous in the stationary system."

We imagine further that with each clock there is a moving observer, and that these observers apply to both clocks the criterion established in § 1 for the synchronization of two clocks. Let a ray of light depart from A at the time[23] t_A, let it be reflected at B at the time t_B, and reach A again at the time t'_A. Taking into consideration the principle of the constancy of the velocity of light we find that

$$t_B - t_A = \frac{r_{AB}}{c-v} \quad and \quad t_A - t_B = \frac{r_{AB}}{c+v} \tag{10.1.4}$$

where r_{AB} denotes the length of the moving rod—measured in the stationary system. Observers moving with the moving rod would thus find that the two clocks were not synchronous, while observers in the stationary system would declare the clocks to be synchronous.

> **Note**
>
> "Time" here denotes "time of the stationary system" and also "position of hands of the moving clock situated at the place under discussion."—AS

So we see that we cannot attach any *absolute* signification to the concept of simultaneity, but that two events which, viewed from a system of coordinates, are simultaneous, can no longer be looked upon as simultaneous events when envisaged from a system which is in motion relative to that system.

§ 3. Theory of the Transformation of Coordinates and Times from a Stationary System to Another System in Uniform Motion of Translation Relative to the Former

Let us in "stationary" space take two systems of coordinates, i.e., two systems, each of three rigid material lines, perpendicular to one another, and issuing from a point. Let the axes of X of the two systems coincide, and their axes of Y and Z respectively be parallel. Let each system be provided with a rigid measuring-rod and a number of clocks, and let the two measuring-rods, and likewise all the clocks of the two systems, be in all respects alike.

Now to the origin of one of the two systems (k) let a constant velocity v be imparted in the direction of the increasing x of the other stationary system (K), and let this velocity be communicated to the axes of the coordinates, the relevant measuring-rod, and the clocks. To any time of the stationary system K there then will correspond a definite position of the axes of the moving system, and from reasons of symmetry we are entitled to assume that the motion of k may be such that the axes of the moving system are at the time t (this "t" always denotes a time of the stationary system) parallel to the axes of the stationary system.

We now imagine space to be measured from the stationary system K by means of the stationary measuring-rod, and also from the moving system k by means of the measuring-rod moving with it; and that we thus obtain the coordinates x, y, z, and ξ, η, ζ respectively. Further, let the time t of the stationary system be determined for all points thereof at which there are clocks by means of light signals in the manner indicated in § 1; similarly let the time τ of the moving system be determined for all points of the moving system at which there are clocks at rest relative to that system by applying the method, given in § 1, of light signals between the points at which the latter clocks are located.

To any system of values x, y, z, t, which completely defines the place and time of an event in the stationary system, there belongs a system of values ξ, η, ζ, τ, determining that event relative to the system k, and our task is now to find the system of equations connecting these quantities.

In the first place it is clear that the equations must be linear on account of the properties of homogeneity which we attribute to space and time.

If we place x' = x − vt, it is clear that a point at rest in the system k must have a system of values x', y, z, independent of time. We first define τ as a function of x;, y, z, and t. To do this we have to express in equations that τ is nothing else than the summary of the data of clocks at rest in system k, which have been synchronized according to the rule given in § 1.

From the origin of system k let a ray be emitted at the time τ_0 along the X-axis to x', and at the time τ_1 be reflected thence to the origin of the coordinates, arriving there at the time τ_2; we then must have $\frac{1}{2}(\tau_0 + \tau_2) = \tau_1$, or, by inserting the arguments of the function τ and applying the principle of the constancy of the velocity of light in the stationary system:—

$$\frac{1}{2}\left[\tau(0,0,0,t) + \tau\left(0,0,0,t + \frac{x'}{c-v} + \frac{x'}{c+v}\right)\right] = \tau\left(x',0,0,t + \frac{x'}{c-v}\right). \tag{10.1.5}$$

Hence, if x' be chosen infinitesimally small,

$$\frac{1}{2}\left(\frac{1}{c-v} + \frac{1}{c+v}\right)\frac{\partial\tau}{\partial t} = \frac{\partial\tau}{\partial x'} + \frac{1}{c-v}\frac{\partial\tau}{\partial t}, \tag{10.1.6}$$

or

$$\frac{\partial\tau}{\partial x'} + \frac{v}{c^2 - v^2}\frac{\partial\tau}{\partial t} = 0. \tag{10.1.7}$$

It is to be noted that instead of the origin of the coordinates we might have chosen any other point for the point of origin of the ray, and the equation just obtained is therefore valid for all values of x', y, z.

An analogous consideration—applied to the axes of Y and Z—it being borne in mind that light is always propagated along these axes, when viewed from the stationary system, with the velocity $\sqrt{c^2 - v^2}$ gives us

$$\frac{\partial\tau}{\partial y} = 0, \ \frac{\partial\tau}{\partial z} = 0. \tag{10.1.8}$$

Since τ is a linear function, it follows from these equations that

$$\tau = a\left(t - \frac{v}{c^2 - v^2}x'\right) \tag{10.1.9}$$

where a is a function $\phi(v)$ at present unknown, and where for brevity it is assumed that at the origin of k, $\tau = 0$, when t = 0.

With the help of this result we easily determine the quantities ξ, η, ζ by expressing in equations that light (as required by the principle of the constancy of the velocity of light, in combination with the principle of relativity) is also propagated with velocity c when measured in the moving system. For a ray of light emitted at the time $\tau = 0$ in the direction of the increasing ξ

$$\xi = c\tau \quad or \quad \xi = ac\left(t - \frac{v}{c^2 - v^2}x'\right). \tag{10.1.10}$$

But the ray moves relative to the initial point of k, when measured in the stationary system, with the velocity c − v, so that

$$\frac{x'}{c - v} = t. \tag{10.1.11}$$

If we insert this value of t in the equation for ξ, we obtain

$$\xi = a\frac{c^2}{c^2 - v^2}x'. \tag{10.1.12}$$

In an analogous manner we find, by considering rays moving along the two other axes, that

$$\eta = c\tau = ac\left(t - \frac{v}{c^2 - v^2}x'\right) \tag{10.1.13}$$

when

$$\frac{y}{\sqrt{c^2 - v^2}} = t, x' = 0. \tag{10.1.14}$$

Thus

$$\eta = a\frac{c}{\sqrt{c^2 - v^2}}y \quad and \quad \zeta = a\frac{c}{\sqrt{c^2 - v^2}}z. \tag{10.1.15}$$

Substituting for x' its value, we obtain

$$\tau = \phi(v)\beta\left(t - \frac{vx}{c^2}\right),$$
$$\xi = \phi(v)\beta(x - vt),$$
$$\eta = \phi(v)y,$$
$$\zeta = \phi(v)z,$$

where

$$\beta = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}, \tag{10.1.16}$$

and ϕ is an as yet unknown function of v. If no assumption whatever be made as to the initial position of the moving system and as to the zero point of τ, an additive constant is to be placed on the right side of each of these equations.

We now have to prove that any ray of light, measured in the moving system, is propagated with the velocity c, if, as we have assumed, this is the case in the stationary system; for we have not as yet furnished the proof that the principle of the constancy of the velocity of light is compatible with the principle of relativity.

At the time t = $\tau = 0$, when the origin of the coordinates is common to the two systems, let a spherical wave be emitted therefrom, and be propagated with the velocity c in system K. If (x, y, z) be a point just attained by this wave, then

$$x^2 + y^2 + z^2 = c^2t^2. \tag{10.1.17}$$

Transforming this equation with the aid of our equations of transformation we obtain after a simple calculation

$$\xi^2 + \eta^2 + \zeta^2 = c^2\tau^2. \tag{10.1.18}$$

The wave under consideration is therefore no less a spherical wave with velocity of propagation c when viewed in the moving system. This shows that our two fundamental principles are compatible.[24]

In the equations of transformation which have been developed there enters an unknown function ϕ of v, which we will now determine.

For this purpose we introduce a third system of coordinates K', which relative to the system k is in a state of parallel translatory motion parallel to the axis of Ξ,[25] such that the origin of coordinates of system K' moves with velocity −v on the axis of Ξ. At the time t = 0 let all three origins coincide, and when t = x = y = z = 0 let the time t' of the system K' be zero. We call the coordinates, measured in the system K', x', y', z', and by a twofold application of our equations of transformation we obtain

$$
\begin{aligned}
t' &= \phi(-v)\beta(-v)\left(\tau + \frac{v\xi}{c^2}\right) &= \phi(v)\phi(-v)t, \\
x' &= \phi(-v)\beta(-v)(\xi + v\tau) &= \phi(v)\phi(-v)x, \\
y' &= \phi(-v)\eta &= \phi(v)\phi(-v)y, \\
z' &= \phi(-v)\zeta &= \phi(v)\phi(-v)z.
\end{aligned}
$$

Since the relations between x', y', z' and x, y, z do not contain the time t, the systems K and K' are at rest with respect to one another, and it is clear that the transformation from K to K' must be the identical transformation. Thus

$$\phi(v)\phi(-v) = 1. \tag{10.1.19}$$

We now inquire into the signification of ϕ(v). We give our attention to that part of the axis of Y of system k which lies between $\xi = 0$, $\eta = 0$, $\zeta = 0$ and $\xi = 0$, $\eta = l$, $\zeta = 0$. This part of the axis of Y is a rod moving perpendicularly to its axis with velocity v relative to system K. Its ends possess in K the coordinates

$$x_1 = vt, \quad y_1 = \frac{l}{\phi(v)}, \quad z_1 = 0 \tag{10.1.20}$$

and

$$x_2 = vt, \quad y_2 = 0, \quad z_2 = 0. \tag{10.1.21}$$

The length of the rod measured in K is therefore $\frac{l}{\phi}$(v); and this gives us the meaning of the function ϕ(v). From reasons of symmetry it is now evident that the length of a given rod moving perpendicularly to its axis, measured in the stationary system, must depend only on the velocity and not on the direction and the sense of the motion. The length of the moving rod measured in the stationary system does not change, therefore, if v and −v are interchanged. Hence follows that $\frac{l}{\phi}$(v) $= \frac{l}{\phi}$(−v), or

$$\phi(v) = \phi(-v). \tag{10.1.22}$$

It follows from this relation and the one previously found that ϕ(v) = 1, so that the transformation equations which have been found become

$$\tau = \beta \left(t - \frac{vx}{c^2} \right),$$
$$\xi = \beta(x - vt),$$
$$\eta = y,$$
$$\zeta = z,$$

where

$$\beta = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}. \tag{10.1.23}$$

Contributors

This page titled 10.1: Appendix A (Part 1) is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Benjamin Crowell via source content that was edited to the style and standards of the LibreTexts platform.

10.2: Appendix A (Part 2)

§ 4. Physical Meaning of the Equations Obtained in Respect to Moving Rigid Bodies and Moving Clocks

We envisage a rigid sphere[26] of radius R, at rest relative to the moving system k, and with its centre at the origin of coordinates of k. The equation of the surface of this sphere moving relative to the system K with velocity v is

$$\xi^2 + \eta^2 + \zeta^2 = R^2. \tag{10.2.1}$$

> **Note**
>
> That is, a body possessing spherical form when examined at rest.—AS

The equation of this surface expressed in x, y, z at the time t = 0 is

$$\frac{x^2}{\left(\sqrt{1 - \frac{v^2}{c^2}}\right)^2} + y^2 + z^2 = R^2. \tag{10.2.2}$$

A rigid body which, measured in a state of rest, has the form of a sphere, therefore has in a state of motion—viewed from the stationary system—the form of an ellipsoid of revolution with the axes

$$R\sqrt{1 - \frac{v^2}{c^2}}, R, R. \tag{10.2.3}$$

Thus, whereas the Y and Z dimensions of the sphere (and therefore of every rigid body of no matter what form) do not appear modified by the motion, the X dimension appears shortened in the ratio $1 : sqrt1 - \frac{v^2}{c^2}$, i.e., the greater the value of v, the greater the shortening. For v = c all moving objects—viewed from the "stationary" system—shrivel up into plane figures.[27] For velocities greater than that of light our deliberations become meaningless; we shall, however, find in what follows, that the velocity of light in our theory plays the part, physically, of an infinitely great velocity.

> **Note**
>
> That is, a body possessing spherical form when examined at rest.—AS

It is clear that the same results hold good of bodies at rest in the "stationary" system, viewed from a system in uniform motion.

Further, we imagine one of the clocks which are qualified to mark the time t when at rest relative to the stationary system, and the time τ when at rest relative to the moving system, to be located at the origin of the coordinates of k, and so adjusted that it marks the time τ. What is the rate of this clock, when viewed from the stationary system?

Between the quantities x, t, and τ, which refer to the position of the clock, we have, evidently, x = vt and

$$\tau = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}\left(t - \frac{vx}{c^2}\right). \tag{10.2.4}$$

Therefore,

$$\tau = t\sqrt{1 - \frac{v^2}{c^2}} = t - \left(1 - \sqrt{1 - \frac{v^2}{c^2}}\right)t \tag{10.2.5}$$

whence it follows that the time marked by the clock (viewed in the stationary system) is slow by $1 - \sqrt{1 - \frac{v^2}{c^2}}$ seconds per second, or—neglecting magnitudes of fourth and higher order—by $\frac{1}{2}\frac{v^2}{c^2}$.

From this there ensues the following peculiar consequence. If at the points A and B of K there are stationary clocks which, viewed in the stationary system, are synchronous; and if the clock at A is moved with the velocity v along the line AB to B, then on its

arrival at B the two clocks no longer synchronize, but the clock moved from A to B lags behind the other which has remained at B by $\frac{1}{2}\frac{tv^2}{c^2}$ (up to magnitudes of fourth and higher order), t being the time occupied in the journey from A to B.

It is at once apparent that this result still holds good if the clock moves from A to B in any polygonal line, and also when the points A and B coincide.

If we assume that the result proved for a polygonal line is also valid for a continuously curved line, we arrive at this result: If one of two synchronous clocks at A is moved in a closed curve with constant velocity until it returns to A, the journey lasting t seconds, then by the clock which has remained at rest the travelled clock on its arrival at A will be $\frac{1}{2}\frac{tv^2}{c^2}$ second slow. Thence we conclude that a spring-clock at the equator must go more slowly, by a very small amount, than a precisely similar clock situated at one of the poles under otherwise identical conditions.[28]

> **Note**
>
> Einstein specifies a spring-clock ("unruhuhr") because the effective gravitational field is weaker at the equator than at the poles, so a pendulum clock at the equator would run more slowly by about two parts per thousand than one at the north pole, for nonrelativistic reasons. This would completely mask any relativistic effect, which he expected to be on the order of $\frac{v^2}{c^2}$, or about 10^{-13}. In any case, it later turned out that Einstein was mistaken about this example. There is also a gravitational time dilation that cancels the kinematic effect. See example 10. The two clocks would actually agree.—BC

§ 5. The Composition of Velocities

In the system k moving along the axis of X of the system K with velocity v, let a point move in accordance with the equations

$$\xi = w_\xi \tau, \quad \eta = w_\eta \tau, \quad \zeta = 0, \tag{10.2.6}$$

where w_ξ and w_η denote constants.

Required: the motion of the point relative to the system K. If with the help of the equations of transformation developed in § 3 we introduce the quantities x, y, z, t into the equations of motion of the point, we obtain

$$x = \frac{w_\xi + v}{1 + \frac{vw_\xi}{c^2}} t,$$

$$y = \frac{\sqrt{1 - \frac{v^2}{c^2}}}{1 + \frac{vw_\xi}{c^2}} w_\eta t,$$

$$z = 0.$$

Thus the law of the parallelogram of velocities is valid according to our theory only to a first approximation. We set[29]

$$V^2 = \left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2,$$

$$w^2 = w_\xi^2 + w_\eta^2,$$

$$a = \tan^{-1}\frac{w_\eta}{w_\xi},$$

a is then to be looked upon as the angle between the velocities v and w. After a simple calculation we obtain

$$V = \frac{\sqrt{(v^2 + w^2 + 2vw\cos a) - (vw\sin\frac{a}{c})^2}}{1 + vw\cos\frac{a}{c^2}}. \tag{10.2.7}$$

> **Note**
>
> This equation was incorrectly given in Einstein's original paper and the 1923 English translation as $a = \tan^{-1}\frac{w_y}{w_x}$.—JW

It is worthy of remark that v and w enter into the expression for the resultant velocity in a symmetrical manner. If w also has the direction of the axis of X, we get

$$V = \frac{v+w}{1 + \frac{vw}{c^2}}. \tag{10.2.8}$$

It follows from this equation that from a composition of two velocities which are less than c, there always results a velocity less than c. For if we set v = c − κ, w = c − λ, κ and λ being positive and less than c, then

$$V = c\frac{2c - \kappa - \lambda}{2c - \kappa - \lambda + \frac{\kappa\lambda}{c}} < c. \tag{10.2.9}$$

It follows, further, that the velocity of light c cannot be altered by composition with a velocity less than that of light. For this case we obtain

$$V = \frac{c+w}{1 + \frac{w}{c}} = c. \tag{10.2.10}$$

We might also have obtained the formula for V, for the case when v and w have the same direction, by compounding two transformations in accordance with § 3. If in addition to the systems K and k figuring in § 3 we introduce still another system of coordinates k' moving parallel to k, its initial point moving on the axis of Ξ[30] with the velocity w, we obtain equations between the quantities x, y, z, t and the corresponding quantities of k', which differ from the equations found in § 3 only in that the place of "v" is taken by the quantity

$$\frac{v+w}{1 + \frac{vw}{c^2}}; \tag{10.2.11}$$

from which we see that such parallel transformations—necessarily—form a group.

> **Note**
>
> "X" in the 1923 English translation.—JW

We have now deduced the requisite laws of the theory of kinematics corresponding to our two principles, and we proceed to show their application to electrodynamics.[31]

> **Note**
>
> The remainder of the paper is not given here, but can be obtained from John Walker's web site at www.fourmilab.ch.—BC

Contributors and Attributions

This page titled 10.2: Appendix A (Part 2) is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Benjamin Crowell via source content that was edited to the style and standards of the LibreTexts platform.

10.3: Appendix A (Part 3)

"Does the Inertia of a Body Depend Upon its Energy Content?"

A. Einstein, Annalen der Physik. 18 (1905) 639.

The results of the previous investigation lead to a very interesting conclusion, which is here to be deduced.

I based that investigation on the Maxwell-Hertz equations for empty space, together with the Maxwellian expression for the electromagnetic energy of space, and in addition the principle that:—

> *The laws by which the states of physical systems alter are independent of the alternative, to which of two systems of coordinates, in uniform motion of parallel translation relative to each other, these alterations of state are referred (principle of relativity).*

With these principles[32] as my basis I deduced *inter alia* the following result (§ 8):—

Let a system of plane waves of light, referred to the system of coordinates (x, y, z), possess the energy l; let the direction of the ray (the wave-normal) make an angle ϕ with the axis of x of the system. If we introduce a new system of coordinates (ξ, η, ζ) moving in uniform parallel translation with respect to the system (x, y, z), and having its origin of coordinates in motion along the axis of x with the velocity v, then this quantity of light—measured in the system (ξ, η, ζ)—possesses the energy[33]

$$l^* = l \frac{1 - \frac{v}{c}\cos\phi}{\sqrt{1 - \frac{v^2}{c^2}}} \tag{10.3.1}$$

where c denotes the velocity of light. We shall make use of this result in what follows.

> **Notes**
>
> [32] The principle of the constancy of the velocity of light is of course contained in Maxwell's equations.—AS
>
> [33] See chapter 4 problem 11.—BC

Let there be a stationary body in the system (x, y, z), and let its energy—referred to the system (x, y, z) be E_0. Let the energy of the body relative to the system (ξ, η, ζ) moving as above with the velocity v, be H_0.

Let this body send out, in a direction making an angle ϕ with the axis of x, plane waves of light, of energy $\frac{1}{2}$L measured relative to (x, y, z), and simultaneously an equal quantity of light in the opposite direction. Meanwhile the body remains at rest with respect to the system (x, y, z). The principle of energy must apply to this process, and in fact (by the principle of relativity) with respect to both systems of coordinates. If we call the energy of the body after the emission of light E_1 or H_1 respectively, measured relative to the system (x, y, z) or (ξ, η, ζ) respectively, then by employing the relation given above we obtain

$$E_0 = E_1 + \frac{1}{2}L + \frac{1}{2}L,$$
$$H_0 = H_1 + \frac{1}{2}L\frac{1 - \frac{v}{c}\cos\phi}{\sqrt{1 - \frac{v^2}{c^2}}} + \frac{1}{2}L\frac{1 + \frac{v}{c}\cos\phi}{\sqrt{1 - \frac{v^2}{c^2}}}$$
$$= H_1 + \frac{L}{\sqrt{1 - \frac{v^2}{c^2}}}.$$

By subtraction we obtain from these equations

$$H_0 - E_0 - (H_1 - E_1) = L\left\{\frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} - 1\right\}. \tag{10.3.2}$$

The two differences of the form H − E occurring in this expression have simple physical significations. H and E are energy values of the same body referred to two systems of coordinates which are in motion relative to each other, the body being at rest in one of the two systems (system (x, y, z)). Thus it is clear that the difference H − E can differ from the kinetic energy K of the body, with

respect to the other system (ξ, η, ζ), only by an additive constant C, which depends on the choice of the arbitrary additive constants[34] of the energies H and E. Thus we may place

$$H_0 - E_0 = K_0 + C,$$
$$H_1 - E_1 = K_1 + C,$$

since C does not change during the emission of light. So we have

$$K_0 - K_1 = L\left\{\frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} - 1\right\}. \qquad (10.3.3)$$

> **Note**
>
> A potential energy U is only defined up to an additive constant. If, for example, U depends on the distance between particles, and the distance undergoes a Lorentz contraction, there is no reason to imagine that the constant will stay the same.—BC

The kinetic energy of the body with respect to (ξ, η, ζ) diminishes as a result of the emission of light, and the amount of diminution is independent of the properties of the body. Moreover, the difference $K_0 - K_1$, like the kinetic energy of the electron (§ 10), depends on the velocity.

Neglecting magnitudes of fourth and higher orders[35] we may place

$$K_0 - K_1 = \frac{1}{2}\frac{L}{c^2}v^2. \qquad (10.3.4)$$

> **Note**
>
> The purpose of making the approximation is to show that under realistic lab conditions, the effect exactly mimics a change in Newtonian mass.

From this equation it directly follows[36] that:—

If a body gives off the energy L in the form of radiation, its mass diminishes by $\frac{L}{c^2}$. The fact that the energy withdrawn from the body becomes energy of radiation evidently makes no difference, so that we are led to the more general conclusion that

The mass of a body is a measure of its energy-content; if the energy changes by L, the mass changes in the same sense by $\frac{L}{9} \times 10^{20}$, the energy being measured in ergs, and the mass in grammes.

It is not impossible that with bodies whose energy-content is variable to a high degree (e.g. with radium salts) the theory may be successfully put to the test.

If the theory corresponds to the facts, radiation conveys inertia between the emitting and absorbing bodies.

> **Note**
>
> The object has the same velocity v before and after emission of the light, so this reduction in kinetic energy has to be attributed to a change in mass.—BC

"The Foundation of the General Theory of Relativity"

A. Einstein, Annalen der Physik 49 (1916) 769.

[A one-page introduction relating to history and personalities is omitted.—BC]

A. Fundamental Considerations on the Postulate of Relativity

§1. Observations on the Special Theory of Relativity

The special theory of relativity is based on the following postulate, which is also satisfied by the mechanics of Galileo and Newton. If a system of coordinates K is chosen so that, in relation to it, physical laws hold good in their simplest form, the *same* laws also hold good in relation to any other system of coordinates K' moving in uniform translation relative to K. This postulate we call the "special principle of relativity." The word "special" is meant to intimate that the principle is restricted to the case when K' has a

motion of uniform translation[37] relative to K, but that the equivalence of K' and K does not extend to the case of non-uniform motion of K' relative to K.

> **Note**
>
> Here Einstein defines the distinction between special and general relativity according to whether accelerated frames of reference are allowed. The modern tendency is to pose this distinction in terms of flat versus curved spacetime, so that accelerated frames of reference in flat spacetime are considered to be part of special relativity. None of this has anything to do with the ability to describe accelerated *objects*. For example, special relativity is perfectly capable of describing the twin paradox.—BC

Thus the special theory of relativity does not depart from classical mechanics through the postulate of relativity, but through the postulate of the constancy of the velocity of light *in vacuo,* from which, in combination with the special principle of relativity, there follow, in the well-known way, the relativity of simultaneity, the Lorentzian transformation and the related laws for the behaviour of moving bodies and clocks.

The modification to which the special theory of relativity has subjected the theory of space and time is indeed far-reaching, but one important point has remained unaffected. For the laws of geometry, even according to the special theory of relativity, are to be interpreted directly as laws relating to the possible relative positions of solid bodies at rest; and, in a more general way, the laws of kinematics are to be interpreted as laws which describe the relations of measuring bodies and clocks. To two selected material points of a stationary rigid body there always corresponds a distance of quite definite length, which is independent of the locality and orientation of the body, and is also independent of the time. To two selected positions of the hands of a clock at rest relative to the privileged system of reference there always corresponds an interval of time of a definite length, which is independent of place and time. We shall soon see that the general theory of relativity cannot adhere to this simple physical interpretation of space and time.[38]

> **Note**
>
> Einstein is just starting to lay out his argument, and has not yet made clear in what sense these statements about location-independence of clocks and rulers could be empirically tested. It becomes more clear later that he means something like this. We could try to fill spacetime with a lattice of clocks and rulers, to synchronize the clocks, and to construct the lattice so that it consisted of right angles and equal-length line segments. This succeeds in special relativity, so that the geometry of spacetime is compatible with frames of reference that split up spacetime into 3+1 dimensions, where the three dimensions are Euclidean. The same prescription fails in general relativity.—BC

§2. The Need for an Extension of the Postulate of Relativity

In classical mechanics, and no less in the special theory of relativity, there is an inherent epistemological defect which was, perhaps for the first time, clearly pointed out by Ernst Mach. We will elucidate it by the following example:[39] — Two fluid bodies of the same size and nature hover freely in space at so great a distance from each other and from all other masses that only those gravitational forces need be taken into account which arise from the interaction of different parts of the same body. Let the distance between the two bodies be invariable, and in neither of the bodies let there be any relative movements of the parts with respect to one another.

> **Note**
>
> This example was described in section 3.6.—BC

But let either mass, as judged by an observer at rest relative to the other mass, rotate with constant angular velocity about the line joining the masses. This is a verifiable relative motion of the two bodies. Now let us imagine that each of the bodies has been surveyed by means of measuring instruments at rest relative to itself, and let the surface of S_1 prove to be a sphere, and that of S_2 an ellipsoid of revolution. Thereupon we put the question — What is the reason for this difference in the two bodies? No answer can be admitted as epistemologically satisfactory,[40] unless the reason given is an *observable fact of experience*. The law of causality has not the significance of a statement as to the world of experience, except when *observable facts* ultimately appear as causes and effects.

Newtonian mechanics does not give a satisfactory answer to this question. It pronounces as follows: — The laws of mechanics apply to the space R_1, in respect to which the body S_1 is at rest, but not to the space R_2, in respect to which the body S_2 is at rest. But the privileged space R_1 of Galileo, thus introduced, is a merely *factitious*[41] cause, and not a thing that can be observed. It is therefore clear that Newton's mechanics does not really satisfy the requirement of causality in the case under consideration but only apparently does so, since it makes the factitious cause R_1 responsible for the observable difference in the bodies S_1 and S_2.

> **Note**
>
> i.e., artificial —BC

The only satisfactory answer must be that the physical system consisting of S_1 and S_2 reveals within itself no imaginable cause to which the differing behaviour of S_1 and S_2 can be referred. The cause must therefore lie *outside* this system. We have to take it that the general laws of motion, which in particular determine the shapes of S_1 and S_2, must be such that the mechanical behaviour of S_1 and S_2 is partly conditioned in quite essential respects, by distant masses which we have not included in the system under consideration. These distant masses and their motions relative to S_1 and S_2 must then be regarded as the seat of the causes (which must be susceptible to observation) of the different behaviour of our two bodies S_1 and S_2. They take over the role of the factitious cause R_1. Of all imaginable spaces R_1, R_2, etc., in any kind of motion relative to one another there is none which we may look upon as privileged a priori without reviving the above-mentioned epistemological objection. *The laws of physics must be of such a nature that they apply to systems reference in any kind of motion.*[42] Along this road we arrive at an extension at the postulate of relativity.

> **Note**
>
> At this time, Einstein had high hopes that his theory would be fully Machian. He was already aware of the Schwarzschild solution (he refers to it near the end of the paper), which offended his Machian sensibilities because it imputed properties to spacetime in a universe containing only a single point-mass. In the present example of the bodies S_1 and S_2, general relativity actually turns out to give the non-Machian result which Einstein here says would be unsatisfactory.—BC

In addition to this weighty argument from the theory of knowledge, there is a well-known physical fact which favours an extension of the theory of relativity. Let K be a Galilean system of reference, i.e., a system relative to which (at least in the four-dimensional region under consideration) a mass, sufficiently distant from other masses, is moving with uniform motion in a straight line. Let K' be a second system of reference which is moving relative to K in *uniformly accelerated* translation. Then, relative to K', a mass sufficiently distant from other masses would have an accelerated motion such that its acceleration and direction of acceleration are independent of the material composition and physical state of the mass.

Does this permit an observer at rest relative to K' to infer that he is on a "really" accelerated system of reference? The answer is in the negative; for the above-mentioned relation of freely movable masses to K' may be interpreted equally well in the following way. The system of reference K' is unaccelerated, but the space-time territory in question is under the sway of a gravitational field, which generates the accelerated motion of the bodies relative to K'.

This view is made possible for us by the teaching of experience as to the existence of a field of force, namely, the gravitational field, which possesses the remarkable property of imparting the same acceleration to all bodies.[43] The mechanical behaviour of bodies relative to K' is the same as presents itself to experience in the case of systems which we are wont to regard as "stationary" or as "privileged." Therefore, from the physical standpoint, the assumption readily suggests itself that the systems K and K' may both with equal right be looked upon as "stationary" that is to say, they have an equal title as systems of reference for the physical description of phenomena.

> **Note**
>
> Eötvös has proved experimentally that the gravitational field has this property in great accuracy.—AS

It will be seen from these reflections that in pursuing the general theory of relativity we shall be led to a theory of gravitation, since we are able to "produce" a gravitational field merely by changing the system of coordinates. It will also be obvious that the

principle of the constancy of the velocity of light *in vacuo* must be modified, since we easily recognize that the path of a ray of light with respect to K' must in general be curvilinear, if with respect to K light is propagated in a straight line with a definite constant velocity.

§3. The Space-Time Continuum. Requirement of General Covariance for the Equations Expressing General Laws of Nature

In classical mechanics, as well as in the special theory of relativity, the coordinates of space and time have a direct physical meaning. To say that a point-event has the X_1 coordinate x_1 means that the projection of the point-event on the axis of X_1, determined by rigid rods and in accordance with the rules of Euclidean geometry, is obtained by measuring off a given rod (the unit of length) x_1 times from the origin of coordinates along the axis of X_1. To say that a point-event has the X_4 coordinate $x_4 = t$, means that a standard clock, made to measure time in a definite unit period, and which is stationary relative to the system of coordinates and practically coincident in space with the point-event,[44] will have measured off $x_4 = t$ periods at the occurrence of the event.

> **Note**
>
> We assume the possibility of verifying "simultaneity" for events immediately proximate in space, or — to speak more precisely — for immediate proximity or coincidence in space-time, without giving a definition of this fundamental concept.—AS

This view of space and time has always been in the minds of physicists, even if, as a rule, they have been unconscious of it. This is clear from the part which these concepts play in physical measurements; it must also have underlain the reader's reflections on the preceding paragraph for him to connect any meaning with what he there read. But we shall now show that we must put it aside and replace it by a more general view, in order to be able to carry through the postulate of general relativity, if the special theory of relativity applies to the special case of the absence of a gravitational field.

In a space which is free of gravitational fields we introduce a Galilean system of reference K (x, y, z,t), and also a system of coordinates K' (x', y', z', t') in uniform rotation[45] relative to K. Let the origins of both systems, as well as their axes of Z, permanently coincide. We shall show that for a space-time measurement in the system K' the above definition of the physical meaning of lengths and times cannot be maintained. For reasons of symmetry it is clear that a circle around the origin in the X, Y plane of K may at the same time be regarded as a circle in the X', Y' plane of K'. We suppose that the circumference and diameter of this circle have been measured with a unit measure infinitely small compared with the radius, and that we have the quotient of the two results. If this experiment were performed with a measuring-rod[46] at rest relative to the Galilean system K, the quotient would be π. With a measuring-rod at rest relative to K', the quotient would be greater than π. This is readily understood if we envisage the whole process of measuring from the "stationary" system K, and take into consideration that the measuring-rod applied to the periphery undergoes a Lorentzian contraction, while the one applied along the radius does not.[47] Hence Euclidean geometry does not apply to K'. The notion of coordinates defined above, which presupposes the validity of Euclidean geometry, therefore breaks down in relation to the system K'. So, too, we are unable to introduce a time corresponding to physical requirements in K', indicated by clocks at rest relative to K'. To convince ourselves of this impossibility, let us imagine two clocks of identical constitution placed, one at the origin of coordinates, and the other at the circumference of the circle, and both envisaged from the "stationary" system K. By a familiar result of the special theory of relativity, the clock at the circumference — judged from K — goes more slowly than the other, because the former is in motion and the latter at rest. An observer at the common origin of coordinates, capable of observing the clock at the circumference by means of light, would therefore see it lagging behind the clock beside him. As he will not make up his mind to let the velocity of light along the path in question depend explicitly on the time, he will interpret his observations as showing that the clock at the circumference "really" goes more slowly than the clock at the origin. So he will be obliged to define time in such a way that the rate of a clock depends upon where the clock may be.

> **Notes**
>
> [45] This example of a rotating frame of reference was discussed in section 3.5.—BC
>
> [46] Einstein implicitly assumes that the measuring rods are perfectly rigid, but it is not obvious that this is possible. This issue is discussed in section 3.5.—BC
>
> [47] As described in section 3.5, Ehrenfest originally imagined that the circumference of the disk would be reduced by its rotation. His argument was incorrect, because it assumed the ability to start the disk rotating when it had originally been at rest. The present paper marks the first time that Einstein asserted the opposite, that the circumference is increased.—BC

We therefore reach this result: — In the general theory of relativity, space and time cannot be defined in such a way that differences of the spatial coordinates can be directly measured by the unit measuring-rod, or differences in the time coordinate by a standard clock.

The method hitherto employed for laying coordinates into the space-time continuum in a definite manner thus breaks down, and there seems to be no other way which would allow us to adapt systems of coordinates to the four-dimensional universe so that we might expect from their application a particularly simple formulation of the laws of nature. So there is nothing for it but to regard all imaginable systems of coordinates, on principle, as equally suitable for the description of nature.[48] This comes to requiring that: —

> *The general laws of nature are to be expressed by equations which hold good for all the systems of coordinates, that is, are covariant with respect to any substitutions whatever (generally covariant).[49]*

Notes

[48] This is a conceptual leap, not a direct inference from the argument about the rotating frame. Einstein started thinking about this argument in 1912, and concluded from it that he should base a theory of gravity on non-Euclidean geometry. Influenced by Levi-Civita, he tried to carry out this project in a coordinate-independent way, but he failed at first, and for a while explored a theory that was not coordinate-independent. Only later did he return to coordinate-independence. It should be clear, then, that the link between the rotating-frame argument and coordinate-independence was not as clearcut as Einstein makes out here, since he himself lost faith in it for a while.—BC

[49] In this book I've used the more transparent terminology "coordinate independence" rather than "general covariance."—BC

It is clear that a physical theory which satisfies this postulate will also be suitable for the general postulate of relativity.[50] For the sum of all substitutions in any case includes those which correspond to all relative motions of three-dimensional systems of coordinates. That this requirement of general covariance, which takes away from space and time the last remnant of physical objectivity,[51] is a natural one, will be seen from the following reflection. All our space-time verifications invariably amount to a determination of space-time coincidences.[52] If, for example, events consisted merely in the motion of material points, then ultimately nothing would be observable but the meetings of two or more of these points. Moreover, the results of our measurings are nothing but verifications of such meetings of the material points of our measuring instruments with other material points, coincidences between the hands of a clock and points on the clock-dial, and observed point-events happening at the same place at the same time.

Notes

[50] For more on this point, see section 3.7.—BC

[51] This is an extreme interpretation of general covariance, and one that Einstein himself didn't hew closely to later on. He presented an almost diametrically opposed interpretation in a philosophical paper, "On the aether," Schweizerische naturforschende Gesellschaft 105 (1924) 85.—BC

[52] i.e., what this book refers to as incidence measurements (section 3.4)—BC

The introduction of a system of reference serves no other purpose than to facilitate the description of the totality of such coincidences. We allot to the universe four space-time variables x_1, x_2, x_3, x_4 in such a way that for every point-event there is a corresponding system of values of the variables $x_1 \ldots x_4$. To two coincident point-events there corresponds one system of values of the variables $x_1 \ldots x_4$, i.e., coincidence is characterized by the identity of the coordinates. If, in place of the variables $x_1 \ldots x_4$, we introduce functions of them, x'_1, x'_2, x'_3, x'_4, as a new system of coordinates, so that the systems of values are made to correspond to one another without ambiguity, the equality of all four coordinates in the new system will also serve as an expression for the space-time coincidence of the two point-events. As all our physical experience can be ultimately reduced to such coincidences, there is no immediate reason for preferring certain systems of coordinates to others, that is to say, we arrive at the requirement of general covariance.

Contributors and Attributions

This page titled 10.3: Appendix A (Part 3) is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Benjamin Crowell via source content that was edited to the style and standards of the LibreTexts platform.

10.4: Appendix C

Euclidean Geometry

E1 Two points determine a line.

E2 Line segments can be extended.

E3 A unique circle can be constructed given any point as its center and any line segment as its radius.

E4 All right angles are equal to one another.

E5 *Parallel postulate*: Given a line and a point not on the line, exactly one line can be drawn through the point and parallel to the given line.[1]

> **Note**
>
> This is a form known as Playfair's axiom, rather than the version of the postulate originally given by Euclid.

Ordered Geometry

O1 Two events determine a line.

O2 Line segments can be extended: given A and B, there is at least one event such that [ABC] is true.

O3 Lines don't wrap around: if [ABC] is true, then [BCA] is false.

O4 Betweenness: For any three distinct events A, B, and C lying on the same line, we can determine whether or not B is between A and C (and by statement 3, this ordering is unique except for a possible over-all reversal to form [CBA]).

Affine Geometry

In addition to O1-O4, postulate the following axioms:

A1 Constructibility of parallelograms: Given any P, Q, and R, there exists S such that [PQRS], and if P, Q, and R are distinct then S is unique.

A2 Symmetric treatment of the sides of a parallelogram: If [PQRS], then [QRSP], [QPSR], and [PRQS].

A3 Lines parallel to the same line are parallel to one another: If [ABCD] and [ABEF], then [CDEF].

Experimentally Motivated Statements about Lorentzian Geometry

L1 *Spacetime is homogeneous and isotropic.* No point has special properties that make it distinguishable from other points, nor is one direction distinguishable from another.

L2 *Inertial frames of reference exist.* These are frames in which particles move at constant velocity if not subject to any forces. We can construct such a frame by using a particular particle, which is not subject to any forces, as a reference point.

L3 *Equivalence of inertial frames*: If a frame is in constant-velocity translational motion relative to an inertial frame, then it is also an inertial frame. No experiment can distinguish one inertial frame from another.

L4 *Causality*: There exist events 1 and 2 such that $t_1 < t_2$ in all frames.

L5 *Relativity of time*: There exist events 1 and 2 and frames of reference (t, x) and (t', x') such that $t_1 < t_2$, but $t'_1 > t'_2$.

Statements of the Equivalence Principle

Accelerations and gravitational fields are equivalent. There is no experiment that can distinguish one from the other (section 1.5).

It is always possible to define a local Lorentz frame in a particular neighborhood of spacetime (section 1.5).

There is no way to associate a preferred tensor field with spacetime (section 4.4).

Vectors

Coordinates cannot in general be added on a manifold, so they don't form a vector space, but infinitesimal coordinate differences can and do. The vector space in which the coordinate differences exist is a different space at every point, referred to as the tangent space at that point (see section 7.1).

Vectors are written in abstract index notation with upper indices, x^a, and are represented by column vectors, arrows, or birdtracks with incoming arrows, $\rightarrow x$.

Dual vectors, also known as covectors or 1-forms, are written in abstract index notation with lower indices, x_a, and are represented by row vectors, ordered pairs of parallel lines (see section 2.1), or birdtracks with outgoing arrows, $\leftarrow x$.

In concrete-index notation, the x^μ are a list of numbers, referred to as the vector's contravariant components, while x_μ would be the covariant components of a dual vector.

Fundamentally the distinction between the two types of vectors is defined by the tensor transformation laws, section 4.3. For example, an odometer reading is contravariant because converting it from kilometers to meters increases it. A temperature gradient is covariant because converting it from degrees/km to degrees/m decreases it.

In the absence of a metric, every physical quantity has a definite vector or dual vector character. Infinitesimal coordinate differences dx^a and velocities $dx^a/d\tau$ are vectors, while momentum p_a and force F_a are dual (see section 4.3). Many ordinary and interesting real-world systems lack a metric (see section 2.1). When a metric is present, we can raise and lower indices at will. There is a perfect duality symmetry between the two types of vectors, but this symmetry is broken by the convention that a measurement with a ruler is a Δx^a, not a Δx_a.

For consistency with the transformation laws, differentiation with respect to a quantity flips the index, e.g., $\partial_\mu = \frac{\partial}{\partial x^\mu}$. The operators $\partial\mu$ are often used as basis vectors for the tangent plane. In general, expressing vectors in a basis using the Einstein notation convention results in an ugly notational clash described in section 7.1.

Index

Index

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: General Relativity (Crowell)

Webpages: 102

All licenses found:

- CC BY-SA 4.0: 83.3% (85 pages)
- CC BY-SA 3.0: 12.7% (13 pages)
- Undeclared: 3.9% (4 pages)

By Page