Chapter 1
A Geometrical Theory of Spacetime

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

Why does the world need yet another book about relativity? I had in mind the hope not necessarily of writing the best relativity book ever, but perhaps of writing the best *free* book at its level, where "at its level" means that the book is meant to be accessible to advanced undergraduates, as well as to graduate students who put their pants on one leg at a time.

This is mainly a book about general relativity, not special relativity. I've heard the sentiment expressed that books on special relativity generally do a lousy job on special relativity, compared to books on general relativity. This is undoubtedly true, for someone who already has already learned special relativity — but wants to unlearn the parts that are completely wrong in the broader context of general relativity. For someone who has *not* already learned special relativity, I strongly recommend mastering it first, from a book such as Taylor and Wheeler's *Spacetime Physics*. Even an advanced student may be able to learn a great deal from a masterfully written, nonmathematical treatment at an even lower level, such as the ones in Hewitt's, *Conceptual Physics* or the inexpensive paperback by Gardner, *Relativity Simply Explained*.

I should reveal at the outset that I am not a professional relativist. My field of research was nonrelativistic nuclear physics until I became a community college physics instructor. I can only hope that my pedagogical experience will compensate to some extent for my shallow background, and that readers who find mistakes will be kind enough to let me know about them using the contact information provided at `http://www.lightandmatter.com/area4author.html`.

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 t, 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.

Example: 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.

Example: If we had a time machine, 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.

Example: 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.[1]

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

[1] This point is revisited in section 5.1.

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

1.2.1 The Hafele-Keating experiment

In 1971, J.C. Hafele and R.E. Keating[2] 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. (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

[2] Hafele and Keating, Science, 177 (1972), 168

one gained $\Delta t_W = +273 \pm 7$ ns. This establishes that time is not universal and absolute.

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.

a / The atomic clock has its own ticket and its own seat.

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.[3] The signs of the Δt's show that moving clocks were slower.

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 the time's rate of flow also changes with height in a gravitational field. The deeper reasons for such an effect are given in section 1.5.8 on page 21.

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 similar effects have everyday consequences. The satellites of the Global Positioning System (GPS) orbit at a speed of 1.9×10^3 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 7 on page 34.

1.2.2 Muons

Although the Hafele-Keating experiment is impressively direct, it was not the first verification of relativistic effects on time, it did

[3] 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.

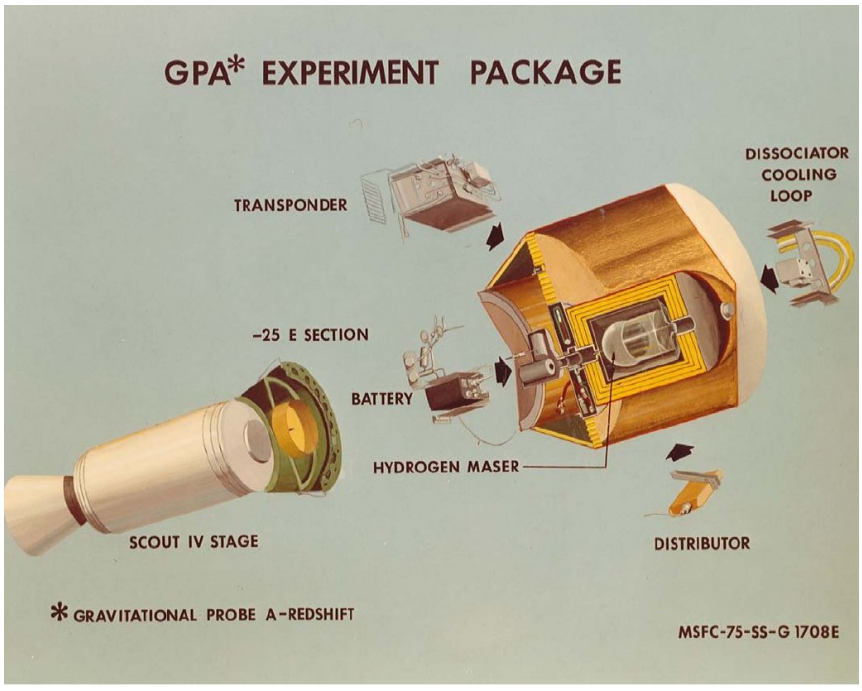
not completely separate the kinematic and gravitational effects, and the effect was small. In 1941, Rossi and Hall detected cosmic-ray muons at the summit and base of Mount Washington in New Hampshire. The muon has a mean lifetime of 2.2 μs, and the time of flight between the top and bottom of the mountain (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.

1.2.3 Gravitational redshifts

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 redshifted, 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 3 on page 82.

b / Gravity Probe A.



The first high-precision experiment of this kind was Gravity Probe A, a 1976 experiment[4] 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, and matched the relativistic prediction to an accuracy of 0.01%.

1.3 Non-simultaneity and the 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 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 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 propa-

[4] Vessot at al., Physical Review Letters 45 (1980) 2081

gate. It is also evident from these considerations that simultaneity itself cannot be a meaningful concept in relativity.

1.4 Ordered geometry

You've probably heard that general relativity is a geometrical theory. 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,[5] 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.[6]

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 self-evident 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

[5] These axioms are summarized for quick reference in the back of the book on page 183.

[6] This is a form known as Playfair's axiom, rather than the version of the postulate originally given by Euclid.

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.

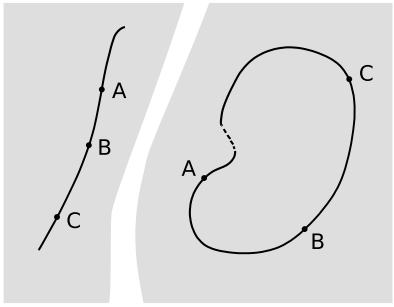
The stripped-down geometry we want is called *ordered geometry*, and was developed by Moritz Pasch around 1882. As suggested by the parable of Alice and Betty, ordered geometry does not have any global, all-encompassing system of measurement. When Betty goes on her trip, she traces out a particular path 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:[7]

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.

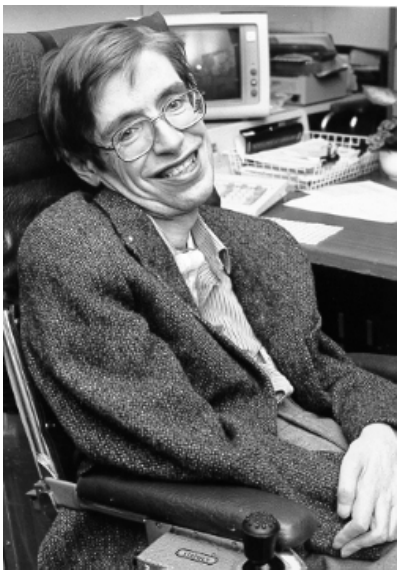


c / Axioms O2 (left) and O3 (right).

[7] The axioms are summarized for convenient reference in the back of the book on page 183. 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 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.

O4 Causality: Any three distinct events A, B, and C lying on the same line can be sorted out in order (and by statement 3, this order is unique).

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. 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 . . . "



d / Stephen Hawking (1942-).

If you're dismayed by the austerity of a system of geometry without any notion of measurement, you may be even more dismayed 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 statement 2. 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.

1.5 The equivalence principle

1.5.1 Proportionality of inertial and gravitational mass

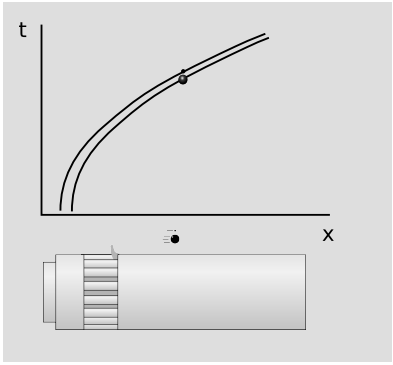
What physical interpretation should we give to the "lines" described in ordered geometry? Galileo described an experiment (which he may or may not have actually performed) in which he simultaneously dropped a cannonball and a musket ball from a tall tower. The two objects hit the ground simultaneously, 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 $a = F/m$, 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 act-

ing 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.

1.5.2 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 objects 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, 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 *geodesics*. The world-line of an object acted on only by gravity is one type of geodesic.

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.[8] 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.



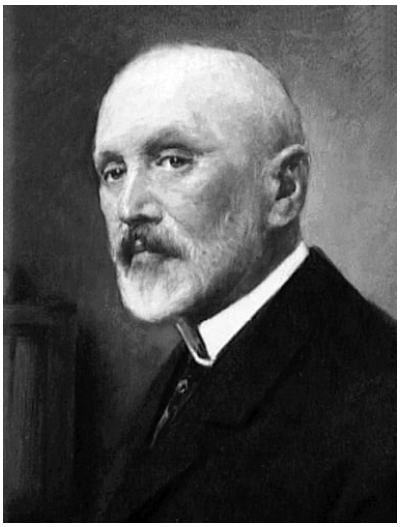
e / The cannonball and the musketball have identical parabolic world-lines. On this type of space-time plot, space is conventionally shown on the horizontal axis, so the tower has to be depicted on its side.

f / 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 e appear curved, but they are the analogs of straight lines in the non-Euclidean geometry used to describe gravitational fields in general relativity.

[8] Regarding infinitesimals, see p. 61.

1.5.3 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. Measurements in the mid-twentieth century refined this to one part in 10^{12}. These are generically referred to as Eötvös experiments.



g / Loránd Eötvös (1848-1919).

Figure i shows a strategy for doing Eötvös experiments that has proved successful. The top panel is a simplified. 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. The bottom panel shows a more realistic drawing of an experiment by Braginskii and Panov.[9] 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.

h / 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.

1.5.4 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 at various times. Suppose that the first explanation is correct. The capsule is initially at rest

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

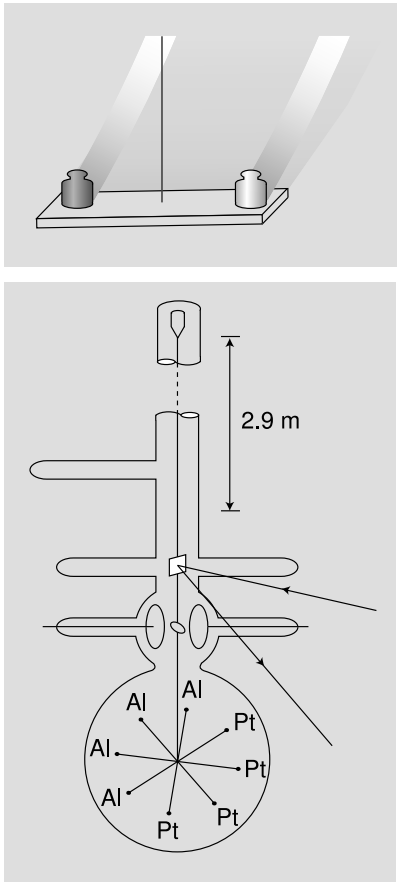
in outer space, where there is no gravity. Betty can release a pencil and a lead ball in the air inside the cabin, and they will stay in place. The capsule then accelerates, and to Betty, who has adopted a frame of reference tied to its deck, ceiling and walls, it appears that the pencil and the ball 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.

1.5.5 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.[10]

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."[11] 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.

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

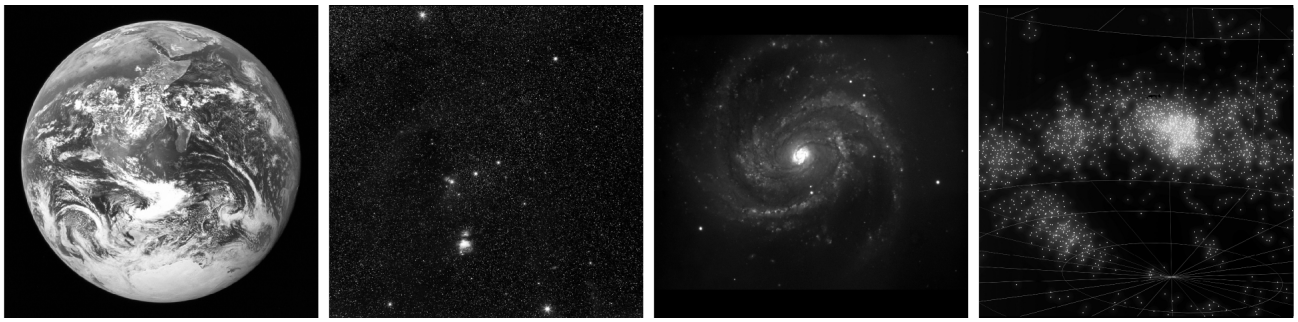


i / An Eötvös experiment. Top: simplified version. Bottom: realistic version by Braginskii and Panov. (Drawing after Braginskii and Panov.)

[10] This statement of the equivalence principle is summarized, along with some other forms of it to be encountered later, in the back of the book on page 184.

[11] paraphrased from a translation by Motte, 1729

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 process 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.



j / Wouldn't it be nice if we could define the meaning of an 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.

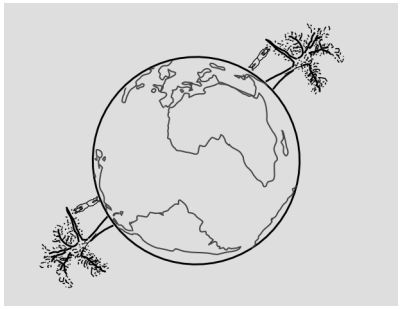
The conclusion is that we need to abandon the entire distinction between inertial and noninertial frames of reference. The best that we can do is to single out certain frames of reference defined by the motion of objects that are not subject to any nongravitational forces. A falling rock defines such a frame of reference. In this frame, the rock is at rest, and the floor is accelerating. The rock's world-line

is a straight line of constant x 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.

1.5.6 Locality of Lorentz frames

It would be nice if we could define a single Lorentz frame that would cover the entire universe, but we can't. In figure k, 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 $2g$.



k / 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.[12] It is not possible to do so on a universal basis.

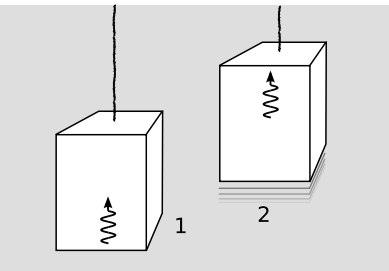
The locality of Lorentz frames can be understood in the analogy of the string stretched across the globe. We don't notice the curvature of the Earth's surface in everyday life because the radius of curvature is thousands of kilometers. On a map of LA, we don't notice any curvature, nor do we detect it 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.

1.5.7 Gravitational redshifts

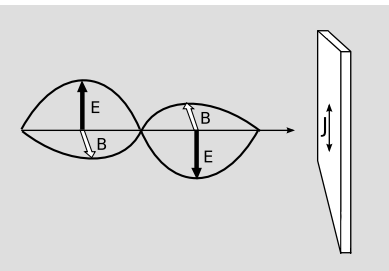
Starting on page 9, we saw experimental evidence that the rate of flow of time changes with height in a gravitational field. We can now see that this is required by the equivalence principle.

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. 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

[12] This statement of the equivalence principle is summarized, along with some other forms of it, in the back of the book on page 184.



l / 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.



m / An electromagnetic wave strikes an ohmic surface. The wave's electric field excites an oscillating current density **J**. The wave's magnetic field then acts on these currents, producing a force in the direction of the wave's propagation. This is a pre-relativistic argument that light must possess inertia.

$\Delta E/E = \Delta f/f = ay/c^2$, where a is the lab's acceleration, y is the height from floor to ceiling, and c is the speed of light.

Self-check: Verify this statement.

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 $\Delta E/E = gy/c^2$. Since the quantity gy is the 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 m, 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 had inertia must also participate in gravitational interactions. Therefore light waves must have weight, and the must lose energy when they rise through a gravitational field.

Self-check: Verify the application of the right-hand rule described above.

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 3.2.2.
- The relation $p = E/c$ 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 3.2.2 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 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.
- 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.

1.5.8 The Pound-Rebka experiment

The 1959 Pound-Rebka experiment at Harvard[13] was the first quantitative test of this effect to be carried out under controlled conditions, and in this section we will discuss it in detail.

When y is on the order of magnitude of the height of a building, the value of $\Delta E/E = gy/c^2$ is $\sim 10^{-14}$, so an extremely high precision experiment is necessary in order to detect a gravitational redshift. A number of other effects are big enough to obscure it entirely, and must be eliminated or compensated for somehow. 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. — $\sim 10^{-6}$

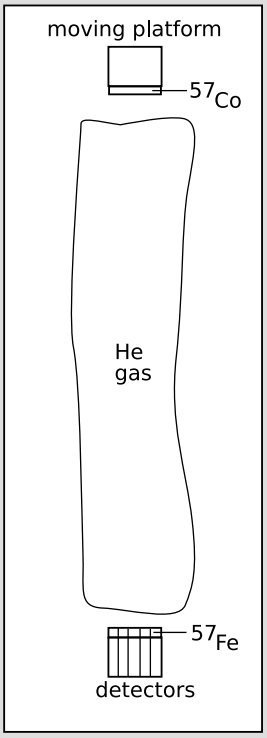
(2) *The recoil Doppler shift.* When an atom emits a photon with energy E and momentum $p = E/c$, conservation of momentum requires that the atom recoil with momentum $p = -E/c$ and energy $p^2/2m$. This causes a downward Doppler shift of the energy of the emitted photon. A similar effect occurs on absorption, 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 $\sim h/\tau$, where h is Planck's constant. — $\sim 10^{-12}$

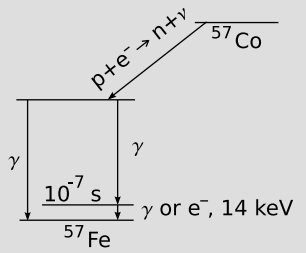
(4) *Special-relativistic Doppler shift due to temperature.* Section 1.2 presented experimental evidence that time flows at a different rate depending on the motion of the observer. Therefore the thermal motion of an atom emitting a photon has an effect on the frequency of the photon, even if the atom's motion is not along the line of emission. The equations needed in order to calculate this effect will not be derived until section 1.7; a quantitative estimate is given in example 8 on page 35. 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 (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 Avo-

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



n / The Pound-Rebka experiment.



o / 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.

gadro'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 time-scale 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.

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 $p^2/2m$, 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 negligible recoil. 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 experiment was carried out in a horizontal plane, then resonant absorption would occur. When the source and absorber are aligned vertically, 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.

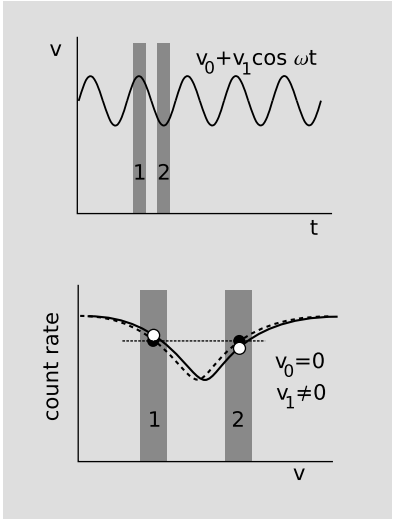
The typical half-life for deexcitation of a nucleus by emission of a gamma-ray with energy E is in the nanosecond range. To measure an gravitational effect at the 10^{-14} level, one would like to have a natural line width, (3), with $\Delta E/E \lesssim 10^{-14}$, which would require a half-life of $\gtrsim 10$ μs. 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 $\Delta E/E \gtrsim 10^{-12}$. As a result, they settled on an experiment in

which 14 keV gammas were emitted by ^{57}Fe nuclei 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 $\Delta E/E \sim 10^{-12}$. 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. 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 $1/\sqrt{\text{count rate}} \propto y$. 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.

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 shoulder of the resonance curve, where the slope of the curve was large. The resulting asymmetry in count rates is shown in figure p. 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.



p / *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.

q / Pound and Rebka at the top and bottom of the tower.

r / Hendrik Antoon Lorentz (1853-1928)

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%.

1.6 Affine properties of Lorentz geometry

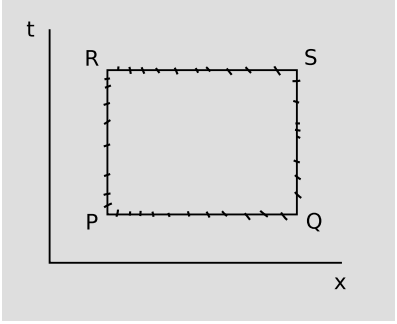
The geometrical treatment of space, time, and gravity only requires as its basis the equivalence of inertial and gravitational mass. That equivalence holds for Newtonian gravity, so it is indeed possible to redo Newtonian gravity as a theory of curved spacetime. This project was carried out by the French mathematician Cartan, as summarized very readably in section 17.5 of *The Road to Reality* by Roger Penrose. 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.

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.

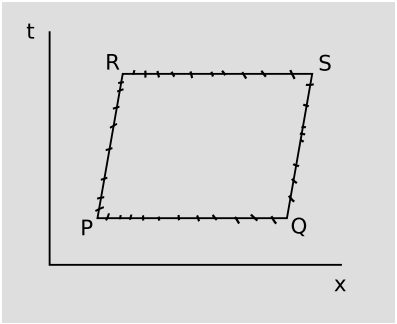
To see how to proceed, let's consider the bare minimum of geometrical apparatus that would be necessary in order to set up frames of reference. The following argument shows that the main missing ingredient is merely a concept of parallelism. We only expect Lorentz frames to be local, but we do need them 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.



s / 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.



t / There is no well-define 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 the third panel of the figure. We do expect, however, that it will remain a parallelogram; a Lorentz frame is one in which the gravitational field, if any, is con-

stant, so the properties of spacetime are uniform, and by symmetry the new frame should still have PR=QS and PQ=RS.

With this motivation, we form the system of *affine geometry* by adding the following axioms to set O1-O4.[14] 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].

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.[15]

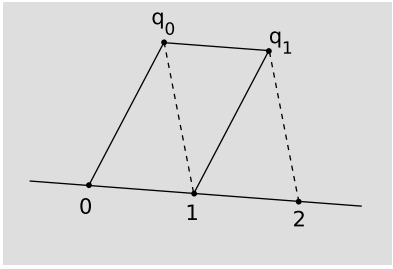
Although these new axioms do nothing more than to introduce the concept of parallelism lacking in ordered geometry, it turns out that they also allow us to build up a concept of measurement. Let ℓ be a line, and suppose we want to define a number system on this line that measures how far apart events 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

[14] The axioms are summarized for convenient reference in the back of the book on page 183. This formulation is essentially the one given by Penrose, *The Road to Reality*, in section 14.1.

[15] 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 $\parallel$ AB $\parallel$ m $\parallel$ BQ. By A3, lines parallel to another line are parallel to each other, so PQ $\parallel$ BQ, but this is a contradiction, since they have Q in common.

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 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.

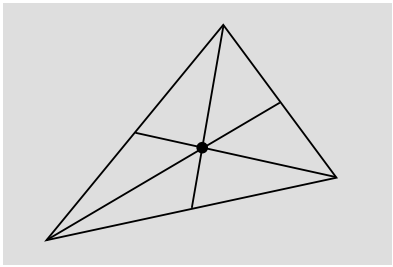
The continuously varying variable constructed in this way is called an *affine parameter*. The time measured by a free-falling clock is an example of an affine parameter, as is the distance measured by the tick marks on a free-falling ruler. Since light rays travel along geodesics, the wave crests on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



u / Construction of an affine parameter.

Centroids *Example: 1*

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.



v / Affine geometry gives a well-defined centroid for the triangle.

Conservation of momentum *Example: 2*

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:

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.

We cannot compare distances along two arbitrarily chosen lines, only along a single line or two parallel lines.

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.

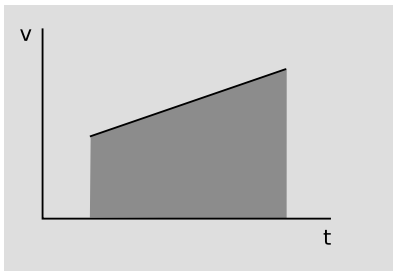
We will eventually want to lift some of these restrictions by adding to our kit a tool called a metric, which allows us to define distances along arbitrary curves in space time, and to compare distances in different directions. The affine parameter, however, will not be entirely superseded. In particular, 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.

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.

w / Example 3. 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.

Area and volume *Example: 3*

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.



x / Example 4.

Area on a graph of v versus t *Example: 4*

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.

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.

1.7 Relativistic properties of Lorentz geometry

We now want to pin down the properties of the Lorentz geometry that are left unspecified by the affine treatment. This can approached either by looking for an appropriate metric, or by finding the appropriate rules for distorting parallelograms when switching from one frame of reference to another frame is in motion relative to the first. In either case, 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:[16]

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:* Observers in different inertial frames agree on the time-ordering of events.

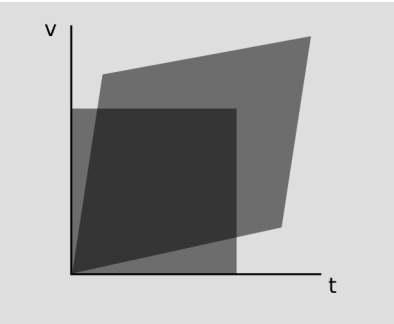
L5 *No simultaneity:* The experimental evidence in section 1.2 shows that observers in different inertial frames do not agree on the simultaneity of events.

Define affine parameters t and x for time and position, and construct a (t, x) plane. 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.

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.

[16] These facts are summarized for convenience on page 183 in the back of the book.

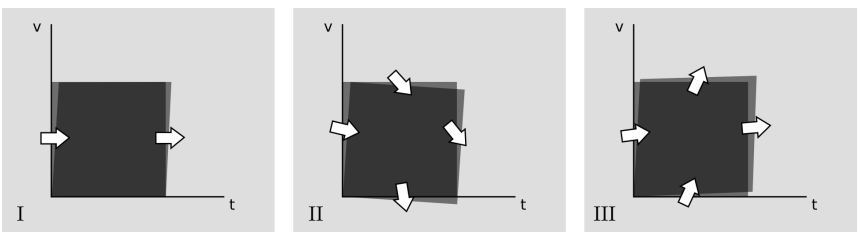


y / Two objects at rest have world-lines that define a rectangle. In a second frame of reference in motion relative to the first one, the rectangle becomes a parallelogram.

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.

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 $A(v) = A(-v)$, $A(v) = \pm 1$ 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. The argument is also purely one about affine geometry (it would apply equally well to a Euclidean space), so there is no reason to expect the area A in the (t, x) plane to have any special physical significance in relativity; it is simply a useful mathematical tool in the present discussion.

If we consider a boost by an infinitesimal velocity $\mathrm{d}v$, 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 $\mathrm{d}v^2$.) 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:



z / 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.[17] 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.

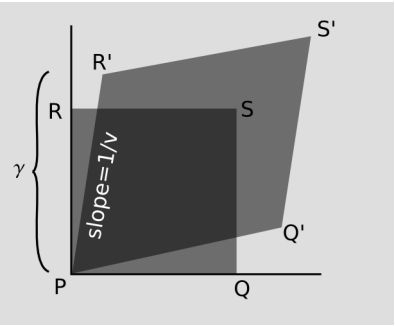
Only case III is possible, and given case III, there must be at least one point P in the first quadrant where area flows neither clockwise nor counterclockwise.[18] The boost simply increases P's distance from the origin by some factor. By the linearity of the transformation, the entire line running 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 $-c$.

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

[17] 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 $\mathbf{r} \times \mathrm{d}\mathbf{r}$, where $\mathrm{d}\mathbf{r}$ is the displacement caused by $L(\mathrm{d}v)$, we find that it depends on $\sin(2\theta+\delta)$, which does not oscillate rapidly enough to have two zeroes in the same quadrant.

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



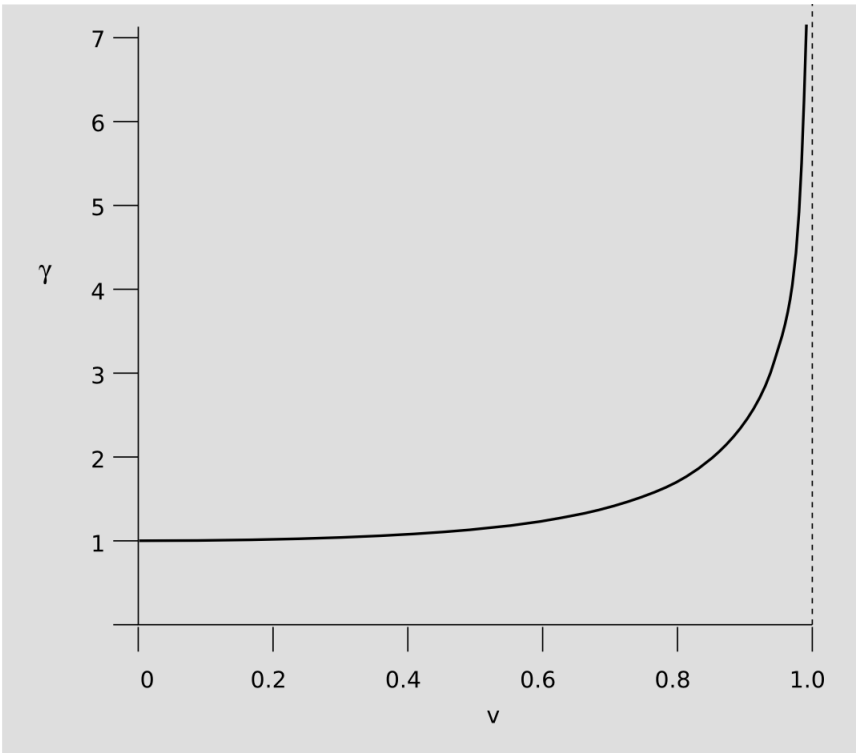
aa / Unit square PQRS is Lorentz-boosted to the parallelogram P′Q′R′S′.

ab / The behavior of the γ factor.

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

$$\gamma = \frac{1}{\sqrt{1 - v^2}} \quad .$$

Self-check: Interpret the dependence of γ on the sign of v.



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

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

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. Although we defined γ in terms of the time coordinate of R′, we could just as easily have used the spatial coordinate of Q′, so γ represents a factor of both time dilation and length contraction. (Clearly it wouldn't have made sense to

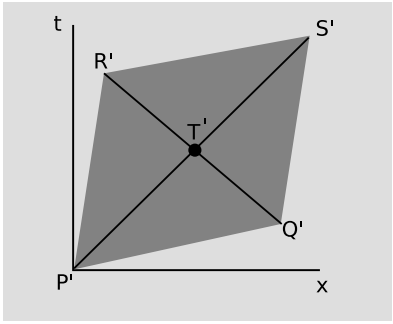
distort one quantity without distorting the other, since the invariant velocity c represents a ratio of a distance to a time.) In summary, a clock runs fastest according to an observer who is at rest relative to the clock, and a measuring rod likewise appears longest in its own rest frame.

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 $v \ll 1$, 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 on page 39 that photons must travel at c, and on page 83 that this must be true for any massless particle.

Example: 5

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.



ac / Example 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.

Example: 6

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 non-accelerating observers.

GPS *Example: 7*

In the Hafele-Keating experiment using atomic clocks aboard airplanes (8), both gravity and motion had effects on the rate of flow of time. Similarly, both effects must be considered in the case of the GPS system. The gravitational effect was found on page 20 to be $\Delta E/E = gy$ (with $c = 1$), based on the equivalence principle. The special-relativistic effect can be found from the Lorentz transformation. 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 $v^2/r = g$, where r is only slightly greater than the radius of the earth. Expanding the Lorentz gamma factor in a Taylor series, we find that the relative effect on the flow of time is $\gamma - 1 \approx v^2/2 = gr/2$. 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 $2y/r$, 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 $\Delta E/E = gy$ to the case where g is not a constant. Integrating the equation $dE/E = g dy$, we find that the time dilation factor is equal to $e^{\Delta\Phi}$, where $\Phi = \int g dy$ is the gravitational potential per unit mass. When $\Delta\Phi$ is small, this causes 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)

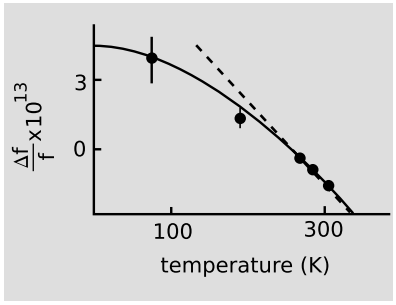
$$\frac{1}{c^2}\left(+\Delta\Phi - \frac{v^2}{2}\right) = 5.2\times10^{-10} - 0.9\times10^{-10} \quad ,$$

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.[19]

Time dilation in the Pound-Rebka experiment *Example: 8*

In the description of the Pound-Rebka experiment on page 21, I postponed the quantitative estimation of the frequency shift due to temperature. Classically, one expects only a *broadening* of the line, since the Doppler shift is proportional to $v_{\|}/c$, where $v_{\|}$, 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,[20] they refer to a preliminary measurement[21] 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



ad / 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.

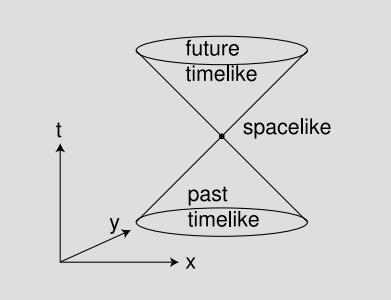
[19] N. Ashby, "Relativity in the Global Positioning System," http://www.livingreviews.org/lrr-2003-1

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

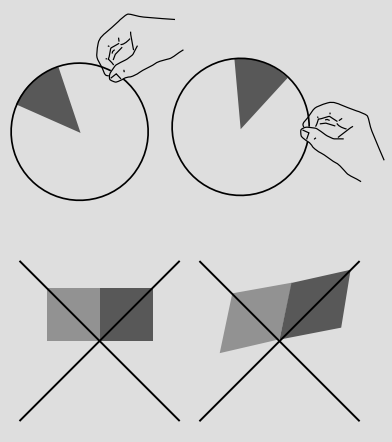
[21] Phys. Rev. Lett. 4 (1960) 274

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 $(3/2)kT$. The velocity is much less than $c(=1)$, so we can use the nonrelativistic expression for kinetic energy, $K = (1/2)mv^2$, which gives a mean value for v^2 of $3kT/m$. In the limit of $v \ll 1$, time dilation produces a change in frequency by a factor of $1/\gamma$, which differs from unity by approximately $-v^2/2$. The relative time dilation is therefore $-3kT/2m$, or, in metric units, $-3kT/2mc^2$. The vertical scale on the graph contains an arbitrary offset, since Pound and Rebka's measurements were the best absolute measurements to date of the frequency. The predicted slope of $-3k/2mc^2$, 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.



ae / The light cone in 2+1 dimensions.



af / 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 Lorentzian geometry. Under a Lorentz boost, the spacetime parallelograms change, but the light cone doesn't.

1.8 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 Lorentzian 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.

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.

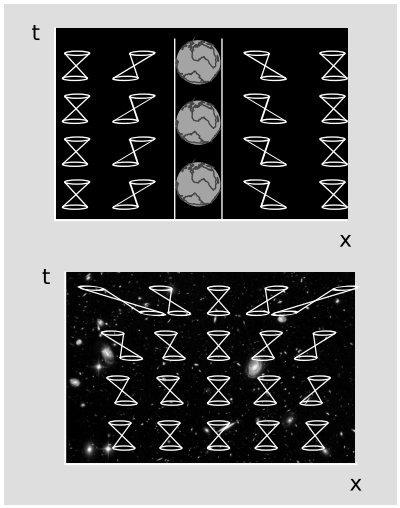
The light cone plays the same role in the Lorentzian geometry that the circle plays in Euclidean geometry. The truth or falsehood of propositions in Euclidean geometry remains the same regardless

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.

Self-check: Under what circumstances is the time-ordering of events P and Q preserved under a Lorentz boost?

In a uniform Lorentzian 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.

Self-check: In the bottom panel of figure ag, 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 k on page 19.

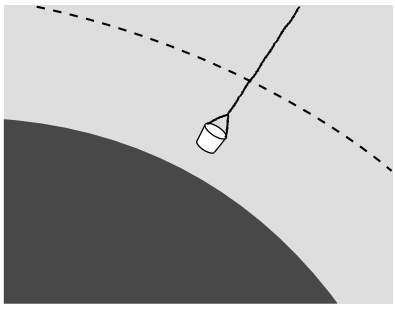


ag / Light cones tip over for two reasons in general relativity: because of the presence of masses, which have gravitational fields, and because of the cosmological constant. The time and distance scales in the bottom figure are many orders of magnitude greater than those in the top.

A Newtonian black hole *Example: 9*

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 $M/r > c^2/2G$. (A fully relativistic argument, as given in section 5.2, agrees on $M/r \propto c^2/G$, 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



ah / 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.

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 $c^2/2$, which is about ten orders of magnitude greater than that of carbon nanotube fibers. (The factor of $1/2$ is not to be taken seriously, since it comes from a nonrelativistic calculation.) The cause-and-effect interpretation of relativity tells us that this is incorrect. A physical object that approaches to within a distance r of a concentration of mass M, with 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. Cf. homework problem 1 and section 2.4.4.

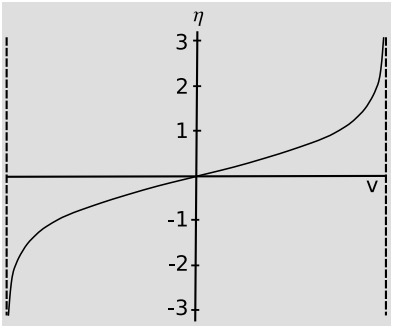
1.8.1 Velocity addition

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 be rewriting the Lorentz transformation as the matrix

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

where $\eta = \tanh^{-1} v$ 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 which is also of this form, with $\eta = \eta_1 + \eta_2$. In other words, rapidities add linearly; velocities don't. In the example of the spaceship and the probe, the rapidities add as $\tanh^{-1}.60 + \tanh^{-1}.60 = .693 + .693 = 1.386$, 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, $\eta_1 + \eta_2 + \ldots + \eta_n$.

Self-check: Interpret the asymptotes of the graph in figure ai.



ai / The rapidity, $\eta = \tanh^{-1} v$, as a function of v.

1.8.2 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 $p \wedge q$ (both p and

q) or $p \vee q$ (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 $p \wedge q$ cannot be checked by physically traveling to P and Q, because no observer would be able to attend both events. The truth-value of $p \wedge q$ 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 $p \wedge q$ 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 observes 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_2, ... of all the other propositions that they intend to test as well. We'll return to these questions on page 153.

1.9 Experimental tests of Lorentzian geometry

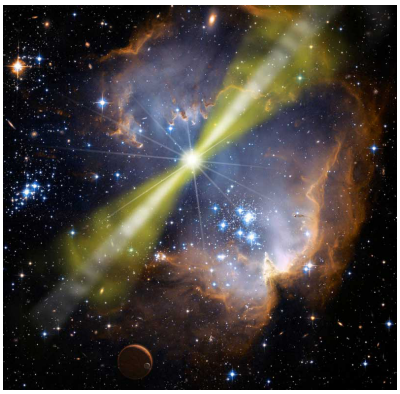
We've already seen, in section 1.2, a variety of evidence for the nonclassical behavior of spacetime. We're now in a position to discuss tests of relativity more quantitatively.

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 relative to a medium, then it must be fixed relative to any observer. The only speed in relativity that is equal in all frames of reference is c, so light must propagate at c. We will see on page 83 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.

1.9.1 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, $\sqrt{\hbar G/c^3} = 10^{-35}$ m, which would naturally 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.



aj / An artist's conception of a gamma-ray burst, resulting from a supernova explosion.

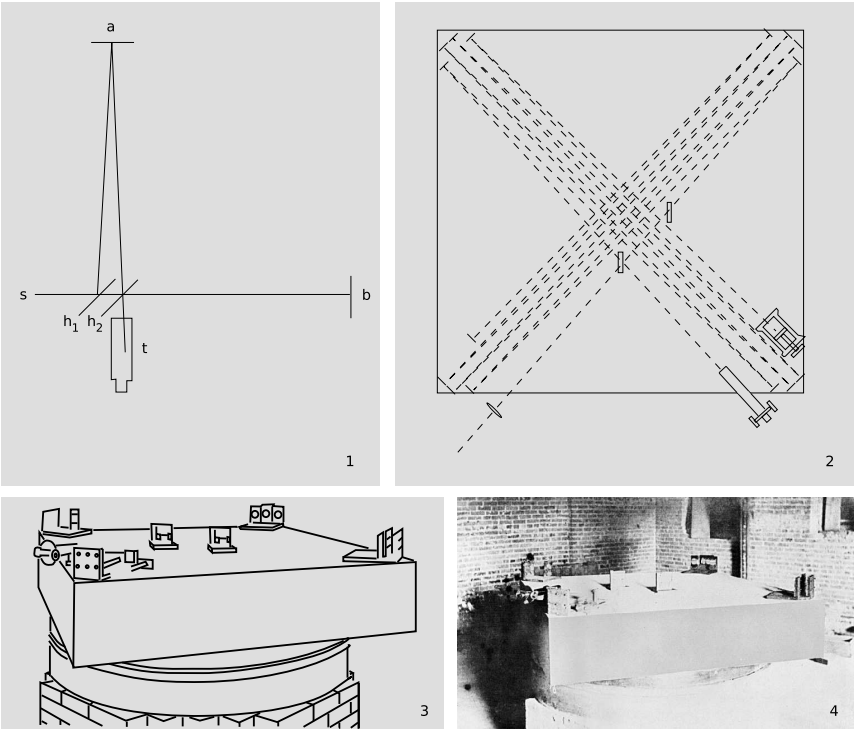
Presently the best experimental tests of the invariance of the speed of light with respect to wavelength come from astronomical observations of gamma-ray bursts, which are sudden outpourings of high-energy photons, believed to originate from a supernova explosion in another galaxy. One such observation, in 2009,[22] 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, ruling out some versions of loop quantum gravity.

1.9.2 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.

Although the Michelson-Morley experiment was nearly two decades in the past by the time Einstein published his first paper on

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



ak / 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 h_1. 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.

relativity in 1905, and Einstein did know about it,[23] 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.

Before Einstein, some physicists who did believe the negative result of the Michelson-Morley experiment came up with explana-

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

al / 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.

tions 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.

1.10 Three spatial dimensions

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

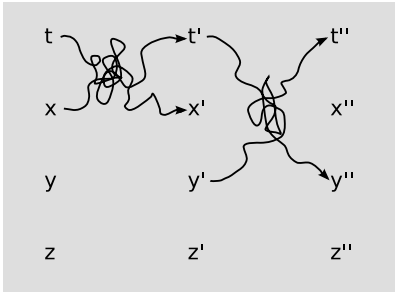
1.10.1 Lorentz boosts in three dimensions

How does a Lorentz boost along one axis, say x, affect the other two spatial coordinates y and z? 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 $L(v\hat{\mathbf{x}})$, a Lorentz boost along the x axis with velocity v, is therefore:

$$\begin{aligned} t' &= \gamma t + v\gamma x \\ x' &= v\gamma t + \gamma x \\ y' &= y \\ z' &= z \end{aligned}$$

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 am. 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.



am / 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.

1.10.2 Gyroscopes and the equivalence principle

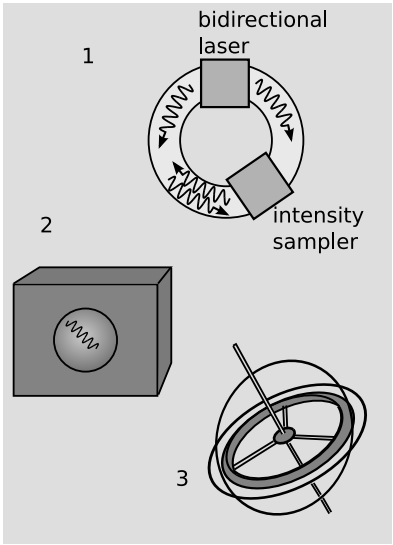
To see how this mathematical fact would play out as a physical effect, we need to consider how to make a physical manifestation of the concept of a direction in space.

In two space dimensions, we can construct a ring laser, which in its simplest incarnation is a closed loop of optical fiber with a bidirectional laser inserted in one place. 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 beam-sampling 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. The loop senses its own angular acceleration 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 described on page 8 were sensitive to Sagnac effects, and it was not practical to maintain their orientations while they were strapped into seats on a passenger jet, so this orientational effect had to be subtracted out of the data at the end of the experiment.

In three spatial dimensions, we could build a spherical cavity with a reflective inner surface, and release a photon inside.

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. For example, one of the classic tests of general relativity is the 2007 Gravity Probe B experiment (discussed in detail on pages 109 and 140), 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.[24] The equivalence



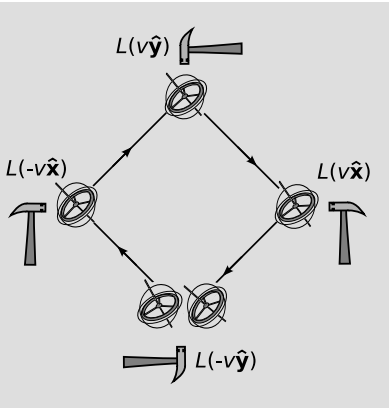
an / Inertial devices for maintaining a direction in space: 1. A ring laser. 2. The photon in a perfectly reflective spherical cavity. 3. A gyroscope.

[24] This is done, for example, in Misner, Thorne, and Wheeler, *Gravitation*, pp. 157-159.

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. Classically, 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 in section 1.10.1 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.

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 on page 88), 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 the two gyroscopes 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.

1.10.3 Boosts causing rotations



ao / Classically, the gyroscope should not rotate as long as the forces from the hammer are all transmitted to it at its center of mass.

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 ao 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 gyroscopes at their centers center of mass, so that they do not result in any torque. Classically, 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.

For notation, let $L(v\hat{\mathbf{x}})$ indicate the boost along the x axis described by the transformation on page 42. This is a transformation that changes to a frame of reference moving in the *negative* x direc-

tion 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 $\mathbf{p}$, we calculate $L\mathbf{p}$, which gives the coordinates of the event as measured in the new frame. An expression like $ML\mathbf{p}$ is equivalent by associativity to $M(L\mathbf{p})$, 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,

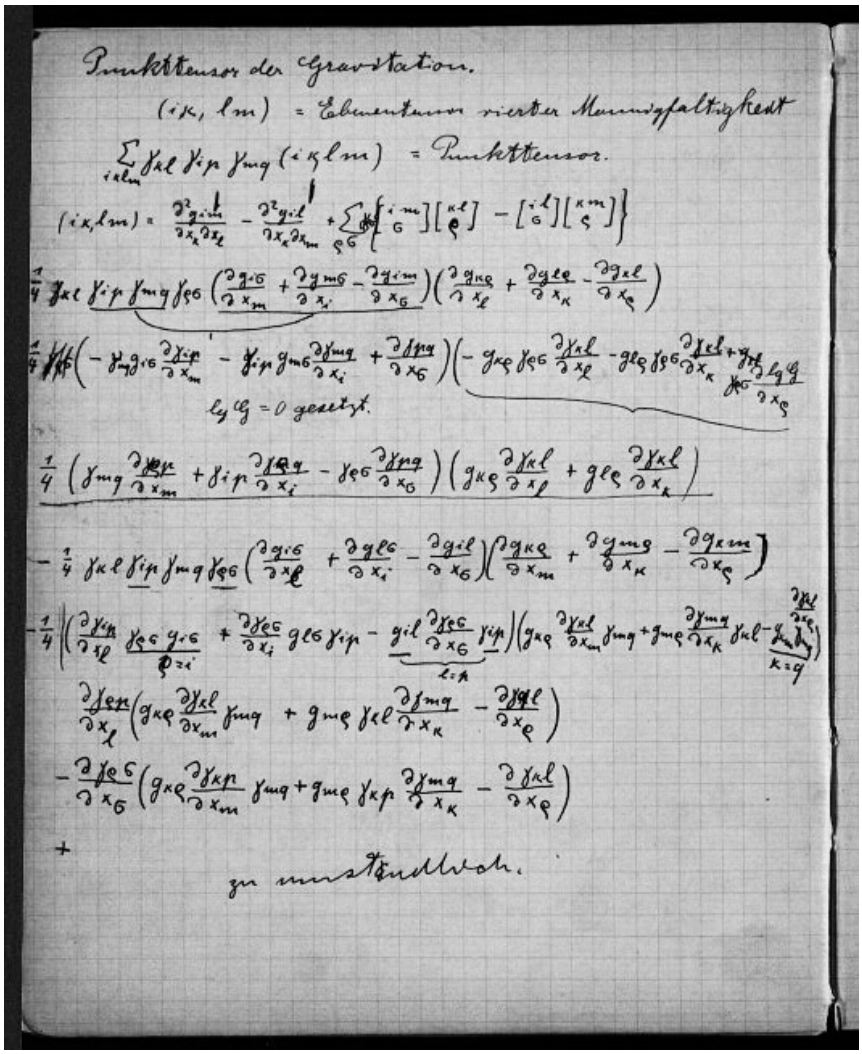
$$T = L(v\hat{\mathbf{x}})\, L(v\hat{\mathbf{y}})\, L(-v\hat{\mathbf{x}})\, L(-v\hat{\mathbf{y}}) \qquad ,$$

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 $L(\ldots)$ to describe a transformation that does not preserve the origin.) The first transformation, $L(-v\hat{\mathbf{y}})$, 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.

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 ap 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." 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. Each
4     is represented by a matrix: */
5  Lx:matrix([gamma,   gamma*v,  0],
6            [gamma*v,  gamma,    0],
7            [0,        0,        1]);
8  Ly:matrix([gamma,   0,        gamma*v],
9            [0,        1,        0],
10           [gamma*v,  0,        gamma]);
11 Lmx:matrix([gamma,  -gamma*v, 0],
12           [-gamma*v, gamma,    0],
13           [0,        0,        1]);
14 Lmy:matrix([gamma,  0,        -gamma*v],
15           [0,        1,        0],
16           [-gamma*v, 0,        gamma]);
```

ap / A page from one of Einstein's notebooks.




```
17  /* Calculate the product of the four matrices: */
18  T:Lx.Ly.Lmx.Lmy;
19  /* Define a point on the same light-cone as the origin,
20     which is equivalent to defining a direction.
21     This is Maxima's notation for the column vector
22     (1,1,0): */
23  P:matrix([1],[1],[0]);
24  /* Find the result of T acting on this direction,
25     expressed as a Taylor series to second order in v: */
26  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 `v`. Lines 5-16 define some 3×3 matrices that represent the L transformations. The basis is $\hat{\mathbf{t}}$, $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$. Line 18 calculates the product of the four matrices; the dots represent matrix multiplication. Line 23 defines a point in spacetime $\mathrm{P} = \hat{\mathbf{t}} + \hat{\mathbf{x}}$; we have in mind the direction that includes the lightlike line OP. P has to be expressed as a column matrix (three rows of one column each) so that Maxima will know how to operate using matrix multiplication by T.

Finally line 26 outputs[25] the result of T acting on P:

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

In other words,

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

where $\dots$ represents higher-order terms in v. The interpretation is as follows. Let $\mathrm{O} = (0, 0, 0)$ be the origin. We can interpret OP as defining a direction. Since O is not affected by T, the transformed version, $\mathrm{P} \to T\mathrm{P}$, output by the program represents the same direction, after being acted on by T. 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. We represent that spatial direction using P. 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 $v \ll 1$ the effect is small. For historical reasons discussed in section 1.10.4, this phenomenon is referred to as the Thomas precession.

[25] 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.

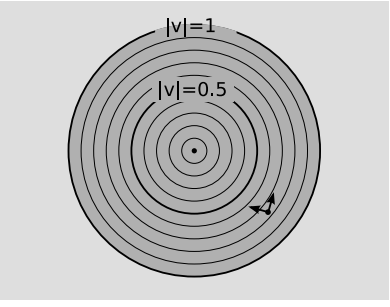
The particular features of this geometry are not necessary. I chose them so that (1) the boosts would be along the Cartesian axes, so that we would be able to write them down easily; (2) it is clear that the effect doesn't arise from any asymmetric treatment of the spatial axes; and (3) the change in 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 classical 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 classical 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.

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

The velocity disk

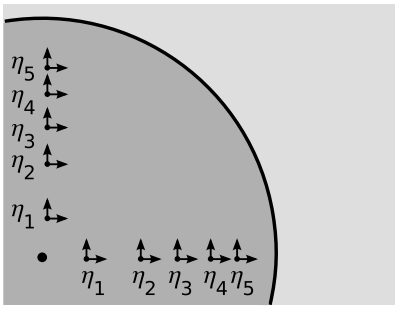
Figure aq shows a useful way of visualizing the combined effects of boosts and rotations in 2+1 dimensions. The disk depicts all possible states of motion relative to some arbitrarily chosen frame of reference. Lack of motion is represented by the point at the center. A point at distance v from the center 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.



aq / The velocity disk.

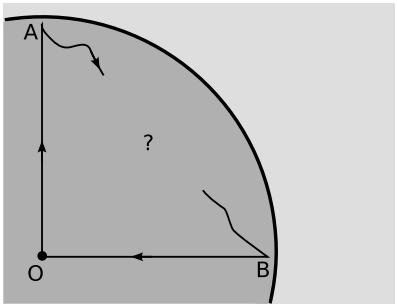
It turns out to be easier to understand the qualitative behavior of our mysterious rotations if we switch from the low-velocity limit to the contrary limit of ultrarelativistic velocities. Suppose we have a rocket-ship with an inertial navigation system consisting of two gyroscopes at right angles to 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 τ_1, τ_2, ..., all separated by equal intervals $\Delta\tau$, the ship's rapidity changes at a constant rate, η_1, η_2, ... This results in a series of frames of reference that appear closer and closer together 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.



ar / Two excursions in a rocket-ship: one along the *y* axis and one along *x*.

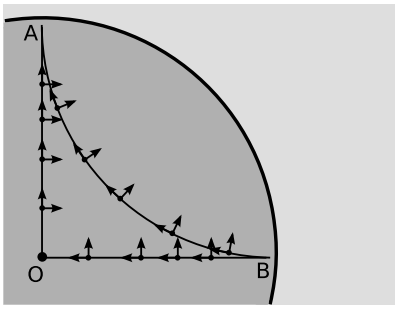
Now suppose that the ship were to accelerate along a route like the one shown in figure **??**. It first accelerates along the y axis at a constant rate (again, as judged by its own sensors), until its velocity is very close to the speed of light, A. It then accelerates, again at a self-perceived constant 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.



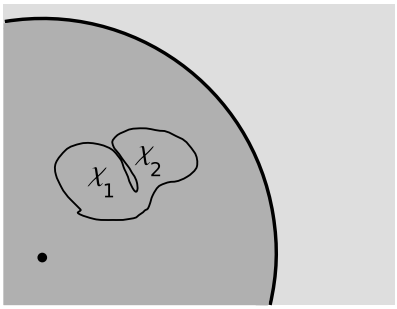
as / A round-trip involving ultrarelativistic velocities. All three legs are at constant acceleration.

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



at / 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.



au / 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 ($\mathbf{v} \to -\mathbf{v}$, but $\mathbf{a} \to +\mathbf{a}$, so $\mathbf{a} \times \mathbf{v} \to -\mathbf{a} \times \mathbf{v}$). 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.

is magnified almost infinitely. As perceived in frame O, the ship's orientation is almost exactly antiparallel to the y axis.[26]

As the ship travels from A to B, its orientation (as judged in frame O) changes from $-\hat{\mathbf{y}}$ to $\hat{\mathbf{x}}$. 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 $\mathbf{u}$ and $\mathbf{v}$ 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.

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 $\mathrm{d}\chi/\mathrm{d}t$ must have the following properties.

> It is odd under a reversal of the direction of motion, $\mathbf{v} \to -\mathbf{v}$. (This corresponds to applying the four hammer hits in the opposite order.)
>
> It is odd under a reversal of the acceleration due to the second boost, $\mathbf{a} \to -\mathbf{a}$.
>
> It is a rotation about the spatial axis perpendicular to the plane of the $\mathbf{v}$ and $\mathbf{a}$ vectors, in the opposite direction compared to the handedness of the curving trajectory.

[26] 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.)

It is approximately linear in $\mathbf{v}$ and $\mathbf{a}$, for small $\mathbf{v}$ and $\mathbf{a}$.

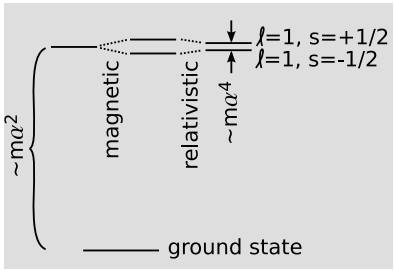
The only rotationally invariant mathematical operation that has these symmetry properties is the vector cross product, so the rate of precession must be $k\mathbf{a}\times\mathbf{v}$, where $k>0$ is nearly independent of $\mathbf{v}$ and $\mathbf{a}$ for small $\mathbf{v}$ and $\mathbf{a}$.

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 $\mathrm{d}\chi/\mathrm{d}t\approx k\mathbf{a}\times\mathbf{v}$ for smooth acceleration. We can do this by considering the physical significance of areas on the velocity disk. As shown in figure au, 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 $\mathrm{d}A$ and integrating their contributions $\mathrm{d}\chi$. What is the relationship between $\mathrm{d}A$ and $\mathrm{d}\chi$? 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 $\mathrm{d}A$ and $\mathrm{d}\chi$ 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, $\mathrm{d}A\approx\mathrm{d}\chi$.

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

$$\boldsymbol{\Omega}\approx\frac{1}{2}\mathbf{a}\times\mathbf{v}\quad,$$

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



av / 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 $s=+1/2$ has its spin angular momentum aligned with its orbital angular momentum, while the $s=-1/2$ state has the two angular momenta in opposite directions. The direction and order of magnitude of the splitting between the two $\ell=1$ 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.

1.10.4 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 $\alpha=ke^2/\hbar\approx1/137$, 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.

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 $\Omega = (1/2)v^2\omega$. Although Thomas was not the first to calculate it, the effect is known as Thomas precession. Since the electron's spin is $\hbar/2$, the energy splitting is $\pm(\hbar/2)\Omega$, 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 $\hbar\omega$ by a factor of $v^2/2$, 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 v/c... Even the cognoscenti of relativity theory (Einstein included!) were quite surprised."

Problems

Key

The notation √indicates that a computerized answer check is available online.

1 Example 9 on page 37 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. (Surprisingly subtle issues can arise in such calculations. See A.Y. Shiekh, Can. J. Phys. 70, 458 (1992).)

Chapter 2
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. 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 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 $\mathrm{d}y$ and $\mathrm{d}x$ are interpreted as infinitesimally small numbers. You may be inclined, based on your previous training, to dismiss infinitesi-

mals 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 2.2. 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.

Areas on a sphere *Example: 1*

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 θ, $\theta + d\theta$, ϕ, and $\phi + d\phi$ has width $r \sin\theta d\theta$ and height $r d\phi$, giving $dA = r^2 \sin\theta d\theta d\phi$. 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.

2.1 The affine parameter revisited, and parallel transport

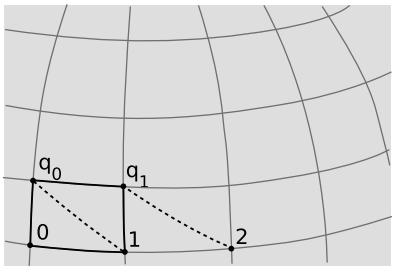
2.1.1 The affine parameter in curved spacetime

An important example of the differential, i.e., local, nature of our geometry is the generalization of the affine parameter to a context broader than affine geometry.

Our construction of the affine parameter with a scaffolding of parallelograms depended on the existence and uniqueness of parallels expressed by 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. To generalize the recipe for the construction, the first obstacle is the ambiguity of the instruction to construct parallelogram $01q_0q_1$, 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.

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.



a / Construction of an affine parameter in curved spacetime.

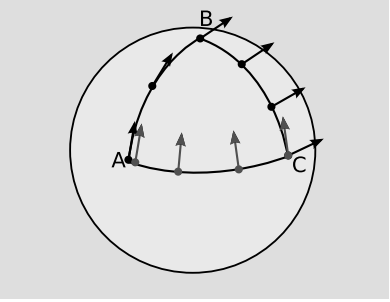
2.1.2 Parallel transport

If you were alert, you may have realized that I cheated you at a crucial point in this construction. We were to make $1q_1$ parallel to $0q_0$ "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 $q_0 01$ and $q_1 12$ 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?

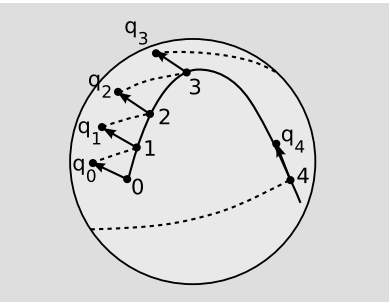
A mathematician's answer would be that our geometry lacks some additional structure called a *connection*, which is a rule that specifies how one locally flat neighborhood is to be joined seamlessly onto another locally flat neighborhood nearby. If you've ever bought two maps and tried to tape them together 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 on page 43. 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 vec-



b / 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.



c / Bad things happen if we try to construct an affine parameter along a curve that isn't a geodesic. This curve is similar to path ABC in figure b. 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.

tor 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 3.2, 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 b.

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 c 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.

Parallel transport compared to Thomas precession

We've now seen two reasons why the orientation of a gyroscope is dependent on the world-line it has been transported along. The path-dependence of parallel transport shows that the gyroscope's orientation can be affected by the curvature of the spacetime. In addition, it can be affected by Thomas precession. These are completely separate effects. In a flat, Lorentzian spacetime, only Thomas precession can occur. On the other hand, we could do geometry on the surface of a sphere, in which case there would be nontrivial effects from parallel transport, but Thomas precession would not exist. This is because Thomas precession is a property of Lorentzian geometry, but a sphere is locally Euclidean, not locally Lorentzian.

Both effects can occur simultaneously, if we have a spacetime that is both curved and locally Lorentzian. When they both occur, the effects add. As an example, both effects applied to the Gravity Probe B experiment, as we will analyze to within an order of

magnitude on page 108, and using an exact treatment on page 140.

2.2 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.

Models provide a way of dealing with both the logical issues and the conceptual ones. Figure a on page 57 "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.

d / 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.

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 of 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 of line segments, 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 propositions, 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

[1] Heath, pp. 195-202

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.

Infinitesimals *Example: 2*

Consider the following axiomatically defined system of numbers:

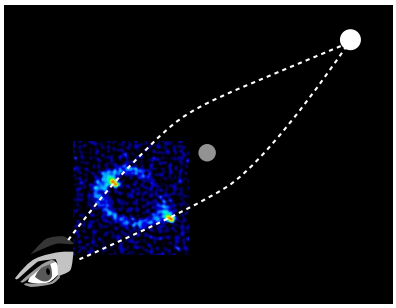
1. It is a field, i.e., it has addition, subtraction, multiplication, and division with the usual properties.
2. It is an ordered geometry in the sense of O1-O4, and the ordering relates to addition and multiplication in the usual way.
3. Existence of infinitesimals: There exists a positive number d 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

[2]The term "elliptic" may be somewhat misleading here. The model is still constructed from a sphere, not an ellipsoid.

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 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."



e / 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.

These two spherical models are not valid models of a general-relativistic spacetime, since they is 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

[3] More on this topic is available in, for example, Keisler's *Elementary Calculus: An Infinitesimal Approach*, Stroyan's *A Brief Introduction to Infinitesimal Calculus*, or my own *Calculus*, all of which are available for free online.

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.

Self-check: 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.

2.3 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.

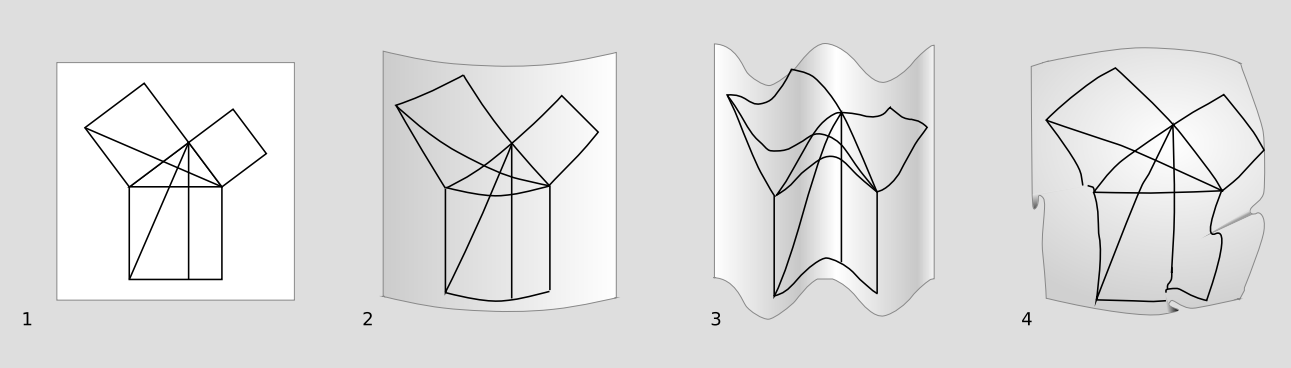
Curvature in elliptic geometry *Example: 3*

What about curvature? In the hemispherical 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 until it returns to the initial point. Since antipodal points are identified, λ equals half the circumference of the sphere, not its whole circumference, so $R = \lambda/\pi$, by wholly intrinsic methods.

Extrinsic curvature *Example: 4*

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 (figure f/1) illustrating illustrating the proof of the Pythagorean theorem in Euclid's *Elements* (proposition I.47) is equally valid if the page is rolled up into 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 f, only the sphere, 4, has intrinsic curvature; the diagram can't be plastered onto the sphere without folding or cutting and pasting.



f / Example 63.

Self-check: How would the ideas of example 4 apply to a cone?

Example 4 shows that it can be difficult to sniff out bogus extrinsic features that seem intrinsic, and example 3 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.

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. 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.

But:

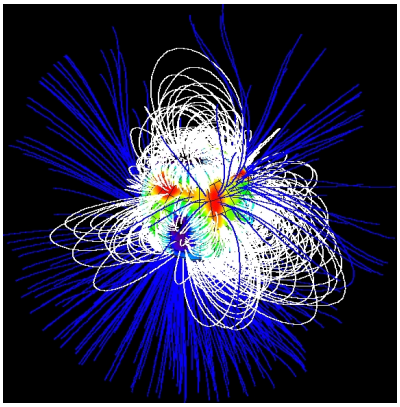
- 2. We can also do measurements in local Lorentz frames.

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 $x \to x' = f(x)$, where f is a smooth one-to-one function. For example, we could have $x \to x + px^3 + q\sin(rx)$ (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.

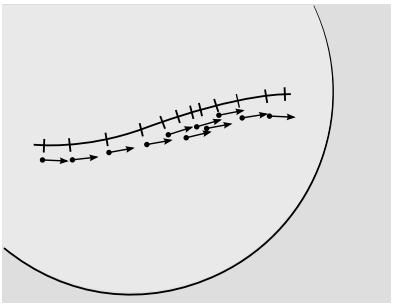
It is instructive to consider this from the point of view of a field theory. Newtonian gravity can be described in three equivalent ways: as a gravitational field $\mathbf{g}$, as a gravitational potential ϕ, or as a set of gravitational field lines. The field lines are never incident on one another, 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,[4] or two sets of field

[4]There is the familiar electrical potential ϕ, measured in volts, but also a



g / 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.



h / 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 coordinate; since the changes in the coordinate are all equal, the arrows are all the same length.

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 4.6.1 on page 112, 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.

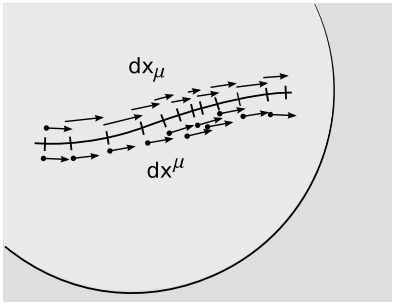
2.4 The metric

Applying these considerations to the creation of a manifestly intrinsic notation, 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; it is not possible to define this type of uniform, Cartesian coordinate system 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.

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, $\mathrm{d}s \neq \sqrt{\mathrm{d}\mathbf{x} \cdot \mathrm{d}\mathbf{x}}$, since the vectors are the wrong length. We therefore need some mathematical rule, some function, that accepts a vector as its input and gives a scalar as its output. This function is a locally adjustable fudge factor that compensates for the wrong lengths of the little vectors. Since the space is locally flat and uniform, the function must be linear, and from linear algebra, we know that the

vector potential $\mathbf{A}$, which you may or may not have encountered. Briefly, the electric field is given not by $-\nabla\phi$ but by $-\nabla\phi - \partial\mathbf{A}/\partial t$, while the magnetic field is the curl of $\mathbf{A}$. This is introduced at greater length in section 3.2.5 on page 85.

most general function of this kind is an inner product. If the little arrow is a row vector, then the function would be represented by taking the row vector's inner product with some column vector to give $\mathrm{d}s^2$. Of course the distinction between row and column vectors is pointless in a one-dimensional space, but it should be clear that this will provide an appropriate foundation for the generalization to more than one coordinate. The row and column vectors are referred to as one another's duals. Figure i shows the resulting picture. Anticipating the generalization to four-dimensional spacetime with coordinates (x^0, x^1, x^2, x^3), we'll start referring to x as x^μ, although in our present one-dimensional example $\mu = 0$ is fixed. The reason for the use of the odd-looking superscripts, rather than subscripts, will become clear shortly.



i / The vectors $\mathrm{d}x^\mu$ and $\mathrm{d}x_\mu$ are duals of each other.

The vectors drawn below the curve are called the contravariant vectors, notated $\mathrm{d}x^\mu$, and the ones above it are the covariant vectors, $\mathrm{d}x_\mu$. It's not particularly important to keep track of which is which, since the relationship between them is symmetric, like the relationship between row and column vectors. Each is the dual of the other. The arc length is given by $\int \mathrm{d}s = \int \sqrt{\mathrm{d}x^\mu \mathrm{d}x_\mu}$, or, equivalently we say $\mathrm{d}s^2 = \mathrm{d}x^\mu \mathrm{d}x_\mu$.

Given a $\mathrm{d}x^\mu$, how do we find its dual $\mathrm{d}x_\mu$, 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 called 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 $\mathrm{d}\phi$ 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 $\mathrm{d}s^2 = \Sigma \mathrm{d}x^\mu \mathrm{d}x_\mu$ for the squared arc length, and in differential geometry we practice the convenient notational convention, introduced by Einstein, of assuming a summation when an index is repeated, so this becomes

$$\mathrm{d}s^2 = \mathrm{d}x^\mu \mathrm{d}x_\mu \quad .$$

2.4.1 The Euclidean metric

In a Euclidean plane, where the distinction between covariant and contravariant vectors is irrelevant, this expression for $\mathrm{d}s^2$ is

simply the Pythagorean theorem, summed over two values of i for the two coordinates:

$$\mathrm{d}s^2 = \mathrm{d}x^i \mathrm{d}x_i = \mathrm{d}x^2 + \mathrm{d}y^2$$

The symbols $\mathrm{d}x$, $\mathrm{d}x^0$, and $\mathrm{d}x_0$ are all synonyms, and likewise for $\mathrm{d}y$, $\mathrm{d}x^1$, and $\mathrm{d}x_1$.

In the non-Euclidean case, the Pythagorean theorem is false; $\mathrm{d}x^\mu$ and $\mathrm{d}x_\mu$ 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 $\mathrm{d}x_\mu = g_{\mu\nu}x^\nu$, we have

$$\mathrm{d}s^2 = g_{\mu\nu}\mathrm{d}x^\mu \mathrm{d}x^\nu \quad ,$$

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.

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

In an introductory course in Newtonian mechanics, one makes a distinction between vectors, which have a direction in space, and scalars, which do not. These are specific examples of *tensors*, which can be expressed as objects with m superscripts and n subscripts. A scalar has $m = n = 0$. A covariant vector has $(m, n) = (0, 1)$, a contravariant 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 3. 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 Newtonian "vector" $\mathbf{u} = (m, T, e)$, where m is the mass of the moon, T is the temperature in Chicago, and e is the charge of the electron. This creature $\mathbf{u}$ doesn't deserve to be called a vector, because it doesn't behave as a vector under rotation. Similarly, a tensor is required to behave in a certain way under rotations and Lorentz boosts.

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

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

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

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 on page 122.

Self-check: Characterize an antisymmetric rank-2 tensor in two dimensions.

A change of scale *Example: 5*

▷ How is the effect of a uniform rescaling of coordinates 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 $g_{\mu\nu} \to \alpha^{-2} g_{\mu\nu}$.

Polar coordinates *Example: 6*

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 $d\theta$ correspond to infinitesimal displacements dr and $r d\theta$ 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 6 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 adopt a convention introduced by Roger Penrose known as the abstract index notation. In this convention, Latin superscripts and subscripts indicate that an equation is of general validity, without regard to any choice of coordinate system, while Greek ones are used for coordinate-dependent equations. For example, we can write the general expression for squared differential arc length with Latin indices,

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

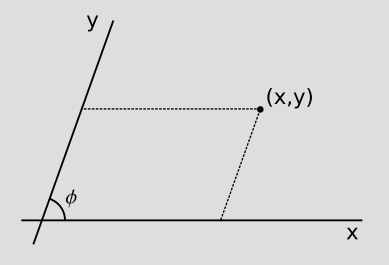
because it holds regardless of the coordinate system, whereas the vanishing of the off-diagonal elements of the metric in Euclidean polar coordinates has to be written as $g_{\mu\nu} = 0$ for $\mu \neq \nu$, since it would in general be false if we used a different coordinate system to describe the same Euclidean plane. The advantages of this notation became widely apparent to relativists starting around 1980, so for example it is used in the text by Wald (1984), but not in Misner, Thorne, and Wheeler (1970). Some of the older literature uses a notation in which the Greek and Latin indices are instead used to distinguish between timelike and spacelike components of a vector, but this usage is dying out, since it inappropriately singles out a distinction between time and space that is not actually preserved under a Lorentz boost.

Oblique Cartesian coordinates *Example: 7*

▷ Oblique Cartesian coordinates are like normal Cartesian coordinates in the plane, but their axes are at at an angle $\phi \neq \pi/2$ 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 $dx^2 = ds \csc\phi$. We then have

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



j / Example 7.

Area *Example: 8*

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 5 on page 69, 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 $dA = \sqrt{g_{11}g_{22}}dx^1dx^2$. 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 $dA = \sqrt{|g|}dx^1dx^2$. 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 7, 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.

Area of a sphere *Example: 9*

For coordinates (θ, ϕ) on the surface of a sphere of radius r, we have, by an argument similar to that of example 6 on page 69, $g_{\theta\theta} = r^2$, $g_{\phi\phi} = r^2 \sin^2\theta$, $g_{\theta\phi} = 0$. The area of the sphere is

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

Inverse of the metric *Example: 10*

▷ 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.

2.4.2 The Lorentz metric

In a locally Euclidean space, the Pythagorean theorem allows us to express the metric in local Cartesian coordinates in the simple form $g_{\mu\mu} = +1$, $g_{\mu\nu} = 0$, i.e., $g = diag(+1, +1, \ldots, +1)$. 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,

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

and this remains true after the Lorentz boost $(t, x) \to (\gamma t, \gamma x)$. 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, and timelike ones have 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 $g = diag(+1, -1, -1, -1)$.

2.4.3 Isometry, inner products, and the Erlangen Program

In Euclidean geometry, the dot product of vectors $\mathbf{a}$ and $\mathbf{b}$ is given by $g_{xx}a_xb_x + g_{yy}a_yb_y + g_{zz}a_zb_z = a_xb_x + a_yb_y + a_zb_z$, and in the special case where $\mathbf{a} = \mathbf{b}$ we have the squared magnitude. In the tensor notation, $a^\mu b_\nu = a^1b_1 + a^2b_2 + a^3b_3$. Like magnitudes, dot products are invariant under rotations. This is because knowing the dot product of vectors $\mathbf{a}$ and $\mathbf{b}$ entails knowing the value of $\mathbf{a}\cdot\mathbf{b} = |\mathbf{a}||\mathbf{a}|\cos\theta_{\mathbf{ab}}$, 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[5] that preserves congruence of triangles.

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 $a^\mu b_\nu = a^0b_0 - a^1b_1 - a^2b_2 - a^3b_3$. 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[6] 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.

The triangle inequality *Example: 11*

In Euclidean geometry, the triangle inequality $|\mathbf{b}+\mathbf{c}| < |\mathbf{b}| + |\mathbf{c}|$ follows from

$$(|\mathbf{b}| + |\mathbf{c}|)^2 - (\mathbf{b}+\mathbf{c})\cdot(\mathbf{b}+\mathbf{c}) = 2(|\mathbf{b}||\mathbf{c}| - \mathbf{b}\cdot\mathbf{c}) \geq 0 \quad .$$

[5] 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.

[6] 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.

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.

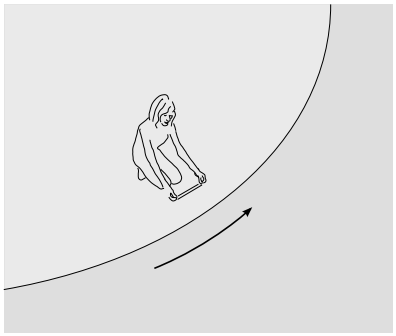
In Lorentzian geometry, the situation is different. Let **b** and **c** be timelike vectors, so that they represent possible world-lines. Then the relation $\mathbf{a} = \mathbf{b} + \mathbf{c}$ suggests the existence of two observers who take two different paths from one event to another. A goes by 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 now becomes $|\mathbf{b} + \mathbf{c}| > |\mathbf{b}| + |\mathbf{c}|$. 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 **b** and **c** are parallel. For timelike vectors, parallelism actually minimizes the inner product rather than maximizing it.[7]

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 3 on page 63, 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.

[7]Proof: Let **b** and **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, but each is still directed forward in time, not backward. Now let **b** and **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.

2.4.4 Einstein's carousel

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



k / Observer A, rotating with the carousel, measures an azimuthal distance with a ruler.

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-sticks in the azimuthal direction they become 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 geometry is non-Euclidean, and the metric differs from the Euclidean one found in example 6 on page 69.

Ehrenfest pointed out in 1909 that 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. This was known as Ehrenfest's paradox, and its resolution comes from considerations we've already discussed on page 37. 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.

Self-check: Can we get around these problems by applying torque uniformly all over the disk, so that the rotation starts smoothly and simultaneously everywhere? What if we build the disk by assembling the building materials so that they are already rotating properly?

Now let's find the metric according to observer A by applying the change of coordinates $\theta \to \theta' = \theta - \omega t$. First we take the Euclidean metric of example 6 on page 69 and rewrite it as a (globally) Lorentzian metric in spacetime for observer B,

$$\mathrm{d}s^2 = \mathrm{d}t^2 - \mathrm{d}r^2 - r^2\mathrm{d}\theta^2 \quad .$$

Applying the transformation into A's coordinates, we find

$$\mathrm{d}s^2 = (1-\omega^2 r^2)\mathrm{d}t^2 - \mathrm{d}r^2 - r^2\mathrm{d}\theta'^2 - 2\omega r^2 \mathrm{d}\theta'\mathrm{d}t \quad .$$

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 $\mathrm{d}t^2$ term. But the $\mathrm{d}r^2$ and $\mathrm{d}\theta'^2$ 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 $\mathrm{d}\theta'\mathrm{d}t$ 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 $r\theta'$ on their odometers and equal proper times $\mathrm{d}s$ on their clocks. But this is not the case, because the sign of the $\mathrm{d}\theta'\mathrm{d}t$ term is opposite for the two world-lines.

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. Our present goal is much more modest. 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?

A trick for improving the situation is to eliminate the $\mathrm{d}\theta'\mathrm{d}t$ cross-term by completing the square. The result is

$$\mathrm{d}s^2 = (1-\omega^2 r^2)\left[\mathrm{d}t + \frac{\omega r^2}{1-\omega^2 r^2}\mathrm{d}\theta'\right]^2 - \mathrm{d}r^2 - \frac{r^2}{1-\omega^2 r^2}\mathrm{d}\theta'^2 \quad .$$

We can now interpret the quantity in square brackets as a time coordinate that is properly synchronized between different points on the carousel. If a meter stick is placed on the carousel, the spacelike negative terms can be used to describe its spatial length, with the understanding that this is not the spatial separation of two events at the same value of the coordinate t but of two events that are at the same universally synchronized clock time. The final two terms can now be thought of as the non-Euclidean metric of the space. The factor of $(1-\omega^2 r^2)^{-1} = \gamma^2$ in the $\mathrm{d}\theta'^2$ term is simply the expected Lorentz-contraction factor. In other words, the circumference is, as expected, greater than 2π by a factor of γ.

2.5 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.

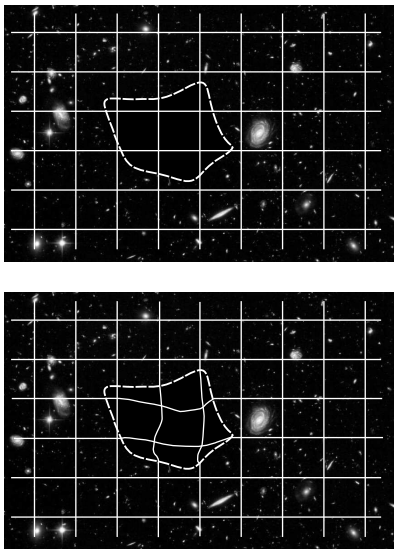
With the benefit of hindsight, we can consider these issues in terms of the general description of measurements in relativity given on page 64:

1. We can tell whether events and world-lines are incident.

2. We can do measurements in local Lorentz frames.

2.5.1 The hole argument

The main factor that led Einstein to his false start is known as the hole argument. Suppose that we know all 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 $g(\mathbf{x})$ that solves the field equations for this set of sources, where $\mathbf{x}$ is some set of coordinates. Now if the field equations are coordinate-independent, we can introduce a new set of coordinates $\mathbf{x}'$, which is identical to $\mathbf{x}$ outside the hole, but differs from it on the inside. If we reexpress the metric in terms of these new coordinates as $g'(\mathbf{x}')$, then we are guaranteed that $g'(\mathbf{x}')$ is also a solution. But furthermore, we can substitute $\mathbf{x}$ for $\mathbf{x}'$, and $g'(\mathbf{x})$ 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.



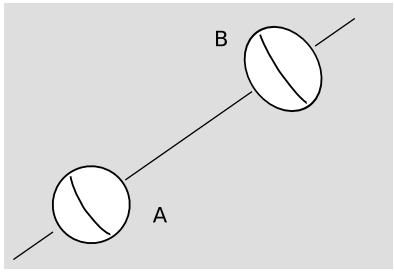
I / Einstein's hole argument.

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 us do is measurements of the incidence of world-lines. Relabeling all the points inside the hole doesn't change any of the incidence relations. For example, if two test particles sent into the region collide at a point $\mathbf{x}$ inside the hole, then changing the point's name to $\mathbf{x}'$ doesn't change the observable fact that they collided.

2.5.2 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. 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.



m / 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.

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 (see page 24).

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."

Chapter 3
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 on page 68. A Lorentz scalar is a tensor of rank 0, and a Lorentz vector is a rank-1 tensor.

3.1 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.

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 $\mathrm{d}s^2 = \ldots$, the quantity $\mathrm{d}s$ is a Lorentz scalar. In the special case of a timelike world-line, $\mathrm{d}s$ and $\mathrm{d}\tau$ are the same thing. (In books that use a $-+++$ metric, one has $\mathrm{d}s = -\mathrm{d}\tau$.)

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,

[1] Some older books define mass as transforming according to $m \to \gamma m$, which can be made to give a self-consistent theory, but is ugly.

but so is $2\tau+7$. 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 5.2.6, page 146, 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.

3.2 Four-vectors

3.2.1 The velocity and acceleration four-vectors

Our basic Lorentz vector is the spacetime displacement $\mathrm{d}x^i$. Any other quantity that has the same behavior as $\mathrm{d}x^i$ 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. Dividing the infinitesimal displacement by an infinitesimal proper time interval, we have the four-velocity vector $v^i = \mathrm{d}x^i/\mathrm{d}\tau$, whose components in a Lorentz coordinate system are $(\gamma, \gamma v^1, \gamma v^2, \gamma v^3)$, where v^μ, $\mu = 1, 2, 3$, 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 if the particle is not moving at the speed of light.

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 1.8.1 on page 38 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 1.10.3 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.

A zero velocity vector? *Example: 1*

▷ 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?

▷ 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 zero magnitude. 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 $(1, 0, 0, 0)$.

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 classical acceleration three-vector, in the limit of small velocities.

Constant acceleration *Example: 2*

▷ 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.

▷ 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

$$\dot{t}^2 - \dot{x}^2 = 1 \quad .$$

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

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

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

$$x = \frac{1}{a}\left(\sqrt{1 + a^2t^2} - 1\right) \quad .$$

As t approaches infinity, dx/dt approaches the speed of light.

3.2.2 The momentum four-vector

Multiplying by the particle's mass, we have the four-momentum $p^i = mv^i$, which in Lorentz coordinates is $(m\gamma, m\gamma v^1, m\gamma v^2, m\gamma v^3)$. 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 classical value. 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.

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, by a factor of $\gamma - 1$ averaged over all the particles. The same must be true for the gravitational mass, based on the equivalence principle as verified

by Eötvös experiments. 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. The timelike component of the four-momentum, $m\gamma$, is interpreted as the mass-energy of the particle, consisting of its mass m plus its kinetic energy $m(\gamma - 1)$.

Gravitational redshifts *Example: 3*

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 $\Delta E/E = \Delta\phi$. In metric units, this becomes $\Delta E/E = \Delta\phi/c^2$, and in the field near the Earth's surface we have $\Delta E/E = gh/c^2$. This is the same result that was found in section 1.5.7 based only on the equivalence principle, and verified experimentally by Pound and Rebka as described in section 1.5.8.

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 $p^2 - E^2 = m^2$.

A common source of confusion for beginners in relativity is the distinction between quantities that are conserved and quantities that are the same in all frames. There is nothing relativistic about this distinction. Before Einstein, physicists already knew that observers in different frames of reference would agree on the mass of a particle. That is, m was known to be frame-invariant. They also knew that energy was conserved. But just because energy was conserved, that didn't mean that it had to be the same for observers in all frames of reference. The kinetic energy of the chair you're sitting in is millions of joules in a frame of reference tied to the axis of the earth. In relativity, m is frame-invariant (i.e., a Lorentz scalar), but the conserved quantity is the momentum four-vector, which is not frame-invariant.

Applying $p^2 - E^2 = m^2$ to the special case of a massless particle, we have $|p| = E$, which demonstrates, for example, that a beam of light exerts pressure when it is absorbed or reflected by a surface. A massless particle must also travel at exactly the speed of light, since $|p| \to E$ requires $m\gamma v \to m\gamma$; conversely, a massive particle always has $|v| < 1$.

Massive neutrinos *Example: 4*

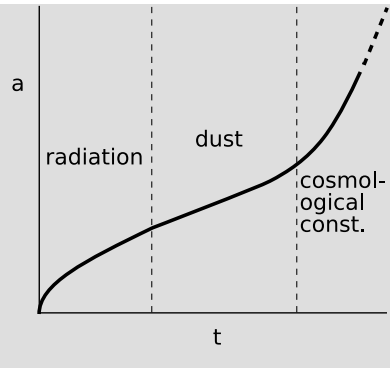
Neutrinos were long thought to be massless, but are now believed to have masses in the eV range. If they had been massless, they would always have had to propagate at the speed of light. Although they are now thought to have mass, that mass is six orders of magnitude less than the MeV energy scale of 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.

Dust and radiation in cosmological models *Example: 5*

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 mass-energy 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 mass-energy 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



a / Example 5.

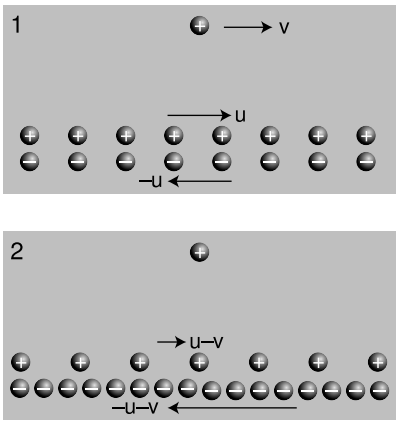
to a qualitatively different behavior resulting from the change in the equation of state.

In the present era, the universe's equation of state is dominated by neither dust nor radiation but by the cosmological constant (see page 157). Figure a shows the evolution of the size of the universe for the three different regimes. Some of the simpler cases are derived in sections 6.2.3 and 6.2.4, starting on page 163.

3.2.3 The frequency four-vector

Frequency is to time as the wavenumber $k = 1/\lambda$ is to space, so when treating waves relativistically it is natural to conjecture that there is a four-frequency f_a made by assembling $(f, \mathbf{k})$, 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 $A = A_\text{o} \exp[2\pi i f_a x^a]$ the partial derivative operator is identical to multiplication by $2\pi f_a$.

3.2.4 A non-example: electric and magnetic fields



b / Magnetism is a purely relativistic effect.

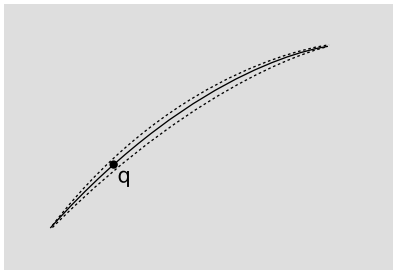
It is fairly easy to see that the electric and magnetic fields cannot be the spacelike parts of two four-vectors. Consider the arrangement shown in figure b/1. We have a two infinite trains of moving charges superimposed on the same line, and a single charge alongside the line. Even though the line charges 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.

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.

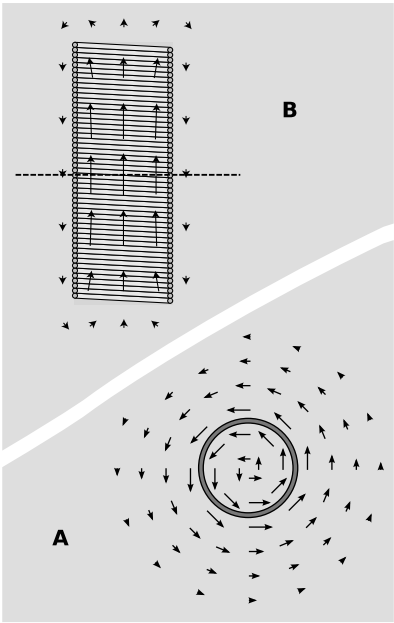
Although the example shows that the electric and magnetic fields do transform when we change from one frame to another, it is easy to show that they do not transform as the spacelike parts of a relativistic four-vector. This is because transformation between frames 1 and 2 is along the axis parallel 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.



c / 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.

3.2.5 The electromagnetic potential four-vector

An electromagnetic quantity that *does* transform as a four-vector is the potential. On page 65, 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 $(\mathbf{E}, \mathbf{B})$ needs a pair of potentials, Φ and $\mathbf{A}$. It's easy to see that Φ can't be a Lorentz scalar. Electric charge q is a scalar, so if Φ were a scalar as well, then the product $q\Phi$ 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 Φ is not a scalar.



d / 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 $1/r$ on the outside.

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

Consider the example of figure c 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 $\mathbf{k} \cdot \mathrm{d}\mathbf{s}$, where $\mathbf{k}$ is the wavenumber. To make this relativistic, we need to use the frequency four-vector to form $f_b \mathrm{d}x^b$, which can also be expressed as $f_b v^b \mathrm{d}\tau = \gamma(f - \mathbf{k} \cdot \mathbf{v})\mathrm{d}\tau$. If the charge is at rest and there are no magnetic fields, then the quantity in parentheses is $f = E/h = (q/h)\Phi$. The correct relativistic generalization is clearly $f_b = (q/h)A_b$.

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

$$\begin{aligned} \mathbf{E} &= -\nabla\Phi - \frac{\partial \mathbf{A}}{\partial t} \\ \mathbf{B} &= \nabla\mathbf{B} \qquad . \end{aligned}$$

An excellent discussion of the vector potential from a purely classical point of view is given in the classic *Feynman Lectures*.[2] Figure d shows an example.

3.3 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, $x'^\mu = ax^\mu + b$. 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

[1] $$v'^\mu = v^\kappa \frac{\partial x'^\mu}{\partial x^\kappa}$$

[2] $$v'_\mu = v_\kappa \frac{\partial x^\kappa}{\partial x'^\mu} \qquad .$$

Note the inversion of the partial derivative in one equation compared to the other.

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

Self-check: 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 [2], 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 $\partial/\partial x^i$ and $\partial/\partial x_i$, the superscripts and subscripts should be understood as being turned upside-down. Similarly, [1] 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., $\mathrm{d}^2x/\mathrm{d}t^2$ having units of meters per second squared.

A quantity v that transforms according to [1] or [2] is referred to as a rank-1 tensor, which is the same thing as a vector.

The identity transformation *Example: 6*

In the case of the identity transformation $x'^\mu = x^\mu$, equation [1] clearly gives $v' = v$, since all the mixed partial derivatives $\partial x'^\mu/\partial x^\kappa$ with $\mu \neq \kappa$ are zero, and all the derivatives for $\kappa = \mu$ equal 1.

In equation [2], it is tempting to write

$$\frac{\partial x^\kappa}{\partial x'^\mu} = \frac{1}{\frac{\partial x'^\mu}{\partial x^\kappa}} \qquad \text{(wrong!)} \qquad ,$$

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:

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

Self-check: Write the similar expressions for $g^{\mu\nu}$, g^μ_ν, 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.

An accelerated coordinate system? *Example: 7*

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

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

The inverse transformation is

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

The tensor transformation law gives

$$\begin{aligned} g'_{t't'} &= 1 - (at')^2 \\ g'_{x'x'} &= -1 \\ g'_{x't'} &= -at' \qquad . \end{aligned}$$

Clearly something bad happens at $at' = \pm 1$, when the relative velocity surpasses the speed of light: the $t't'$ 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 $a \to 0$ 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.

Physically meaningful constant acceleration *Example: 8*

To make a more physically meaningful version of example 7, we need to use the result of example 2 on page 81. The somewhat messy derivation of the coordinate transformation is given by Semay.[3] The result is

$$\begin{aligned} t' &= \left(x + \frac{1}{a} \right) \sinh at \\ x' &= \left(x + \frac{1}{a} \right) \cosh at \end{aligned}$$

Applying the tensor transformation law gives (problem 6, page 130):

$$\begin{aligned} g'_{t't'} &= (1 + ax')^2 \\ g'_{x'x'} &= -1 \end{aligned}$$

Unlike the result of example 7, this one never misbehaves.

3.4 Experimental tests

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

3.4.1 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 on page 43 the 2007 Gravity Probe B experiment (discussed in detail on pages 109 and 140), 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.[4]

3.4.2 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.[5]

Assuming Lorentz invariance, a photon cannot decay into an electron and a positron, $\gamma \to e^+ + e^-$, 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. Suppose, however, that material particles have a

[4]This statement of the equivalence principle, along with the others we have encountered, is summarized in the back of the book on page 184.

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

maximum speed $c_m = 1$, while photons have a maximum speed $c_p > 1$. Then the photon's momentum four-vector, $(E, E/c_p)$ is timelike, so a frame does exist in which its three-momentum is zero. The detection of cosmic-ray gammas from distant sources with energies on the order of 10 TeV puts an upper limit on the decay rate, implying $c_p - 1 \lesssim 10^{-15}$.

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 $c_p - 1 \gtrsim -10^{-23}$.

3.4.3 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 $P \sim npv/V$, where p is the typical momentum.

Nondegenerate gas: In an ordinary gas such as air, the parti-

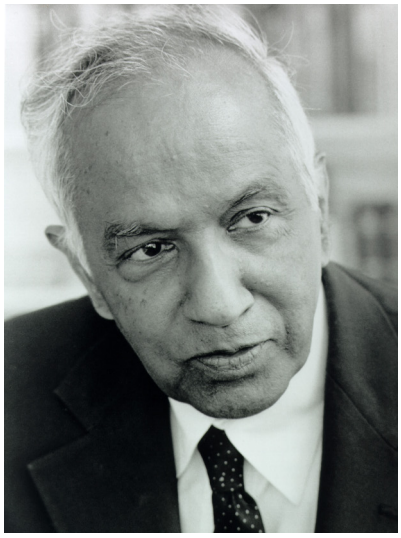
cles are nonrelativistic, so $v = p/m$, and the thermal energy per particle is $p^2/2m \sim kT$, so the pressure is $P \sim nkT/V$.

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 $\sim V/n$, so its wavelength is no more than $\sim (V/n)^{1/3}$, leading to $p = h/\lambda \sim h(n/V)^{1/3}$. If the speeds of the particles are still nonrelativistic, then $v = p/m$ still holds, so the pressure becomes $P \sim (h^2/m)(n/V)^{5/3}$.

Relativistic, degenerate gas: If the compression is strong enough to cause highly relativistic motion for the particles, then $v \approx c$, and the result is $P \sim hc(n/V)^{4/3}$.

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 $P \sim \rho g r \sim GM^2/r^4$, 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 $K \approx 4000$. 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 $\sim (hc/G)^{3/2}(Km)^{-2} = 6M_\odot$. A less sloppy calculation gives something more like $1.4M_\odot$. The self-consistency of this solution is investigated in homework problem 2 on page 97.



e / 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

$$\begin{aligned} \mathrm{n} &\to \mathrm{p} + e^- + \bar{\nu} \\ \mathrm{p} + e^- &\to \mathrm{n} + \nu \end{aligned} \quad ,$$

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 neutron-to-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 $T = \sqrt{3\pi/G\rho}$, 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 1). In reality, astronomical bodies of planetary size and greater are held together by their own gravity, so we have $T \gtrsim 1/\sqrt{G\rho}$ 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:

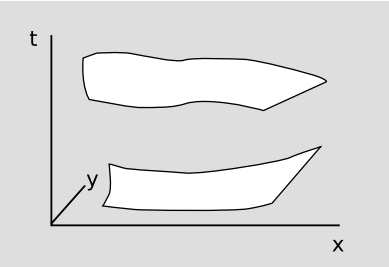
$$1.4M_\odot \left(\frac{Km_e}{m_n}\right)^2 \approx 5M_\odot$$

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 push 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.

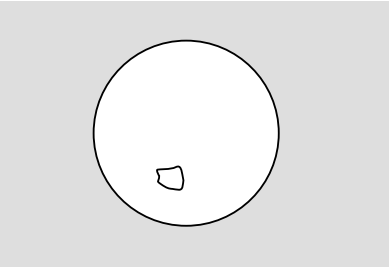
For stars with masses above the Tolman-Oppenheimer-Volkoff limit, theoretical predictions become even more speculative. A variety of bizarre objects has been proposed, including gravastars, fuzzballs, black stars, quark stars, Q-balls. It seems likely, however, both on theoretical and observational grounds, that objects with masses of about 3 to 20 solar masses end up as black holes; see section 5.3.3.

3.5 Conservation laws

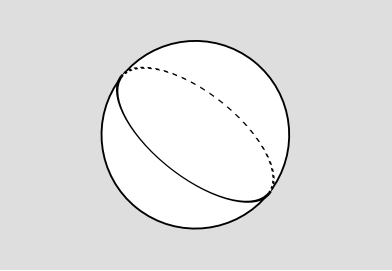
Some of the first tensors we discussed were mass and charge, both rank-0 tensors, and the rank-1 momentum tensor, which contains both the classical energy and the classical momentum. Physicists originally decided that mass, charge, energy, and momentum were interesting because these things were found 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



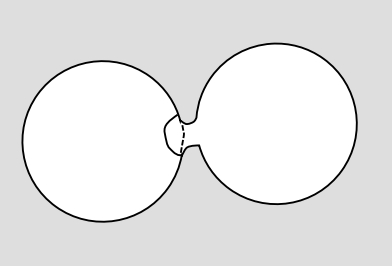
f / Two spacelike surfaces.



g / We define a boundary around a region whose charge we want to measure.



h / This boundary cuts the sphere into equal parts.



i / The circumference of this boundary is small, but the region it contains is large.

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 this way 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 f, 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.

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 6.2 for reassurance that this isn't physically unreasonable), and we can visualize it as a two-sphere.

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.

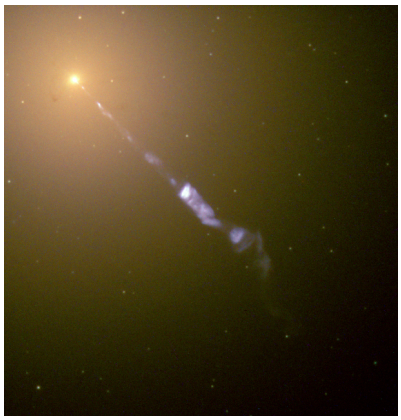
But now we're in real trouble. How would we go about actually measuring the total electric charge of the universe? The only way to do it is to measure electric fields, and then apply Gauss's law. This requires us to single out some surface that we can integrate the flux over, as in g. This would really 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, h. If we measure a nonvanishing total flux across it, how do we know where the charge is? It could be on either side.

You might protest that this is an artificial example. In reality wouldn't we conduct a "charge survey" by chopping up space into small regions, as in g? Figure i shows that this won't work either. The boundary curve has a small circumference, but it contains a very large interior region.

The conclusion is that conservation laws only make sense in rel-

ativity under very special circumstances. We do not have anything like over-arching principles of conservation.

As an example, section 5.2.5 shows how to define conserved quantities, which behave like energy and momentum, for the motion of a test particle in a particular metric.



j / A relativistic jet.

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 asymptotically flat spacetimes, we can also get around the problems that crop up with conserved vectors, such as momentum. 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.

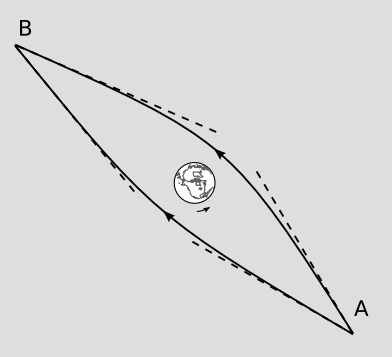
As an example, figure j shows a jet of matter being ejected from the galaxy M87 at ultrarelativistic fields. The blue color of the jet in the visible-light image comes from synchrotron radiation, which is the electromagnetic radiation emitted by relativistic charged particles accelerated by a magnetic 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 the angular momentum preserves information about their formation, and that the angular momentum is externally detectable via its effect on the spacetime metric.

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

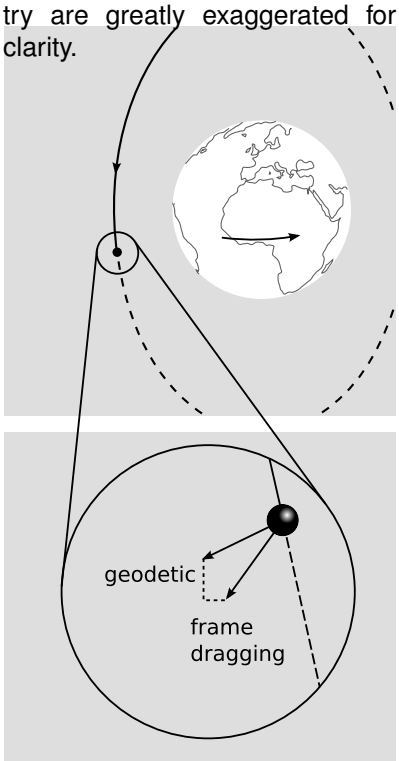
$$\mathrm{d}s^2 = \mathrm{d}t^2 - \mathrm{d}r^2 - r^2\mathrm{d}\theta^2 - r^2\sin^2\theta\mathrm{d}\phi^2 \quad .$$

We'll see in chapter 5 that for a non-rotating black hole, the metric is of the form

$$\mathrm{d}s^2 = (\ldots)\mathrm{d}t^2 - (\ldots)\mathrm{d}r^2 - r^2\mathrm{d}\theta^2 - r^2\sin^2\theta\mathrm{d}\phi^2 \quad ,$$



k / 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 for clarity.



l / Gravity Probe B verified the existence of frame-dragging, 90 years after it was predicted. The rotational axis of the gyroscope precesses in two perpendicular planes due to the two separate effects: geodetic and frame-dragging.

where $(\ldots)$ 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 azimuthal coordinate ϕ.
3. The metric should have axial symmetry, i.e., it should be independent of ϕ.

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

$$f(\ldots)L\,\mathrm{d}\phi\,\mathrm{d}t \quad ,$$

where $(\ldots)$ 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 j 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, k. The deflections will be different, because the sign of $\mathrm{d}\phi\mathrm{d}t$ 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.

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.

Problems

Key

The notation √indicates that a computerized answer check is available online.

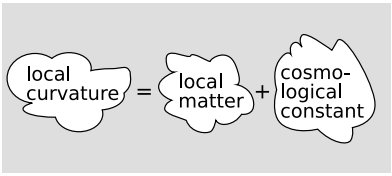
1 Derive the equation $T = \sqrt{3\pi/G\rho}$ given on page 92 for the period of a rotating, spherical object that results in zero apparent gravity at its surface.

2 Section 3.4.3 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 9 on page 37, a reasonable order-of-magnitude check on this result is that we should have $M/r \ll c^2/G$.

3 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.

Chapter 4
Curvature

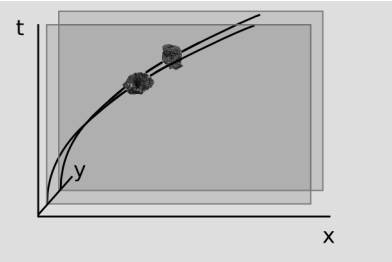
General relativity describes gravitation as a curvature of spacetime, with matter acting as the source of the curvature in the same way that electric charge acts as the source of electric fields. Our goal is to arrive at Einstein's field equations, which relate the local intrinsic curvature to the locally 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 $c(=1)$.



a / The expected structure of the field equations in general relativity.

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 4 on page 63 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.

As an example, drop two rocks side by side, b. 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.



b / 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 is not a measure of intrinsic curvature. In example 8 on page 88, we found the metric for an accelerated observer to be

$$g'_{t't'} = (1 + ax')^2 \qquad g_{x'x'} = -1 \qquad ,$$

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 measurements of any tensor 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 g nor its first derivatives indicate curvature, we can reasonably conjecture that the curvature might be encoded in its second derivatives.

c / Tidal forces disrupt comet Shoemaker-Levy.

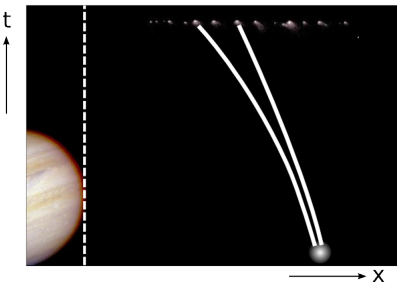
4.1 Tidal curvature versus curvature caused by local sources

A further complication is the need to distinguish tidal curvature from curvature caused by local sources. Figure c 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.

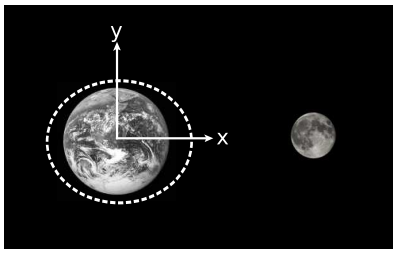
If initially parallel lines lose their parallelism, that is clearly an indication of intrinsic curvature. We call it a measure of *sectional curvature*, because the loss of parallelism occurs within a particular plane, in this case the (t, x) plane represented by figure d.

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, e.



d / 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.

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, f. 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 $1/r^2$ forces such as gravity that they conserve volume in this way; this is essentially a restatement of Gauss's law in a vacuum.

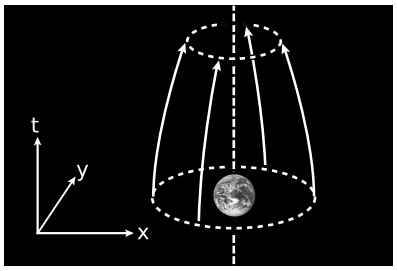


e / The moon's gravitational field causes the Earth's oceans to be distorted into an ellipsoid. The sign of the sectional curvature is negative in the $x-t$ plane, but positive in the $y-t$ plane.

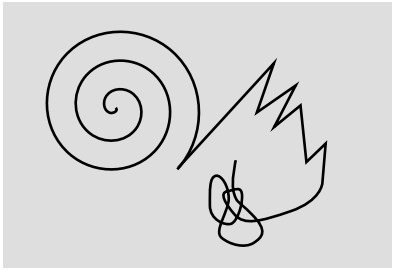
4.2 The energy-momentum tensor

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 (page 24), 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

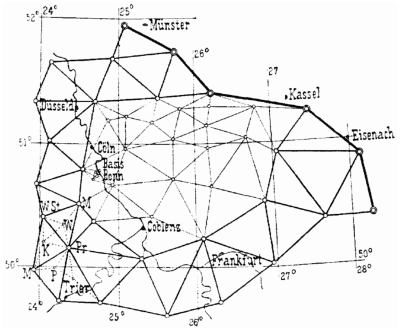


f / A cloud of test masses is released at rest in a spherical shell around the earth, shown here as a circle because the z axis is omitted. The volume of the shell contracts over time, which demonstrates that the local curvature of spacetime is generated by a local source — the earth — rather than some distant one.



g / This curve has no intrinsic curvature.

h / A surveyor on a mountaintop uses a heliotrope.



i / 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.

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 energy-momentum tensor T^{ij} (also known as the stress-energy tensor), 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^{00} represents the mass density.

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 energy-momentum tensor is discussed further in section 6.1.2 on page 157.

4.3 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, g, 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.

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

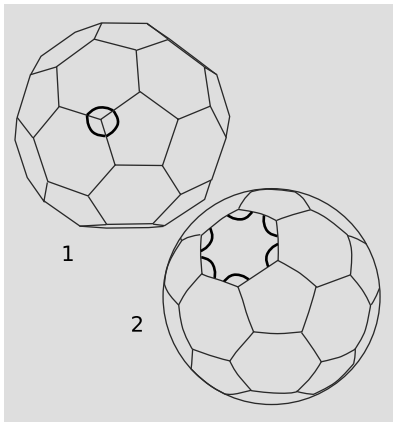
of them as approximations to geodesics painted directly on the dirt, which would be intrinsic rather than extrinsic. The angular defect around a vertex now vanishes, because the space is locally Euclidean, but we now pick up a different kind of angular defect, which is that the interior angles of a triangle no longer add up to the Euclidean value of π.

A polygonal survey of a soccer ball *Example: 1*

Figure j applies similar ideas to a soccer ball, the only difference being the use of pentagons and hexagons rather than triangles.

In j/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 j/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.



j / Example 1.

Angular defect on the earth's surface *Example: 2*

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 k has angles that are each 90 degrees, and the sum is 270. This shows that the Earth's surface has intrinsic curvature.

This example suggests another way of measuring intrinsic curvature, in terms of the ratio 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 k, $r = (\pi/2)\rho$. (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 C/r is equal to 4, which is smaller than the Euclidean value of 2π.

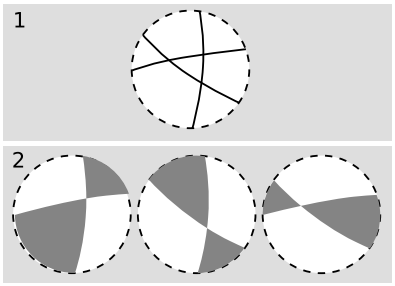
k / Example 2.

Let $\epsilon = \Sigma\theta - \pi$ 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.

Self-check: In elliptic geometry, what is the minimum possible

value of the quantity 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.



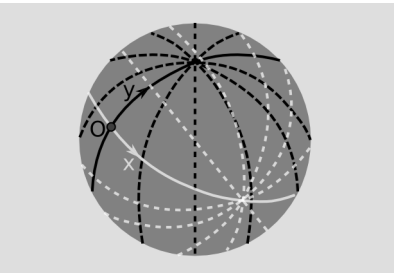
l / Proof that the angular defect of a triangle in elliptic geometry is proportional to its area. Each white circle represents the entire elliptic plane. The dashed line at the edge is not really a boundary; lines that go off the edge simply wrap back around. In the spherical model, the white circle corresponds to one hemisphere, which is identified with the opposite hemisphere.

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 l/1. Each pair of lines crosses at only one point (E1) and divides the plane into two lunes with its two vertices touching at this point, figure l/2. 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 P/π. The sum of the areas of our three lunes is $(P/\pi)(\alpha+\beta+\gamma)$, 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.

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, $K = \epsilon/A$. 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 $K = 1/\rho^2$.

Self-check: Verify the equation $K = 1/\rho^2$ by considering a triangle covering one octant of the sphere, as in example 2.

It is useful to introduce *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.



m / 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 $(0, \mathrm{d}y)$ at an angle that differs from perpendicular by the infinitesimal amount $\mathrm{d}\alpha$ (figure n). Construct the line $x' =$

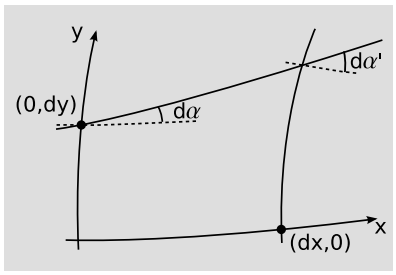
$\mathrm{d}x$, and let $\mathrm{d}\alpha'$ be the angle its perpendicular forms with ℓ. Then[1] the Gaussian curvature at O is

$$K = \frac{\mathrm{d}^2\alpha}{\mathrm{d}x\mathrm{d}y} \quad ,$$

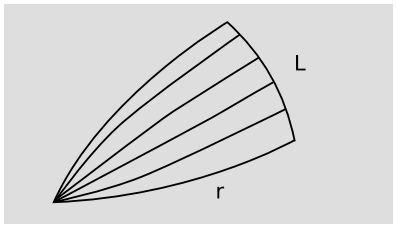
where $\mathrm{d}^2\alpha = \mathrm{d}\alpha' - \mathrm{d}\alpha$.

2. From a point P, emit a fan of rays at angles filling a certain range θ of angles in Gaussian polar coordinates (figure o). 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]

$$K = -3\frac{\mathrm{d}^2}{\mathrm{d}r^2}\left(\frac{L}{L_e}\right) \quad .$$



n / 1. Gaussian curvature can be interpreted as the failure of parallelism represented by $\mathrm{d}^2\alpha/\mathrm{d}x\mathrm{d}y$.



o / 2. Gaussian curvature as $L \neq r\theta$.

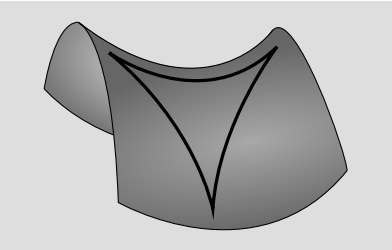
Let's now generalize beyond elliptic geometry. Consider a space modeled by a surface embedded in three dimensions, with geodesics defined as curves of extremal length, i.e., the curves made by a piece of string stretched taut across the surface. At a particular point P, we can always pick a coordinate system (x, y, z) such that the surface $z = \frac{1}{2}k_1x^2 + \frac{1}{2}k_2y^2$ 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 4 on page 63, 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_1k_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 $k_1^2 + k_2^2$ and k_1k_2. (We can't have, for example, just k_1^2, because that would change under an extrinsic rotation about the z axis.) Only k_1k_2 vanishes on a cylinder, so it is the only possible intrinsic curvature.

Eating pizza *Example: 3*

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.

[1] 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 $x = \mathrm{d}x$, so that $K = \mathrm{d}\epsilon/\mathrm{d}A = \mathrm{d}^2\alpha/\mathrm{d}x\mathrm{d}y$, where third powers of infinitesimals have been discarded.

[2] 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 $L/L_E = \sin u/u$, whose second derivative with respect to u is $-1/3$. Since $r = \rho u$, the second derivative of the same quantity with respect to r equals $-1/3\rho^2 = -K/3$.

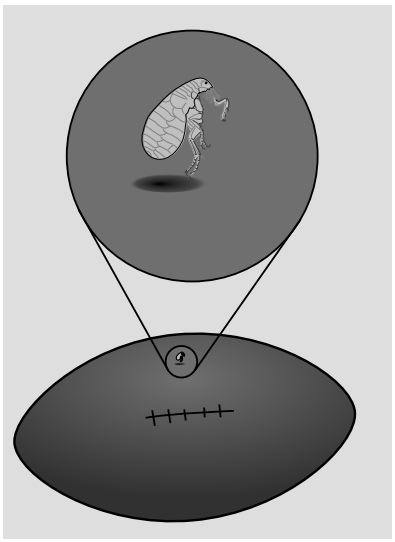


p / A triangle in a space with negative curvature has angles that add to less than π.

Elliptic and hyperbolic geometry *Example: 4*

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 $z = k_1x^2 + k_2y^2$ 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 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.



q / A flea on the football cannot orient himself by intrinsic, local measurements.

A flea on a football *Example: 5*

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 $z = \frac{1}{2}k_1x^2 + \frac{1}{2}k_2y^2$ is invariant under the parity inversion $x \to -x$, $y \to -y$. 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.

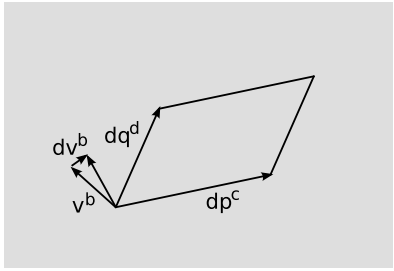
4.4 Curvature tensors

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. More elegantly, we expect that under a uniform rescaling of coordinates by a factor of μ, a curvature tensor should scale down by μ^{-2}.

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 in-

dices 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 $\mathrm{d}p^c$ and $\mathrm{d}q^d$ be two infinitesimal vectors, and use them to form a quadrilateral that is a good approximation to a parallelogram.[3] Parallel-transport vector v^b all the way around the parallelogram. When it comes back to its starting place, it has a new value $v^b \to v^b + \mathrm{d}v^b$. Then the Riemann curvature tensor is defined as the tensor that computes $\mathrm{d}v^a$ according to $\mathrm{d}v^a = R^a{}_{bcd} v^b \mathrm{d}p^c \mathrm{d}q^d$. (There is no standardization in the literature of the order of the indices.)



r / The definition of the Riemann tensor. The vector v^b changes by $\mathrm{d}v^b$ when parallel-transported around the approximate parallelogram. (v^b is drawn on a scale that makes its length comparable to the infinitesimals $\mathrm{d}p^c$, $\mathrm{d}q^d$, and $\mathrm{d}v^b$; in reality, its size would be greater than theirs by an infinite factor.)

A symmetry of the Riemann tensor *Example: 6*

If vectors $\mathrm{d}p^c$ and $\mathrm{d}q^d$ lie along the same line, then $\mathrm{d}v^a$ must vanish, and interchanging $\mathrm{d}p^c$ and $\mathrm{d}q^d$ simply reverses the direction of the circuit around the quadrilateral, giving $\mathrm{d}v^a \to -\mathrm{d}v^a$. This shows that $R^a{}_{bcd}$ must be antisymmetric under interchange of the indices c and d, $R^a{}_{bcd} = -R^a{}_{bdc}$.

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 $K = \mathrm{d}\epsilon/\mathrm{d}A$.

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 $\mathrm{d}p^c$ and $\mathrm{d}q^d$. 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.

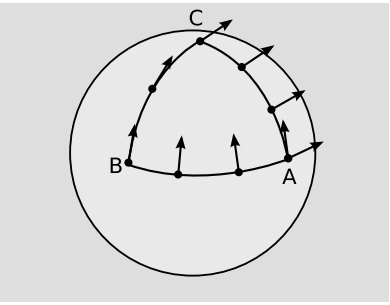
Curvature tensors on a sphere *Example: 7*

Let's find the curvature tensors on a sphere of radius ρ.

Construct normal coordinates (x, y) with origin O, and let vectors $\mathrm{d}p^c$ and $\mathrm{d}q^d$ 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 magni-

[3] Section 4.8 discusses the sense in which this approximation is good enough.

tude, then $R^x{}_{yxy}$ equals the angular deficit of the quadrilateral. Comparing with the definition of the Gaussian curvature, we find $R^x{}_{yxy} = K = 1/\rho^2$. 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.



s / The change in the vector due to parallel transport around the octant equals the integral of the Riemann tensor over the interior.

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 s. The area of the octant is $(\pi/2)\rho^2$, and multiplying it by the Riemann tensor, we find that the defect in parallel transport is $\pi/2$, 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 $dx/d\phi = \rho \sin\theta$ and $dy/d\theta = \rho$. The result is $R^\phi{}_{\theta\phi\theta} = R^x{}_{yxy}(dy/d\theta)^2 = 1$ and $R^\theta{}_{\phi\theta\phi} = R^y{}_{xyx}(dx/d\phi)^2 = \sin^2\theta$. 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 $\mathrm{diag}(K, K)$.

4.5 Some order-of-magnitude estimates

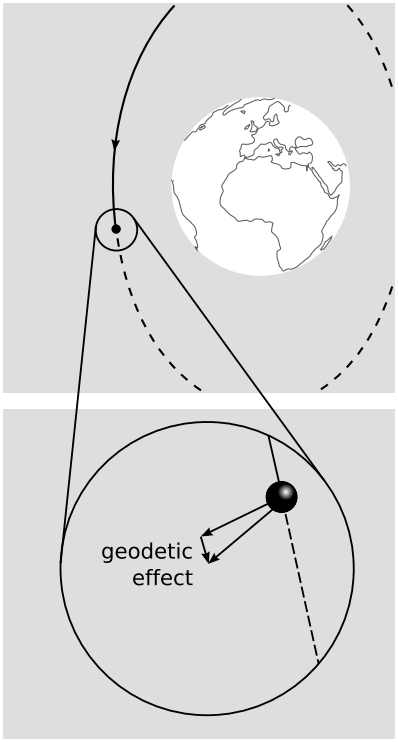
As a general proposition, calculating an order-of-magnitude estimate of a physical effect requires an understanding of 50% of the physics, while an exact calculation requires about 75%.[4] We've reached

[4] 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 point where it's reasonable to attempt a variety of order-of-magnitude estimates.

4.5.1 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.



t / 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 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 $1/c^2$. In units with $c = 1$, this factor is unnecessary. In ordinary metric units, the $1/c^2$ 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

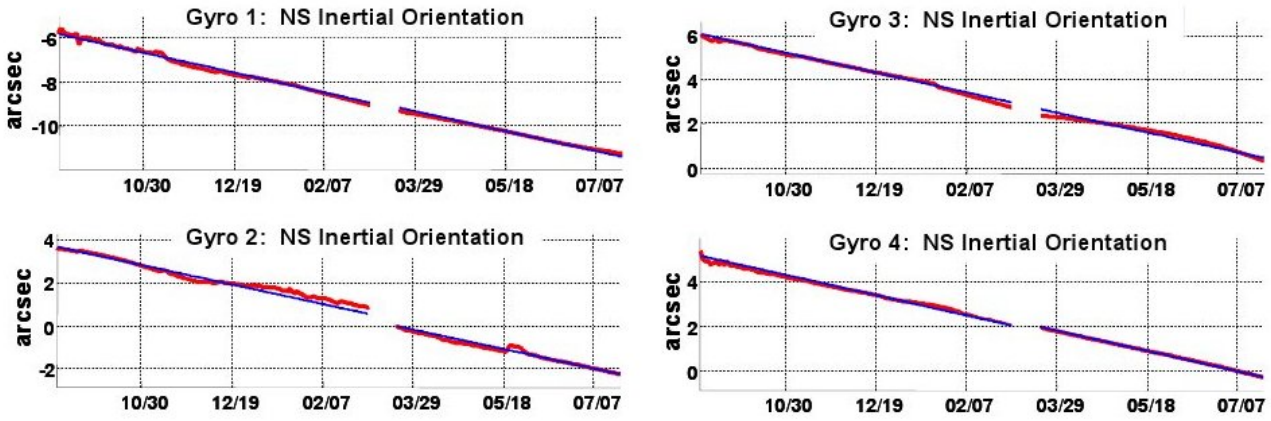
$$\Delta\theta \sim \frac{nGM}{c^2 r} \quad .$$

There is also a Thomas precession. Like the spacetime curvature effect, it is proportional to 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 $r = r_e + (650\text{ km})$ and $n \approx 5000$ (orbiting once every 90 minutes for the 353-day duration of the experiment), giving $\Delta\theta \sim 3 \times 10^{-6}$ radians. Figure u shows the actual results[5] the four gyroscopes aboard the probe. The precession was about 6 arc-seconds, or 3×10^{-5} radians. Our crude estimate

[5] http://einstein.stanford.edu/content/final_report/GPB_Final_NASA_Report-020509-web.pdf

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 on page 140.

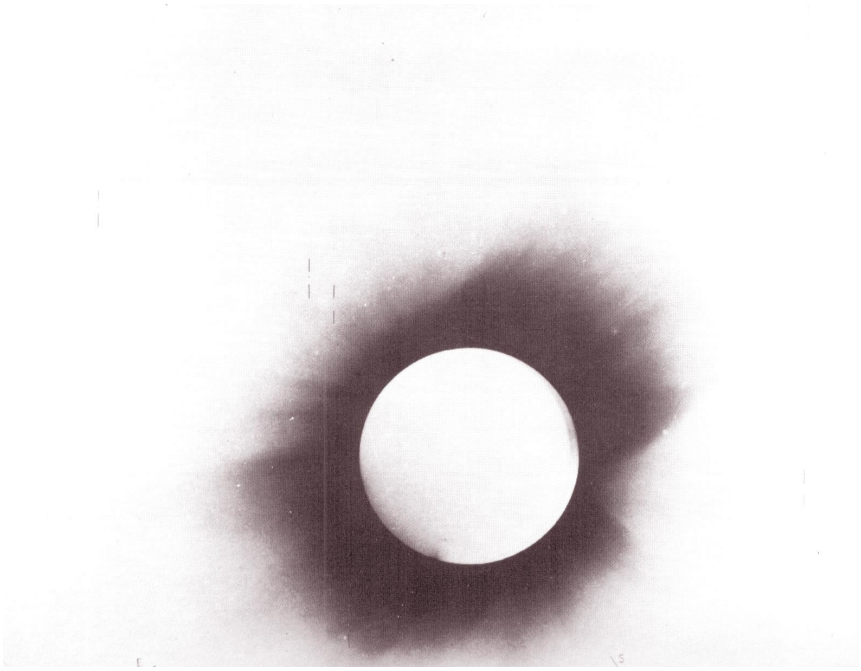


u / Precession angle as a function of time as measured by the four gyroscopes aboard Gravity Probe B.

4.5.2 Deflection of light rays

In the discussion of the momentum four vector in section 3.2.2, 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 $K = \mathrm{d}\epsilon/\mathrm{d}A$. 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 in section 4.5.1, and the result is $\epsilon \sim GM/c^2r$, 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 GM/c^2r, where M is the sun's mass, and r



v / 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.

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 on page 146.

4.6 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: $R_{ab} = 0$. Einstein posited this equation based essentially on the considerations laid out in section 4.1.

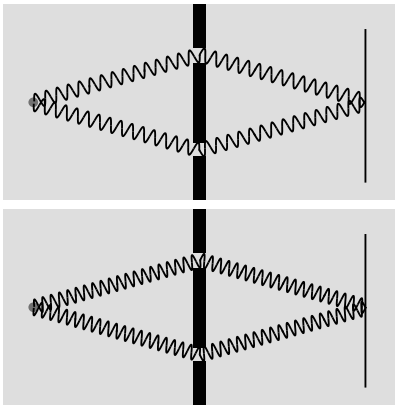
But just knowing that a certain tensor vanishes identically in the space surrounding the earth clearly doesn't tell us anything explicit about the structure of the spacetime in that region. We want to know the metric. As suggested at the beginning of the chapter, we

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}.

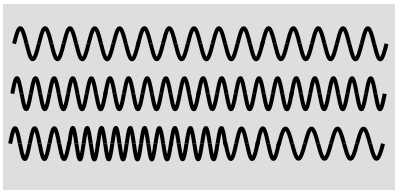
4.6.1 The covariant derivative in electromagnetism

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 $\Psi \rightarrow -\Psi$ does not change the results of experiments. As a less trivial example, we can redefine the ground of our electrical potential, $\Phi \rightarrow \Phi + \delta\Phi$, 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 $E = hf$. 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.



w / 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.



x / 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.

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 $\Psi \rightarrow \Psi' = e^{i\alpha}\Psi$, where α is a constant. The transformation $\Phi \rightarrow \Phi - \delta\Phi$ is also allowed, and it gives $\alpha(t) = (q\delta\Phi/\hbar)t$, 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 $\Phi \rightarrow \Phi - \delta\Phi$ is now $A_b \rightarrow A_b - \partial_b\alpha$, where A_b is the electromagnetic potential four-vector (section 3.2.5, page 85).

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:

$$\begin{aligned}\partial_b\Psi &\rightarrow \partial_b\left(e^{i\alpha}\Psi\right)\\ &= e^{i\alpha}\partial_b\Psi + i\partial_b\alpha\left(e^{i\alpha}\Psi\right)\\ &= \left(\partial_b + A'_b - A_b\right)\Psi'\end{aligned}$$

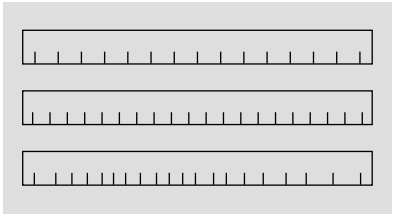
To avoid getting incorrect results, we have to do the substitution $\partial_b \rightarrow \partial_b + ieA_b$, where the correction term compensates for the change of gauge. We call the operator ∇ defined as

$$\nabla_b = \partial_b + ieA_b$$

the *covariant derivative*. It gives the right answer regardless of a change of gauge.

4.6.2 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.



y / These three rulers represent three choices of coordinates. As in figure x on page 112, 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 $\mathrm{d}v/\mathrm{d}x$, 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 m pn page 104 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 $V = v\,\mathrm{d}X/\mathrm{d}x$, and differentiation with respect to X will not give zero, because the factor $\mathrm{d}X/\mathrm{d}x$ 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 $\mathrm{d}/\mathrm{d}X$, 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 $\mathrm{d}/\mathrm{d}X$ with

$$\nabla_X = \frac{\mathrm{d}}{\mathrm{d}X} - G^{-1}\frac{\mathrm{d}G}{\mathrm{d}X} \quad .$$

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 $\nabla_X T = \mathrm{d}T/\mathrm{d}X - G^{-1}(\mathrm{d}G/\mathrm{d}X)T$.

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 $\mathrm{d}V/\mathrm{d}T = 9.8\ \mathrm{m/s^2}$. This acceleration cannot be a tensor, because we could make it vanish by changing from Earth-fixed 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 $(1/y)\mathrm{d}y/\mathrm{d}x$, which is known as a logarithmic derivative, since it equals $\mathrm{d}(\ln y)/\mathrm{d}x$. 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,

$$\nabla_X = \frac{\mathrm{d}}{\mathrm{d}X} - \frac{1}{2}G^{-1}\frac{\mathrm{d}G}{\mathrm{d}X} \qquad .$$

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:

$$\begin{aligned}\nabla_a v^b &= \partial_a v^b + \Gamma^b{}_{ac} v^c \\ \nabla_a v_b &= \partial_a v_b - \Gamma^c{}_{ba} v_c \qquad ,\end{aligned}$$

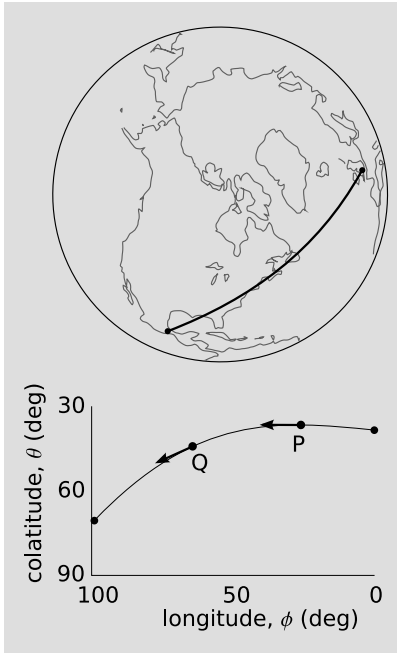
where $\Gamma^b{}_{ac}$, called the Christoffel symbol, does not transform like a tensor, and involves derivatives of the metric. The explicit computation of the Christoffel symbols from the metric is deferred until section 4.9, but the intervening sections 4.7 and 4.8 can be omitted on a first reading without loss of continuity.

Christoffel symbols on the globe *Example: 8*

As a qualitative example, consider the geodesic airplane trajectory shown in figure z, 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 $\theta = 0$.)

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.



z / Example 8.

$\Gamma^{\theta}{}_{\phi\phi}$ is computed in example 9 on page 125.

Symmetry also requires that this Christoffel symbol be independent of ϕ, and it must also be independent of the radius of the sphere.

For higher-rank tensors, we just add more correction terms, e.g.,

$$\nabla_a U_{bc} = \partial_a U_{bc} - \Gamma^d{}_{ba} U_{dc} - \Gamma^d{}_{ca} U_{bd}$$

or

$$\nabla_a U^c_b = \partial_a U^c_b - \Gamma^d{}_{ba} U^c_d + \Gamma^c{}_{ad} U^d_b \quad .$$

With the partial derivative ∂_a, it does not make sense to use the metric to raise the index and form ∂^a. It *does* make sense to do so with covariant derivatives, so $\nabla^a = g^{ab}\nabla_b$ is a correct identity.

Comma and semicolon notation

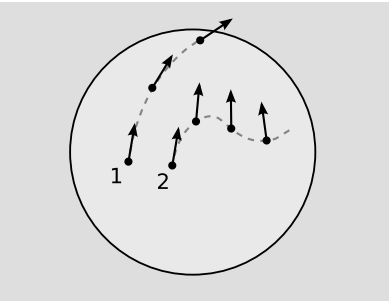
Some authors use superscripts with commas and semicolons to indicate partial and covariant derivatives. The following equations give equivalent notations for the same derivatives:

$$\partial_a X_b = X_{b,a}$$
$$\nabla_a X_b = X_{b;a}$$
$$\nabla^a X_b = X_b{}^{;a}$$

4.7 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.

4.7.1 Characterization of the geodesic



aa / 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 geodesic can be defined as a world-line that preserves tangency under parallel transport, aa. 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."

A curve can be specified by giving functions $x^i(\lambda)$ for its coordinates, where λ is a real parameter. A vector lying tangent to the curve can then be calculated using partial derivatives, $T^i = \partial x^i/\partial\lambda$. 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^i, which is tangent by construction. It would therefore be convenient if T^i 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^i 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^i to be purely spatial coordinates, and λ to be a universal time coordinate. We would then interpret T^i 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.

4.7.2 Covariant derivative with respect to a parameter

The notation of section 4.6 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^i with respect to λ as $\nabla_\lambda T^i$, 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:

$$\frac{\mathrm{d}f}{\mathrm{d}\lambda} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial \lambda} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial \lambda} \quad .$$

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

$$\begin{aligned}\nabla_\lambda T^i &= (\nabla_b T^i)\frac{\mathrm{d}x^b}{\mathrm{d}\lambda} \\ &= \left(\partial_b T^i + \Gamma^i{}_{bc}T^c\right)\frac{\mathrm{d}x^b}{\mathrm{d}\lambda} \quad .\end{aligned}$$

4.7.3 The geodesic equation

Recognizing $\partial_b T^i \mathrm{d}x^b/\mathrm{d}\lambda$ as a total non-covariant derivative, we find

$$\nabla_\lambda T^i = \frac{\mathrm{d}T^i}{\mathrm{d}\lambda} + \Gamma^i{}_{bc}T^c\frac{\mathrm{d}x^b}{\mathrm{d}\lambda} \quad .$$

Substituting $\partial x^i/\partial\lambda$ for T^i, and setting the covariant derivative equal to zero, we obtain

$$\frac{\mathrm{d}^2x^i}{\mathrm{d}\lambda^2} + \Gamma^i{}_{bc}\frac{\mathrm{d}x^c}{\mathrm{d}\lambda}\frac{\mathrm{d}x^b}{\mathrm{d}\lambda} = 0.$$

This is known as the geodesic equation.

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 7). 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 2 on page 166.

4.7.4 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 on page 15; 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^i and their derivatives $\mathrm{d}x^i/\mathrm{d}\lambda$. Also, set a small step-size $\Delta\lambda$ by which to increment λ at each step below.
2. For each i, calculate $\mathrm{d}^2x^i/\mathrm{d}\lambda^2$ using the geodesic equation.
3. Add $(\mathrm{d}^2x^i/\mathrm{d}\lambda^2)\Delta\lambda$ to the currently stored value of $\mathrm{d}x^i/\mathrm{d}\lambda$.
4. Add $(\mathrm{d}x^i/\mathrm{d}\lambda)\Delta\lambda$ to x^i.
5. Add $\Delta\lambda$ to λ.
6. Repeat steps 2-5 until the 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 4.9.2 on page 126.

4.8 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.

Self-check: Interpret the mathematical meaning of the equation $\Gamma^a{}_{[bc]} = 0$, which is expressed in the notation introduced on page 68.

4.8.1 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.

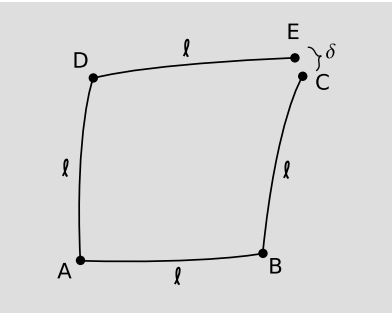
But just because derivatives of scalars don't require special treatment *for this particular reason*, that doesn't mean they are guaranteed to behave as we intuitively expect, in the strange world of coordinate-invariant relativity.

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.

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 $\partial^2 T/\partial x \partial y$. 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.[6]



ab / Measuring $\partial^2 T/\partial x \partial y$ for a scalar T.

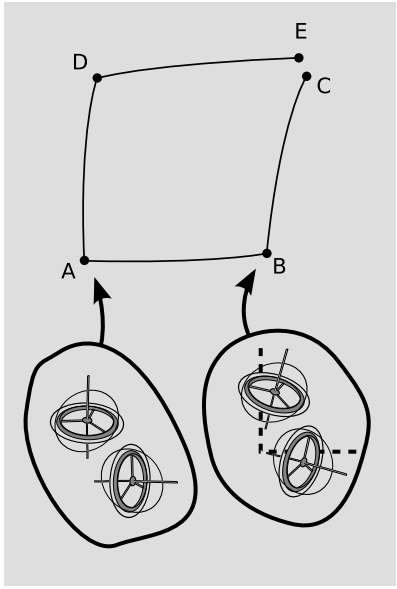
None of this seems surprising yet, but there are now two qualitatively different ways that her analysis of her data could turn out, indicating qualitatively different things about the laws of physics in her universe. The definition of the derivative as a limit requires that she repeat the experiment at smaller scales. As $\ell \to 0$, the result for $\partial^2 T/\partial x \partial y$ 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

[6] This point was mentioned on page 107, in connection with the definition of the Riemann tensor.

$\nabla_x\nabla_y - \nabla_y\nabla_x$ 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.



ac / 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.

4.8.2 The torsion tensor

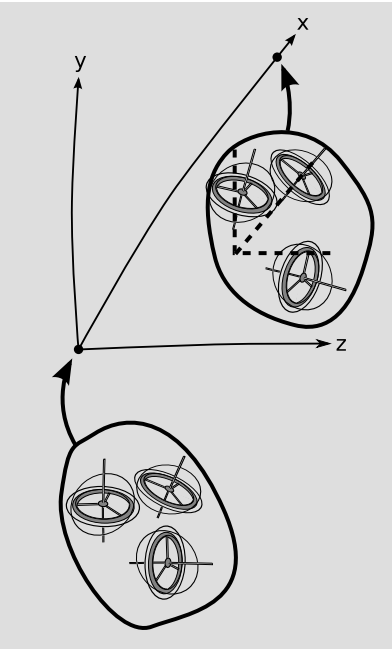
Since torsion is odd under parity, it must be represented by an odd-rank tensor, which we call $\tau^c{}_{ab}$ and define according to

$$(\nabla_a\nabla_b - \nabla_b\nabla_a)f = -\tau^c{}_{ab}\nabla_c f \qquad ,$$

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 $\Gamma^a{}_{bc} = \Gamma^a{}_{cb}$ of the Christoffel symbols. Using the notation introduced on page 68, $\Gamma^a{}_{[bc]} = 0$ if $\tau = 0$.

Self-check: Use an argument similar to the one in example 5 on page 106 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 $x_a \to -x_a$, i.e., a parity inversion plus a time-reversal, PT.



ad / 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 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 λ, μ, 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 page 68, we have

$$T_{(abc)} = \frac{1}{3!}\Sigma T_{abc}$$
$$T_{[abc]} = \frac{1}{3!}\Sigma \epsilon^{abc} T_{abc} \quad ,$$

where the sums are over all permutations of the indices, and in the second line the totally antisymmetric Levi-Civita tensor is defined by $\epsilon^{abc} = \epsilon^{bca} = \epsilon^{cab} = +1$, $\epsilon^{bac} = \epsilon^{acb} = \epsilon^{cba} = -1$, and $\epsilon^{\lambda\mu\nu} = 0$ for all other combinations of indices. In other words, $\epsilon^{\lambda\mu\nu}$ is $+1$ if $\lambda\mu\nu$ represent an even permutation of abc (i.e., they can be obtained from abc by an even number of pairwise swaps), -1 for odd permutations, and 0 otherwise (i.e., if $\lambda\mu\nu$ is not a permutation of abc). 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.

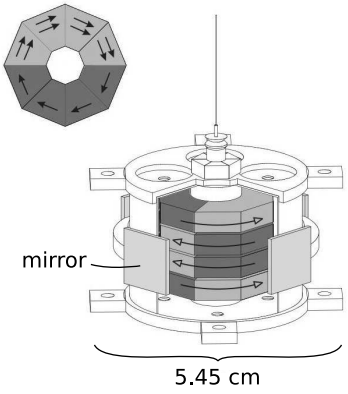
In two dimensions, there are no totally antisymmetric rank-3 tensors, because we can't write three indices without repeating one. In three dimensions, an antisymmetric rank-3 tensor is simply a multiple of the Levi-Civita tensor, so a totally antisymmetric torsion, if it exists, is represented by a single 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, τ_{xyz}, τ_{tyz}, τ_{txz}, and τ_{txy}. In other words, an antisymmetric torsion of 3+1 spacetime can be represented by a four-vector, $\tau^a = \epsilon^{abcd}\tau_{bcd}$.

4.8.3 Experimental searches for torsion

One way of stating the equivalence principle (see p. 89) 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, $\sqrt{\hbar G/c^3} \sim 10^{-35}$ 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 ae shows a torsion pendulum used in an experiment by the Eöt-Wash group at the University of Washington.[7] If torsion exists, then the intrinsic spin $\boldsymbol{\sigma}$ of an electron should have an energy $\boldsymbol{\sigma} \cdot \boldsymbol{\tau}$, where $\boldsymbol{\tau}$ 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 $\boldsymbol{\sigma} \cdot \boldsymbol{\tau}$ 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



ae / The University of Washington torsion pendulum used to search for torsion. The light gray wedges are Alnico, the darker ones $SmCo_5$. 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.

[7] http://www.npl.washington.edu/eotwash/publications/pdf/lowfrontier2.pdf

will tend to precess around the axis defined by $\boldsymbol{\tau}$.

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 $m_e^2/m_P \sim 10^{-17}$ eV, where m_e is the mass of the electron and $m_P = \sqrt{\hbar c/G} \approx 10^{19}$ 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.[8]

4.9 From metric to curvature

4.9.1 Finding Γ given g

We've already found the Christoffel symbol in terms of the metric in one dimension. Expressing it in tensor notation, we have

$$\Gamma^d{}_{ba} = \frac{1}{2} g^{cd} \left(\partial_? g_{??} \right) \qquad ,$$

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

$$\Gamma^b{}_{ac} = \frac{1}{2} g^{db} \left(L \partial_c g_{ab} + M \partial_a g_{cb} + N \partial_b g_{ca} \right) \qquad ,$$

[8] Carroll and Field, `http://arxiv.org/abs/gr-qc/9403058`

where L, M, and N are constants. Consistency with the one-dimensional expression requires $L + M + N = 1$, 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 $1/\sqrt{k}$. 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 the M correction, so $|M| = |N|$. We find $L = M = -N = 1$.

Self-check: 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

$$\Gamma^c{}_{ab} = \frac{1}{2} g^{cd} \left(\partial_a g_{bd} + \partial_b g_{ad} - \partial_d g_{ab} \right) \qquad .$$

One can readily go back and check that this gives $\nabla_c g_{ab} = 0$. In fact, the calculation 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.[9]

Self-check: In the case of 1 dimension, show that this reduces to the earlier result of $-(1/2)\mathrm{d}G/\mathrm{d}X$.

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.

Christoffel symbols on the globe, quantitatively *Example: 9*

In example 8 on page 115, we inferred the following properties

[9] http://arxiv.org/abs/0903.2085

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 $ds^2 = R^2 d\theta^2 + R^2 \sin^2\theta d\phi^2$. 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 $\Gamma^{\theta}{}_{\phi\phi} = -\sin\theta\cos\theta$, which can be verified to have the properties claimed above.

4.9.2 Numerical solution of the geodesic equation

On page 118 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 on page 118 and its concrete realization below. For polar coordinates in a Euclidean plane, one can compute $\Gamma^{r}{}_{\phi\phi} = -r$ and $\Gamma^{\phi}{}_{r\phi} = 1/r$ (problem 1, page 130). 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).

Self-check: 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 on page 5.2.6, where we calculate the deflection of light rays in a gravitational field, one of the classic observational tests of general relativity.

4.9.3 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 on page 107 is a measure of the failure of covariant derivatives to commute:

$$(\nabla_a \nabla_b - \nabla_b \nabla_a) A^c = A^d R^c{}_{dab}$$

A tedious calculation now gives R in terms of the Γs:

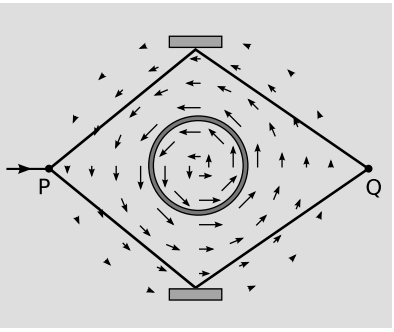
$$R^a{}_{bcd} = \partial_c \Gamma^a{}_{db} - \partial_d \Gamma^a{}_{cb} + \Gamma^a{}_{ce}\Gamma^e{}_{db} - \Gamma^a{}_{de}\Gamma^e{}_{cb}$$

This is given as another example later in Brewin's manual for applying Cadabra to general relativity.[10] (Brewin writes the upper index in the second slot of R.)

4.9.4 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	$\Gamma^c{}_{ab}$
A second differentiation gives the directly observable field(s) ...	**E** and **B**	$R^c{}_{dab}$



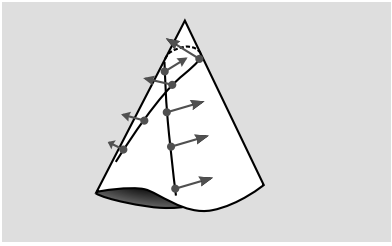
af / The Aharonov-Bohm effect. An electron enters a beam splitter at P, and is sent out in two different directions. The two parts of the wave are reflected so that they reunite at Q. The arrows represent the vector potential **A**. The observable magnetic field **B** is zero everywhere outside the solenoid, and yet the interference observed at Q depends on whether the field is turned on. See page 85 for further discussion of the **A** and **B** fields of a solenoid.

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. The solenoid in figure af has $\mathbf{B} = 0$ 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

[10] http://arxiv.org/abs/0903.2085

potential does not vanish outside the solenoid, and as we've seen on page 86, 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 ag 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.



ag / 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 $\mathbf{A}$ 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 $\mathbf{A}$ 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.

Geodetic effect and structure of the source *Example: 10*

▷ In section 4.5.1 on page 109, 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.

Problems

Key

The notation √indicates that a computerized answer check is available online.

1 Show, as claimed on page 126, that for polar coordinates in a Euclidean plane, $\Gamma^r{}_{\phi\phi} = -r$ and $\Gamma^\phi{}_{r\phi} = 1/r$.

2 Partial derivatives commute with partial derivatives. Covariant derivatives don't commute with covariant derivatives. Do covariant derivatives commute with partial derivatives?

3 (a) For an object moving in a circle at constant speed, the dot product of the classical three-vectors $\mathbf{v}$ and $\mathbf{a}$ 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.

4 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.

5 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.

6 Verify the transformation of the metric given in example 8 on page 88.

7 Show that if the differential equation for geodesics on page 116 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.

8 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.

Chapter 5
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 4.5.1 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 4.1, page 100, the field equations in a vacuum are simply $R_{ab} = 0$.

a / A Swiss commemorative coin shows the vacuum field equation.

5.1 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.

5.1.1 The event horizon of an accelerated observer

Consider the uniformly accelerated observer described in examples 2 on page 81 and 8 on page 88. Recalling these earlier results, we have for the ship's equation of motion in an inertial frame

$$x = \frac{1}{a}\left(\sqrt{1+a^2t^2} - 1\right) \quad ,$$

and for the metric in the ship's frame

$$\begin{aligned} g'_{t't'} &= (1+ax')^2 \\ g'_{x'x'} &= -1 \quad . \end{aligned}$$

Since this metric was derived by a change of coordinates from a flat-space metric, and the Ricci curvature is an intrinsic property, we

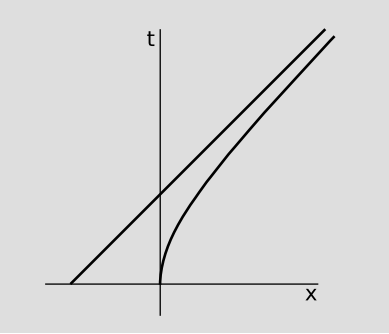
expect that this one also has zero Ricci curvature. This is straightforward to verify. The nonvanishing Christoffel symbols are

$$\Gamma^{t'}{}_{x't'} = \frac{a}{1+ax'} \quad \text{and} \quad \Gamma^{x'}{}_{t't'} = a(1+ax') \qquad .$$

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.

Self-check: Verify these facts.

This seemingly routine exercise now leads us into some very interesting territory. Way back on page 6, 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.



b / A spaceship (curved world-line) moves with an acceleration perceived as constant by its passengers. The photon (straight world-line) come closer and closer to the ship, but will never quite catch up.

We observe that $x(t)$ approaches the asymptote $x = t - 1/a$. 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 5.3.2 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 $x' = -1/a$. An observer aboard the ship reasons as follows. If I start out with a head-start of $1/a$ 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 $\mathrm{d}s^2 > 0$. If I were to detect a particle emitted at that event, it would violate the laws of physics, since material particles must have $\mathrm{d}s^2 > 0$, 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.

We've already seen in example 9 on page 37 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.

5.1.2 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 of 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

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

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:

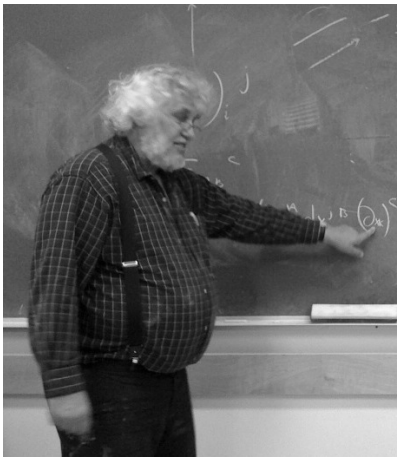
$$\begin{pmatrix} 1 & 0 \\ 0 & \sqrt{2} \end{pmatrix}$$

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 time-reversal 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.

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.

5.1.3 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, $\Delta E \Delta t \lesssim h$. 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 $\sqrt{g'_{t't'}} = \sqrt{(1+ax')^2}$. Say the pair is created right near the horizon, at $x' = -1/a$. 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 $a\Delta x' = a\Delta t'$, which by the uncertainty principle should be at least ha/E, so that when the photon is observed by B, its energy is $E(ha/E) = ha$.



c / Bill Unruh (1945-).

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 $E \sim kT$, where k is Boltzmann's constant. We conclude that the horizon acts like a black-body radiator with a temperature $T \sim ha/k$. The more careful treatment by Unruh shows that the exact relation is $T = ha/4\pi^2 k$, or $ha/4\pi^2 kc$ 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

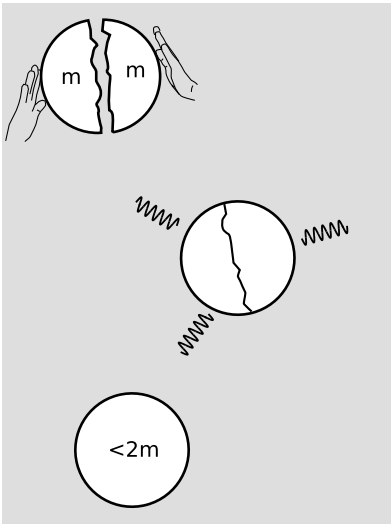
To detect Unruh radiation experimentally, we would ideally like to be able to accelerate a detector and let it detect the radiation. This is clearly impractical. The third line shows that it is possible to impart very large linear accelerations to subatomic particles, but then one can only hope to infer 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.[1] 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.

5.2 The Schwarzschild metric

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. 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 $\hat{\mathbf{r}}/r^2$, 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 $R_{ab} = 0$ 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 d, 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



d / The field equations of general relativity are nonlinear.

[1] http://xxx.lanl.gov/abs/gr-qc/9605032

proportional to m^2, so the gravitational mass of the combined system has been reduced from $2m$ to $2m-(\ldots)m^2$, where $\ldots$ is roughly G/c^2r. There is a nonlinear dependence of the gravitational field on the masses.

Self-check: 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.

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 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 $G/c^2 = 7 \times 10^{-28}$ 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 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.

5.2.1 The zero-mass case

First let's demonstrate the trivial solution with flat spacetime. In spherical coordinates, we have

$$\mathrm{d}s^2 = \mathrm{d}t^2 - \mathrm{d}r^2 - r^2\mathrm{d}\theta^2 - r^2\sin^2\theta\mathrm{d}\phi^2 \qquad .$$

The nonvanishing Christoffel symbols (ignoring swaps of the lower indices) are:

$$\begin{aligned}
\Gamma^\theta{}_{r\theta} &= \frac{1}{r} \\
\Gamma^\phi{}_{r\phi} &= \frac{1}{r} \\
\Gamma^r{}_{\theta\theta} &= -r \\
\Gamma^r{}_{\phi\phi} &= -r\sin^2\theta \\
\Gamma^\theta{}_{\phi\phi} &= -\sin\theta\cos\theta \\
\Gamma^\phi{}_{\theta\phi} &= \cot\theta
\end{aligned}$$

Self-check: 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 ctensor

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 on page 45.

The following Maxima program calculates the Christoffel symbols found in section 5.2.1.

```
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 `lg` 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 ${\Gamma'}_{bc}{}^{a}$ with the indices swapped around a little compared to the convention ${\Gamma^{a}}_{bc}$ 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
```

```
(%t10)                    mcs        = - r sin (theta)
                             4, 4, 2

(%t11)                 mcs        = - cos(theta) sin(theta)
                          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.

5.2.2 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: $-r^2\mathrm{d}\theta^2 - r^2\sin^2\theta\mathrm{d}\phi^2$. 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 $t \to -t$, $\theta \to -\theta$, and $\phi \to -\phi$, so we can't have any off-diagonal elements. The result is that we have narrowed the metric down to something of the form

$$\mathrm{d}s^2 = h(r)\mathrm{d}t^2 - k(r)\mathrm{d}r^2 - r^2\mathrm{d}\theta^2 - r^2\sin^2\theta\mathrm{d}\phi^2 \quad ,$$

where both h and k approach 1 for $r \to \infty$, where spacetime is flat.

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 $-m/r^2$, since nonrelativistic physics applies there. We have

$$\nabla_t v^r = \partial_t v^r + \Gamma^r{}_{tc} v^c \quad .$$

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

$(t, r, \ldots)$ 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 $-\Gamma^r{}_{tt} = -\frac{1}{2}g^{rr}\partial_r g_{tt} = -\frac{1}{2}h'/k$. Setting this equal to $-m/r^2$, we find $h'/k = 2m/r^2$ for $r \gg m$. Since $k \approx 1$ for large r, we have

$$h' \approx \frac{2m}{r^2} \quad \text{for } r \gg m \qquad .$$

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 tensor gymnastics, in example 3 on page 82.

Integrating, we find that for large r, $h = 1 - 2m/r$.

5.2.3 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 m/r to expand around, so let's try to write h and k as series of the form

$$h = \Sigma_{n=0}^{\infty} a_k \left(\frac{m}{r}\right)^n$$
$$k = \Sigma_{n=0}^{\infty} b_k \left(\frac{m}{r}\right)^n$$

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 $R_t t$, which looked as simple as any of them. It appears to vary as r^{-4} for large r, so let's evaluate $\lim_{r\to\infty}\left(r^4 R_{tt}\right)$:

```
9   limit(r^4*ric[1,1],r,inf);
```

The result is $(b_1-2)/2$, so let's set $b_1 = 2$. The approximate solution we've found so far (reinserting the m's),

$$\mathrm{d}s^2 \approx \left(1 - \frac{2m}{r}\right)\mathrm{d}t^2 - \left(1 + \frac{2m}{r}\right)\mathrm{d}r^2 - r^2\mathrm{d}\theta^2 - r^2\sin^2\theta\mathrm{d}\phi^2 \quad ,$$

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:

$$\begin{array}{ll} a_0 = 1 & b_0 = 1 \\ a_1 = -2 & b_1 = 2 \\ a_2 = 0 & b_2 = 4 \\ a_3 = 0 & b_2 = 8 \end{array}$$

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 $1/(1-2/r)$, and when we reinsert the factor of m in the only way that makes the units work, we get $1/(1-2m/r)$. The result for the metric is

$$\mathrm{d}s^2 = \left(1 - \frac{2m}{r}\right)\mathrm{d}t^2 - \left(\frac{1}{1-2m/r}\right)\mathrm{d}r^2 - r^2\mathrm{d}\theta^2 - r^2\sin^2\theta\mathrm{d}\phi^2 \quad .$$

A quick calculation in Maxima demonstrates that this 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.

5.2.4 Geodetic effect

As promised in section 4.5.1, we now calculate the geodetic effect on Gravity Probe B, including all the niggling factors of 3 and π.

As a warmup, consider the flat-space case, in plane Euclidean polar coordinates (r, ϕ). Parallel-transport of a gyroscope's angular momentum around a circle of constant r gives

$$\begin{aligned} \nabla_\phi L^\phi &= 0 \\ \nabla_\phi L^r &= 0 \end{aligned} \quad .$$

Computing the covariant derivatives, and we have

$$\begin{aligned} 0 &= \partial_\phi L^\phi + \Gamma^\phi{}_{\phi r} L^r \\ 0 &= \partial_\phi L^r + \Gamma^r{}_{\phi\phi} L^\phi \end{aligned} \quad .$$

The Christoffel symbols are $\Gamma^{\phi}{}_{\phi r} = 1/r$ 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

$$\begin{aligned} P' &= Q \\ Q' &= -P \end{aligned} \quad .$$

The solution is $P = \sin\phi$, $Q = \cos\phi$. For each orbit (2π change in ϕ), the basis vectors rotate by 2π, so the angular momentum vector once again has the same components. In other words, it hasn't really changed at all. However, the gyroscopes do *not* return to the same orientation relative to a distant reference point, such as the star IM Pegasi that was used by Gravity Probe B. Due to Thomas precession (section 1.10.4), they precess at the rate $|(1/2)\mathbf{a} \times \mathbf{v}| = \omega^3 r^2/2$. Integrating this over n periods, $\Delta T = 2\pi n/\omega$, gives $n\pi\omega^2 r^2 = n\pi g r$, where g is the acceleration of gravity. Substituting $g = m/r^2$, the Thomas precession is $\Delta\theta = n\pi m/r$. The direction of the Thomas precession effect is always counter the direction of motion, i.e., if the satellite is orbiting clockwise, the precession is counterclockwise The Thomas precession will end up accounting for one third of the total observed effect.

Using the actual 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

$$\begin{aligned} P' &= Q \\ Q' &= -(1-\epsilon)P \end{aligned} \quad ,$$

where $\epsilon = 2m$. The solution is $P = \sin\omega'\phi$, $Q = \omega'\cos\omega'\phi$, where $\omega' = \sqrt{1-\epsilon}$. 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 $1 - m/r$. The effect is in the same direction as the effect of the Thomas precession, and the precession accumulated over n periods is $2\pi nm/r$.

Adding the contributions from the Thomas precession and spacetime curvature, we have $3\pi nm/r$. In SI units, this is $3\pi nGm/c^2 r$. Using the (relatively crude) data from section 1.10.4, we find $\Delta\theta = 3 \times 10^{-5}$ radians, in excellent agreement with the data shown in figure u on page 110.

5.2.5 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. In other words, Mercury spends more time near perihelion than it should classically. During this time, it sweeps out a greater angle than classically expected, so that when it flies back out and away from the sun, its orbit has rotated counterclockwise.

We can easily understand why there should be such an effect according to general relativity, by passing to the limit in which the relativistic effect is very large. Consider the orbits of material objects in the vicinity of a black hole. An object that passes through the event horizon spends infinite time at its "perihelion" and never emerges again. Applying this reasoning to the case of Mercury's orbit, we find that an effect in the observed direction is expected.

Based on the examples in section 4.5, we also expect that the effect will be of order 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. Classically, these are energy and angular momentum.

Consider a rock falling directly toward the sun. The Schwarzschild metric is of the special form

$$\mathrm{d}s^2 = h(r)\mathrm{d}t^2 - k(r)\mathrm{d}r^2 - \dots \quad .$$

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 $\Delta r_1, \dots$ or $\mathrm{d}r_1, \dots$, but we avoid the Δ's or d's for convenience.) Approximating the geodesic

using two line segments, the proper time is

$$\begin{aligned} s &= s_1 + s_2 \\ &= \sqrt{h_1 t_1^2 - k_1 r_1^2} + \sqrt{h_2 t_2^2 - k_2 r_2^2} \\ &= \sqrt{h_1 t_1^2 - k_1 r_1^2} + \sqrt{h_2 (T - t_1)^2 - k_2 r_2^2} \end{aligned} \quad ,$$

where $T = t_1 + t_2$ is fixed. If this is to be extremized with respect to t_1, then $\mathrm{d}s/\mathrm{d}t_1 = 0$, which leads to

$$0 = \frac{h_1 t_1}{s_1} - \frac{h_2 t_2}{s_2} \quad ,$$

which means that

$$h \frac{\mathrm{d}t}{\mathrm{d}s} = g_{tt} \frac{\mathrm{d}x^t}{\mathrm{d}s} = \frac{\mathrm{d}x_t}{\mathrm{d}s}$$

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. 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.

Writing these two quantities out explicitly in terms of the contravariant coordinates, in the case of the Schwarzschild spacetime, we have

$$E = \left(1 - \frac{2m}{r}\right) \frac{\mathrm{d}t}{\mathrm{d}s}$$

and

$$L = r^2 \frac{\mathrm{d}\phi}{\mathrm{d}s}$$

for the conserved energy per unit mass and angular momentum per unit mass. In interpreting the energy 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; E includes all of these, and turns out to be less than the rest mass (i.e., less than 1) for a planet orbiting the sun.

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 $\theta = \pi/2$,

$$\begin{aligned} 1 &= g^{tt}p_t^2 - g^{rr}p_r^2 - g^{\phi\phi}p_\phi^2 \\ &= g^{tt}p_t^2 - g_{rr}(p^r)^2 - g^{\phi\phi}p_\phi^2 \\ &= \frac{1}{1-2m/r}E^2 - \frac{1}{1-2m/r}\left(\frac{\mathrm{d}r}{\mathrm{d}s}\right)^2 - \frac{1}{r^2}L^2 \qquad . \end{aligned}$$

Rearranging terms and writing $\dot{r}$ for $\mathrm{d}r/\mathrm{d}s$, this becomes

$$\dot{r}^2 = E^2 - (1-2m/r)(1+L^2/r^2)$$

or

$$\dot{r}^2 = E^2 - U^2$$

where

$$U^2 = (1-2m/r)(1+L^2/r^2) \qquad .$$

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 classical and nearly circular.

Classically, a circular orbit has radius $r = L^2/m$ and period $T = 2\pi L^3/m^2$.

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

$$\begin{aligned} r &= \frac{L^2}{2m}\left(1+\sqrt{1-12m^2/L^2}\right) \\ &\approx \frac{L^2}{m}(1-\epsilon) \qquad , \end{aligned}$$

where $\epsilon = 3(m/L)^2$. 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

$$\begin{aligned} k &= \frac{\mathrm{d}^2(U^2)}{\mathrm{d}r^2} \\ &= \frac{\mathrm{d}^2}{\mathrm{d}r^2}\left(1 - \frac{2m}{r} + \frac{L^2}{r^2} - \frac{2mL^2}{r^3}\right) \\ &= -\frac{4m}{r^3} + \frac{6L^2}{r^4} - \frac{24mL^2}{r^5} \\ &= 2L^{-6}m^4(1+2\epsilon) \end{aligned}$$

The period of the oscillations is

$$\begin{aligned}\Delta s_{osc} &= 2\pi\sqrt{2/k} \\ &= 2\pi L^3 m^{-2}(1-\epsilon) \qquad .\end{aligned}$$

The period of the azimuthal motion is

$$\begin{aligned}\Delta s_{az} &= 2\pi r^2/L \\ &= 2\pi L^3 m^{-2}(1-2\epsilon) \qquad .\end{aligned}$$

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 $2\pi\epsilon = 6\pi(m/L)^2 = 6\pi m/r$. 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 1966, Dicke and Goldenberg measured the oblateness optically, and found an unexpectedly high value, which would have put Mercury's orbit far out of agreement with general relativity. The oblateness turns out to be extremely difficult to measure, and 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 the development of a variety of alternatives to general relativity. This is important because 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.

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,[2] 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.

5.2.6 Deflection of light

As discussed on page 110, 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 m/r, the unitless ratio of the sun's mass to its radius. Expressed in SI units, this is Gm/c^2r, 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 in section 5.2.5. 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 $\mathrm{d}s = 0$; 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 4.9.2 on page 126. 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
```

[2] http://arxiv.org/abs/astro-ph/0407149

```
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 quadrant
58  prop_sec = prop*180.*3600/math.pi
59  print "final direction of propagation = %6.2f arc-seconds" % prop_sec
```

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 $4Gm/c^2r$. The unitless factor of 4 is in fact the correct result in the case of small deflections, i.e., for $m/r \ll 1$.

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 $m/r \ll 1$ fails. For $m/r = 0.3$, the numerical techique gives a deflection of 222 degrees, whereas the weak-field approximation $4Gm/c^2r$ 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 5.3, 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 $\pm$ 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}.[3]

5.3 Black holes

5.3.1 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

[3]For a review article on this topic, see Clifford Will, "The Confrontation between General Relativity and Experiment," `http://relativity.livingreviews.org/Articles/lrr-2006-3/`.

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, $R^a{}_a$, 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. The Maxima command `lriemann(true)` displays the nonvanishing components of R_{abcd}. The component that misbehaves the most severely at $r = 0$ is $R_{trrt} = 2m/r^3$. Because of this, $R^{abcd}R_{abcd}$ 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 $t \to t' = t - 2m\ln(r - 2m)$ (see problem 3 on page 149). 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.)

5.3.2 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, $\mathrm{d}s^2$ 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 $\mathrm{d}s^2 < 0$, 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 5.1.

5.3.3 Expected formation

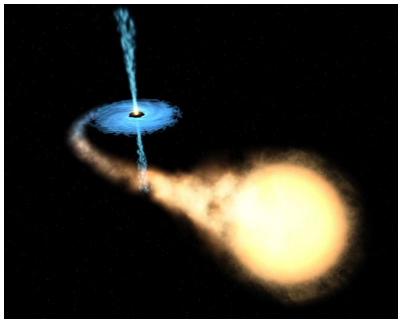
Einstein and Schwarzschild did not believe, however, that any of these features of the Schwarzschild metric were more than a mathematical curiosity, and the term "black hole" was not invented until the 1967, by John Wheeler. Although there is quite a bit of evidence these days that black holes do exist, there is also the related 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 $0.1M_\odot$ (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 $100M_\odot$ are immediately torn apart by their own solar winds. In the range from 0.1 to $100M_\odot$, stars form. As discussed in section 3.4.3, those with masses greater than about a few solar 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.

5.3.4 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 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 semi-major 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.



e / A black hole accretes matter from a companion star.

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 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.

5.3.5 Singularities and cosmic censorship

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 the singularity at the center of a black hole, 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 cannot extract any data about a singularity inside. In a way, this is a good thing. If a singularity exists, it is a point at which all the known laws of physics broke down, and physicists therefore have no way of predicting anything about its behavior. As John Earman of the University of Pittsburgh puts it, anything could pop out of a singularity, including green slime or your lost socks. In more technical language, a naked singularity would constitute an extreme violation of unitarity and an acute instance of the information paradox (see page 133).

As long as singularities are hidden behind event horizons, this has no effect on our ability to make predictions about the physical behavior of the universe. There is no obvious built-in reason that general relativity should not allow *naked* singularities, but neither do we know of any real-world process by which one could be formed. Roger Penrose's 1969 *cosmic censorship* hypothesis states that our universe contains no naked singularities other than the Big Bang. A reasonably readable treatment of these issues is given in Earman, *Bangs, Crunches, Whimpers, and Shrieks: Singularities and Acausalities in Relativistic Spacetimes*, Oxford, 1995; he clearly marks the sections that are highly technical and suggests how the non-specialist reader can navigate the book and absorb the main points.

5.3.6 Black hole radiation

Since event horizons are expected to emit blackbody radiation, a black hole should not be entirely black; it should radiate. 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 black-hole 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.

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 on page 38. 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.

Problems

Key

The notation √indicates that a computerized answer check is available online.

1 The metric of coordinates (θ, ϕ) on the unit sphere is $\mathrm{d}s^2 = \mathrm{d}\theta^2 + \sin^2\theta \mathrm{d}\phi^2$. (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.

2 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 $0.98M_\odot$ and a radius of 5.9×10^6 m. Find the redshift.

3 Show that, as claimed on page 149, 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.

Chapter 6
Sources

6.1 Sources in general relativity

6.1.1 Point sources in a background-independent theory

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 on page 152, 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 electromag-

netism, 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 $1/r^2$ 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 isn't even part of the continuous structure of spacetime (and its location is also hidden behind an event horizon, so it can't be observed from the outside).

6.1.2 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 energy-momentum tensor T^{ab} (page 101) on the other,

$$R_{ab} = 8\pi T_{ab} \quad ,$$

where a factor of 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, conservation of mass-energy requires the divergence-free property $\nabla_b T^{ab} = 0$, because otherwise, for example, we could have a mass-energy tensor that varied as $T^{00} = kt$, describing a region of space in which mass was uniformly appearing or disappearing at a constant rate. But this adds 4 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 the a different tensor G_{ab}, called the Einstein tensor, defined by $G_{ab} = R_{ab} - (1/2)Rg_{ab}$,

$$G_{ab} = 8\pi T_{ab} \quad .$$

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 energy-momentum tensor that satisfies the field equation is automatically divergenceless, and thus no additional constraints need to be applied in order to guarantee conservation of mass-energy.

Self-check: Does replacing R_{ab} with G_{ab} invalidate the Schwarzschild metric?

Further interpretation of the energy-momentum tensor

The energy-momentum tensor was briefly introduced in section 4.2 on page 101. 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. 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. The spacelike components are associated with the pressure, P. The form of the tensor with mixed upper and lower indices has the simple form $T^{\mu}{}_{\nu} = diag(-\rho, P, P, P)$.

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 $r^3/r^2 = r$. 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 rotations.

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:

- The equivalence principle is satisfied.
- It should be coordinate-independent.
- It should be equivalent to Newtonian gravity in the appropriate limit.
- 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.

A failed attempt at tinkering *Example: 1*

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 energy-momentum 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 c_1 and c_2 would be constrained by the requirement that the theory agree with Newtonian gravity in the classical limit.

This particular attempt fails, because it violates the equivalence principle. Consider a beam of light directed along the x axis. Its momentum is equal to its energy (see page 81), so its contributions to the local energy density and pressure are equal. Thus its contribution to the energy-momentum tensor is of the form $T^\mu{}_\nu$ = (constant) $\times$ $diag(-1, 1, 0, 0)$. The trace vanishes, so its coupling to gravity in the c_2 term is zero. But this violates

the equivalence principle, which requires that all forms of mass-energy contribute equally to gravitational mass.

One way in which we *can* change the field equation without violating any of these is to add a term Λg_{ab}, giving

$$G_{ab} = 8\pi T_{ab} + \Lambda g_{ab} \quad ,$$

which is what we will refer to as the Einstein field equation.[1] The universal constant Λ is called the cosmological constant. 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 pseudo-classical argument. Although we've already seen (page 143) 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.

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 the come out right. Therefore we can estimate that the cosmological constant could not have been greater than about $(10^{-8})m/r^3 \sim 10^{-10}$ 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

[1] In books that use a $-+++$ metric rather then our $+---$, the sign of the cosmological constant term is reversed relative to ours.

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.

6.2 Cosmological solutions

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 isotropic (the same in all spatial directions) and homogeneous (the same at all locations in space), 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

$$\mathrm{d}s^2 = \mathrm{d}t^2 - a(t)\mathrm{d}\ell^2 \quad ,$$

where the spatial part is

$$\mathrm{d}\ell^2 = f(r)\mathrm{d}r^2 + r^2\mathrm{d}\theta^2 + r^2\sin^2\theta\mathrm{d}\phi^2 \quad .$$

In these coordinates, the time t is interpreted as the proper time of a particle that has always been at rest. Events that are simultaneous according to this t are events at which the local properties of the universe — i.e., its curvature — are the same. 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 unknown function $f(r)$ has to make the 3-space metric $\mathrm{d}\ell^2$ have 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, $G^t{}_t = (1-1/f)/r^2$. 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 $f = 1/(1-kr^2)$, where $k = -1$, 0, or 1. The form of $\mathrm{d}\ell^2$ shows us that k can be interpreted in terms of the sign of the spatial curvature. The $k = 0$ case gives a flat space. For negative k, a circle of radius r centered on the origin has a circumference $2\pi r f(r)$ that is less than its Euclidean value of $2\pi r$. The opposite occurs for $k > 0$. The resulting metric, called the Robertson-Walker metric, is

$$\mathrm{d}s^2 = \mathrm{d}t^2 - a^2\left(\frac{\mathrm{d}r^2}{1-kr^2} + r^2\mathrm{d}\theta^2 + r^2\sin^2\theta\mathrm{d}\phi^2\right) \quad .$$

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

$$G^t{}_t = 3\left(\frac{\dot{a}}{a}\right)^2 + 3ka^{-2}$$
$$G^r{}_r = G^\theta{}_\theta = G^\phi{}_\phi = 2\frac{\ddot{a}}{a} + \left(\frac{\dot{a}}{a}\right)^2 + ka^{-2} \quad ,$$

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 $g = diag(1,-a^2,0,0)$. The energy-momentum tensor is $T^\mu{}_\nu = diag(-\rho,P,P,P)$. Substituting into $G^a{}_b = 8\pi T^a{}_b + \Lambda g^a{}_b$, we find

$$3\left(\frac{\dot{a}}{a}\right)^2 + 3ka^{-2} - \Lambda = 8\pi\rho$$
$$2\frac{\ddot{a}}{a} + \left(\frac{\dot{a}}{a}\right)^2 + ka^{-2} - \Lambda = -8\pi P \quad .$$

Rearranging a little, we have the Friedmann equations,

$$\begin{aligned}\frac{\ddot{a}}{a} &= \frac{1}{3}\Lambda - \frac{4\pi}{3}(\rho + 3P) \\ \left(\frac{\dot{a}}{a}\right)^2 &= \frac{1}{3}\Lambda + \frac{8\pi}{3}\rho - ka^{-2} \qquad .\end{aligned}$$

6.2.1 Evidence for expansion of the universe

By 1929, Edwin Hubble at Mount Wilson had determined that the universe was expanding rather than static, so that Einstein's original goal of allowing a static cosmology became pointless. The universe, it seemed, had originated in a Big Bang (a concept that originated with the Belgian Roman Catholic priest Georges Lemaître). This now appears natural, since the Friedmann equations would only allow a constant a in the case where Λ was perfectly tuned relative to the other parameters. 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.

Self-check: Why is it not correct to think of the Big Bang as an explosion that occurred at a specific point in space?

The existence of the Big Bang is confirmed directly by looking up in the sky and seeing it. 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.

This is now interpreted as follows. Soon after the Big Bang, 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 around 4×10^5 years after the big bang, 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.

6.2.2 Observability of expansion

The proper interpretation of the expansion of the universe, as described by the Friedmann equations, can be tricky. It might seem as though the expansion would be undetectable, in the sense that general relativity is coordinate-independent, and therefore 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!"

One way to see that the expansion does not apply on every scale would be to solve the Einstein field equations exactly so as to describe the internal structure of the bodies that occupy the space: galaxies, superclusters, etc. This is impractical in general, but can be done in simple cases, as in example 3 on page 167.

Another way to see this is that if meter-sticks expanded along with the universe, then the expansion would be nothing more than a change of coordinates. 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.

So in general a gravitationally bound system does not expand due to the stretching of the cosmological metric, nor does a system bound by electrical or nuclear forces. Note that this is different from the case of a photon traveling across the universe as described above. Such a photon *does* expand, and this, too, 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.

6.2.3 The vacuum-dominated solution

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 non-degenerate 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 (page 91), 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 radiation.

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

$$\begin{aligned} \frac{\ddot{a}}{a} &= \frac{1}{3}\Lambda \\ \left(\frac{\dot{a}}{a}\right)^2 &= \frac{1}{3}\Lambda \end{aligned} \quad .$$

This is referred to as a vacuum-dominated universe. The solution is

$$a = \exp\left[\sqrt{\frac{\Lambda}{3}}\, t\right] \quad .$$

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 on page 160, 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. 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 6.2.2 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

$$\mathrm{d}s^2 = \mathrm{d}t^2 - a^2\mathrm{d}r^2 \qquad ,$$

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 $a(t)$.[2] Let a test particle travel in the radial direction, starting at event $\mathrm{A} = (0, 0)$ and ending at $\mathrm{B} = (t', r')$. 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

[2] A computation of the Einstein tensor with $\mathrm{d}s^2 = \mathrm{d}t^2 - a^2(1 - kr^2)^{-1}\mathrm{d}r^2$ 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 $\Lambda > 0$ and $k = -1$, 0, and 1 turn out in fact to be equivalent except for a change of coordinates.

geodesics relative to a line on an r-t plot is most easily understood in the limit where t' is fairly long compared to the time-scale $T = \sqrt{3/\Lambda}$ 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 $\mathrm{d}r$ 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.

Geodesics in a vacuum-dominated universe *Example: 2*

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 $\mathrm{d}r/\mathrm{d}t = 0$ initially. First we recall the geodesic equation

$$\frac{\mathrm{d}^2x^i}{\mathrm{d}\lambda^2} = \Gamma^i_{jk}\frac{\mathrm{d}x^j}{\mathrm{d}\lambda}\frac{\mathrm{d}x^k}{\mathrm{d}\lambda} \quad .$$

from page 116. The nonvanishing Christoffel symbols for the 1+1-dimensional metric $\mathrm{d}s^2 = \mathrm{d}t^2 - a^2\mathrm{d}r^2$ are $\Gamma^r_{tr} = \dot{a}/a$ 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 $\int \mathrm{d}s$ is simply the same as its time coordinate t, and we can therefore use t as an affine coordinate. Letting $\lambda = t$, we have

$$\begin{aligned}
&\frac{\mathrm{d}^2t}{\mathrm{d}t^2} - \Gamma^t_{rr}\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2 = 0 \\
&0 - \Gamma^t_{rr}\dot{r}^2 = 0 \\
&\dot{r} = 0 \\
&r = \text{constant}
\end{aligned}$$

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.

Schwarzschild-de Sitter space *Example: 3*

The metric

$$ds^2 = \left(1 - \frac{2m}{r} - \frac{1}{3}\Lambda r^2\right) dt^2 - \frac{dr^2}{1 - \frac{2m}{r} - \frac{1}{3}\Lambda r^2} - r^2 d\theta^2 - r^2 \sin^2\theta d\phi^2$$

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. The characteristic scale of the black hole, e.g., the radius of its event horizon, is still set by m, so we can see that cosmological expansion does not affect the size of gravitationally bound systems on smaller scales.

6.2.4 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 equations of state for nonrelativistic matter (p. 83) are

$$P = 0$$
$$\rho \propto a^{-3}$$

so the Friedmann equations become

$$\begin{aligned} \frac{\ddot{a}}{a} &= -\frac{4\pi}{3}\rho \\ \left(\frac{\dot{a}}{a}\right)^2 &= \frac{8\pi}{3}\rho - ka^{-2} \end{aligned} \quad ,$$

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 semiclassically in terms of 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 two-dimensional 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 $\dot{a}/a$ 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 semiclassical analogy, it represents the case in which the universe is moving exactly at escape velocity; as t approaches infinity, we have $a \to \infty$, $\rho \to 0$, and $\dot{a} \to 0$. 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 $-4\pi\rho/3 = -ca^{-3}$. The Friedmann equations are

$$\begin{aligned}\ddot{a} &= -ca^{-2} \\ \dot{a} &= \sqrt{2c}a^{-1/2} \qquad .\end{aligned}$$

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 $\dot{a}/a$ 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 $\rho = \rho_{crit} = 2H^2/8\pi = 2H^2/8\pi G$. The observed value of the Hubble constant is about $1/(14 \times 10^9$ 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.

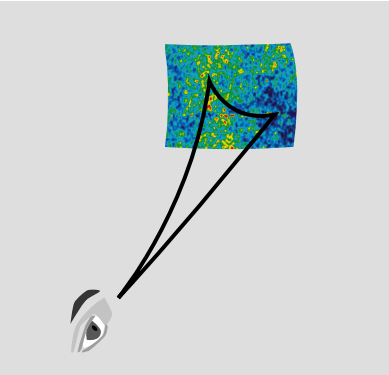
The open universe

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.

6.2.5 Observation

Current observational evidence indicates that we live in an open universe. Historically, it was very difficult to determine the universe's average density, even to within an order of magnitude. Most of the matter in the universe probably doesn't emit light, making it difficult to detect. Astronomical distance scales are also very poorly calibrated against absolute units such as the SI. The observation of the universe's accelerating expansion, however, marked the beginning of a new era of high-precision cosmology. If the universe's density had been significantly higher than the critical density, then it would have begun recontraction before it could enter the vacuum-dominated phase of accelerated expansion. We therefore have $\rho \lesssim \rho_{crit}$.

A further 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. The temperature of the cosmic microwave background radiation is not the same in all directions, and its can



a / The angular scale of fluctuations in the cosmic microwave background can be used to infer the curvature of the universe.

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 backwards and infer the universe's spatial curvature based on the angular scale of the anisotropies.

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.

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."

Chapter 7
Gravitational Waves

7.1 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 that it would be greater than c; in section 1.7 (p. 29), 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, Lorentzian space, it seems likely to apply to general relativity as well, at least for ripples in spacetime that are relatively weak, so that space is approximately Lorentzian. 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 that showed that gravitational effects should propagate at c. 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.

Although extremely well founded theoretically, this turns out to be extremely difficult to test empirically. In a 2003 experiment,[1] Fomalont and Kopeikin used a world-wide array of radio telescopes to observe a conjunction in which Jupiter passed within $3.7'$ of a quasar, so that the quasar's radio waves came within about 3 light-seconds of the planet on their way to the earth. Since Jupiter moves with $v = 4 \times 10^{-5}$, one expects naively that the radio waves passing by it should be deflected by the field produced by Jupiter at the position it *had* 3 seconds earlier. This position differs from its present position by about 10^{-4} light-seconds, and the result should be a difference in propagation time, which should be different when observed from different locations on earth. Fomalont and Kopeikin measured these phase differences with picosecond precision, and found them to be in good agreement with the predictions of general relativity. The real excitement started when they published their result with the interpretation that they had measured, for the first time, the speed

[1] http://arxiv.org/abs/astro-ph/0302294

of gravity, and found it to be within 20% error bars of c. Samuel[2] and Will[3] published refutations, arguing that Kopeikin's calculations contained mistakes, and that what had really been measured was the speed of light, not the speed of gravity. The reason that the interpretation of this type of experiment is likely to be controversial is that although we do have theories of gravity that are viable alternatives to general relativity (e.g., the Brans-Dicke theory, in which the gravitational constant is a dynamically changing variable), such theories have generally been carefully designed to agree with general relativity in the weak-field limit, and in particular every such theory (or at least ever theory that remains viable given current experimental data) predicts that gravitational effects propagate at c in the weak-field limit. Without an alternative theory to act as a framework — one that *disagrees* with relativity about the speed of gravity — it is difficult to know whether an observation that agrees with relativity is a test of this specific aspect of relativity.

7.2 Gravitational radiation

7.2.1 Empirical evidence

So we still don't know, a century after Einstein found the field equations, whether gravitational "ripples" travel at c. Nevertheless, we do have strong empirical evidence that such ripples exist. The Hulse-Taylor system (page 145) contains two neutron stars orbiting around their common center of mass, and the period of the orbit is observed to be lengthening gradually over time, a. This is interpreted as evidence that the stars are losing energy to radiation of gravitational waves.[4]

More dramatic, if less clearcut, evidence is provided by Komossa, Zhou, and Lu's observation`http://arxiv.org/abs/0804.4585` 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.

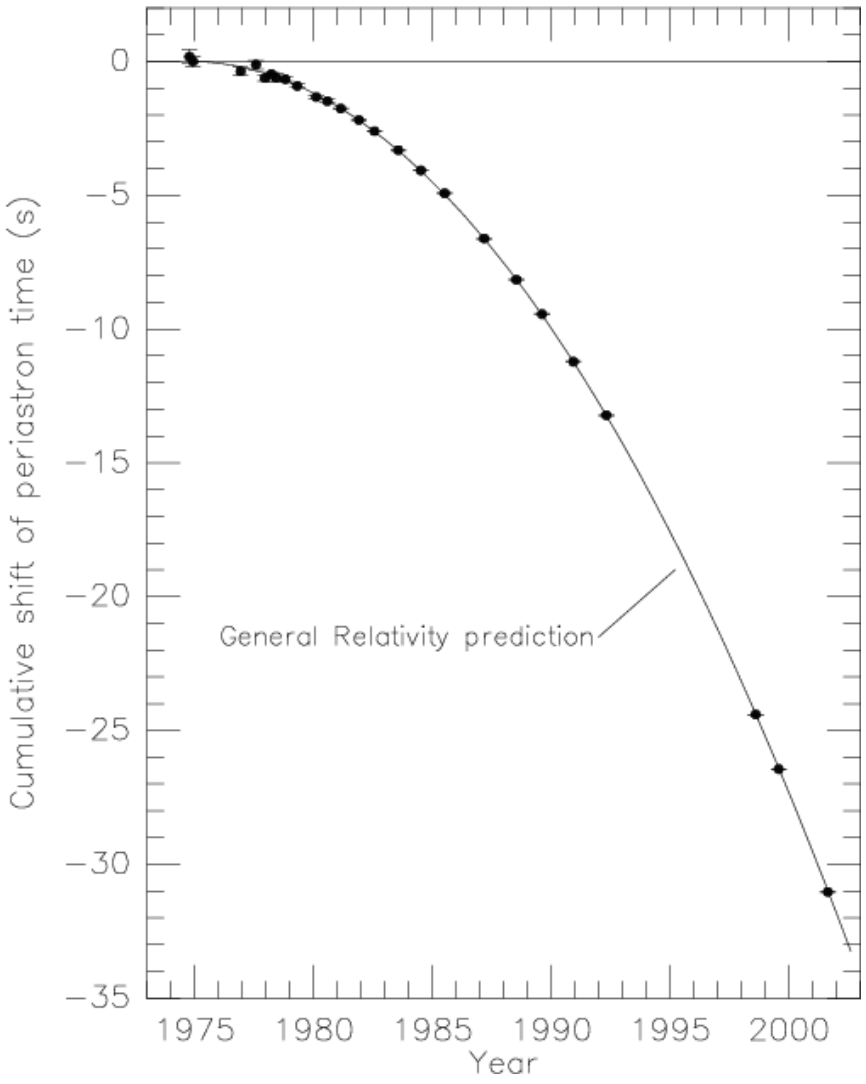
7.2.2 Expected properties

To see what properties we should expect for such radiation, first consider the reasoning that led to the construction of the Ricci and

[2] `http://arxiv.org/abs/astro-ph/0304006`

[3] `http://arxiv.org/abs/astro-ph/0301145`

[4] Stairs, "Testing General Relativity with Pulsar Timing," `http://relativity.livingreviews.org/Articles/lrr-2003-5/`



a / The Hulse-Taylor pulsar's orbital motion is gradually losing energy due to the emission of gravitational waves. The linear increase of the period is integrated on this plot, resulting in a parabola.

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., $-\mathrm{d}^2V/\mathrm{d}t^2$ 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 6.2.2 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 6.2.2; 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.

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.

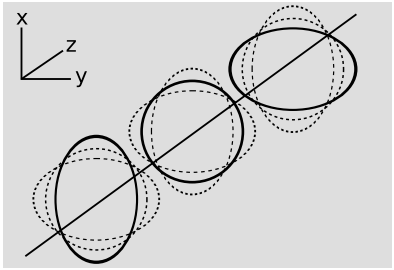
$$\mathrm{d}s^2 = \mathrm{d}t^2 - \left(1 + \frac{1}{10}\sin x\right)\mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2$$

The Christoffel symbols depend on derivatives of the form $\partial_a g_{bc}$, so here the only nonvanishing Christoffel symbol is ${\Gamma^x}_{xx}$. It is then straightforward to check that the Riemann tensor ${R^a}_{bcd} = \partial_c {\Gamma^a}_{db} - \partial_d {\Gamma^a}_{cb} + {\Gamma^a}_{ce}{\Gamma^e}_{db} - {\Gamma^a}_{de}{\Gamma^e}_{cb}$ vanishes by symmetry. Therefore this metric must really just be a flat-spacetime metric that has been subjected to a silly change of coordinates.

To keep the curvature from vanishing, we clearly need to have a metric in which the oscillation is not restricted to a single variable. For example, the metric

$$\mathrm{d}s^2 = \mathrm{d}t^2 - \left(1 + \frac{1}{10}\sin y\right)\mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2$$

does have nonvanishing curvature. In other words, it seems like we should be looking for transverse waves rather than longitudinal ones. 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.



b / As the gravitational wave propagates in the *z* direction, the metric oscillates in the *x* and *y* directions, preserving volume.

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:

$$\mathrm{d}s^2 = \mathrm{d}t^2 - (1 + A\sin(z-t))\,\mathrm{d}x^2 - \frac{\mathrm{d}y^2}{1 + A\sin(z-t)} - \mathrm{d}z^2$$

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

$$G_{tt} = -\frac{\cos^2(z-t)}{2 + 4A\sin(z-t) + 2A^2\sin^2(z-t)}A^2$$

For small values of A, we have $|G_{tt}| \lesssim A^2/2$. 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.[5]

7.2.3 Some exact solutions

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.

An exact solution *Example: 1*

We've already seen, e.g., in the derivation of the Schwarzschild metric in section 5.2.3, 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

$$ds^2 = dt^2 - (1+f)\,dx^2 - \frac{dy^2}{1+f} - dz^2 \quad ,$$

where $f = A\sin(z-t)$. 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

[5] Some of the history is related at `http://en.wikipedia.org/wiki/Sticky_bead_argument`.

haven't found quite the right nonlinear dependence. Suppose we try something of the form

$$\mathrm{d}s^2 = \mathrm{d}t^2 - \left(1 + f + cf^2\right)\mathrm{d}x^2 - \frac{\mathrm{d}y^2}{1 - f + df^2} - \mathrm{d}z^2 \qquad .$$

This approximately conserves volume, since $(1+f+\ldots)(1-f+\ldots)$ 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 $c = d = 3/8$. 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 $e^{+i\ldots}$ and $e^{-i\ldots}$, so as to give a real result. Unfortunately 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 1).

An exact, oscillatory, non-monochromatic solution *Example: 2*

Assume a metric of the form

$$\mathrm{d}s^2 = \mathrm{d}t^2 - p(z-t)^2\mathrm{d}x^2 - q(z-t)^2\mathrm{d}y^2 - \mathrm{d}z^2 \qquad ,$$

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.

```
load(ctensor);
ct_coords:[t,x,y,z];
depends(p,[z,t]);
depends(q,[z,t]);
lg:matrix([1,0,0,0],
          [0,-p^2,0,0],
          [0,0,-q^2,0],
          [0,0,0,-1]);
cmetric();
einstein(true);
```

The result is proportional to $\ddot{q}/q + \ddot{p}/p$, so any functions p and q that satisfy the differential equation $\ddot{q}/q + \ddot{p}/p = 0$ will result in a solution to the field equations. Setting $p(u) = 1 + A\cos u$, for example, we find that q is oscillatory, but with a period longer than 2π (problem 2).

An exact, plane, monochromatic wave *Example: 3*

Any metric of the form

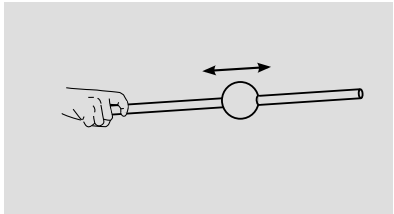
$$ds^2 = (1-h)dt^2 - dx^2 - dy^2 - (1+h)dz^2 + 2h\,dz\,dt \quad ,$$

where $h = f(z-t)xy$, and f is any function, is an exact solution of the field equations (problem 3).

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.

7.2.4 Energy content

To show that these waves carry momentum and energy, we can use the nonmathematical "sticky bead argument" (figure c), which was originated by Feynman in 1957 and later popularized by Bondi.



c / The 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.

We would next like to find an expression for the energy of the wave in terms of its amplitude. Alas, this was not meant to be.

It seems like it ought to be straightforward. We have such expressions in other classical field theories. In electromagnetism, we have energy densities $+(1/8\pi k)|\mathbf{E}|^2$ and $+(1/2\mu_o)|\mathbf{B}|^2$ associated with the electric and magnetic fields. In Newtonian gravity, we can assign an energy density $-(1/8\pi G)|\mathbf{g}|^2$ 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*, `http://arxiv.org/abs/gr-qc/?9712019`.

7.2.5 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 $\omega^2 x_o^2$.

In electromagnetism, conservation of charge forbids the existence of an oscillating electric monopole. The simplest radiating source is therefore an oscillating electric dipole $q = q_o \sin \omega t$. 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 3.2.5) is proportional to $q_o\omega$, the fields to $q_o\omega^2$, and the radiated power to $q_o^2\omega^4$. 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.

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, $Q = Q_o \sin \omega t$. 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 $Q_o^2 \omega^6$.

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 $(m/r)^5$. In units with $G = 1$ and $c = 1$, power is unitless, so the units of this expression check out. Reinserting the proper constants to give an equation that allows practical calculation in SI units, we have

$$P = k \frac{G^4}{c^5} \left(\frac{m}{r} \right)^5 \quad ,$$

where k is a unitless constant of order unity.

For the Hulse-Taylor pulsar,[6] we have $m \sim 3 \times 10^{30}$ kg (about one and a half solar masses) and $r \sim 10^9$ m. The binary pulsar is made to order our purposes, since 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 $\alpha = 2.418 \times 10^{-12}$ seconds per second. To compare this with our crude theoretical estimate, we take the Newtonian energy of the system 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.

[6] http://arxiv.org/abs/astro-ph/0407149

Problems

Key

The notation √indicates that a computerized answer check is available online.

1 Show that the metric $\mathrm{d}s^2 = \mathrm{d}t^2 - A\mathrm{d}x^2 - B\mathrm{d}y^2 - \mathrm{d}z^2$ 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.

2 Verify the claims made in example 2. Characterize the (somewhat complex) behavior of the function q obtained when $p(u) = 1 + A\cos u$.

3 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 $2h\mathrm{d}z\mathrm{d}t$ 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.

Photo Credits

Cover Galactic center: NASA, ESA, SSC, CXC, and STScI **9** *Atomic clock:* USNO official photograph, public domain. **10** *Gravity Probe A:* I believe this diagram to be public domain, due to its age and the improbability of its copyright having been renewed. **14** *Stephen Hawking:* unknown NASA photographer, 1999, public-domain product of NASA. **16** *Eotvos:* Unknown source. Since Eötvös died in 1919, the painting itself would be public domain if done from life. Under U.S. law, this makes photographic reproductions of the painting public domain. **18** *Earth:* NASA, Apollo 17. Public domain. **18** *Orion:* Wikipedia user Mouser, GFDL. **18** *M100:* European Southern Observatory, CC-BY-SA. **18** *Supercluster:* Wikipedia user Azcolvin429, CC-BY-SA. **24** *Pound and Rebka photo:* Harvard University. I presume this photo to be in the public domain, since it is unlikely to have had its copyright renewed. **24** *Lorentz:* Jan Veth (1864-1925), public domain. **37** *Galaxies:* Hubble Space Telescope. Hubble material is copyright-free and may be freely used as in the public domain without fee, on the condition that NASA and ESA is credited as the source of the material. The material was created for NASA by STScI under Contract NAS5-26555 and for ESA by the Hubble European Space Agency Information Centre. **40** *Gamma-Ray burst:* NASA/Swift/Mary Pat Hrybyk-Keith and John Jones. **59** *Levi-Civita:* Believed to be public domain. Source: `http://www-history.mcs.st-and.ac.uk/PictDisplay/Levi-Civita.html`. **62** *Einstein's ring:* I have lost the information about the source of this image. I would be grateful to anyone who could put me in touch with the copyright owners. **66** *SU Aurigae's field lines:* P. Petit, GFDL 1.2. **76** *Galaxies:* Hubble Space Telescope. Hubble material is copyright-free and may be freely used as in the public domain without fee, on the condition that NASA and ESA is credited as the source of the material. The material was created for NASA by STScI under Contract NAS5-26555 and for ESA by the Hubble European Space Agency Information Centre. **91** *Chandrasekhar:* University of Chicago. I believe the use of this photo in this book falls under the fair use exception to copyright in the U.S. **95** *Relativistic jet:* Biretta et al., NASA/ESA, public domain. **99** *Rocks:* Siim Sepp, CC-BY-SA 3.0. **100** *Jupiter and comet:* Hubble Space Telescope, NASA, public domain. **101** *Earth:* NASA, Apollo 17. Public domain. **101** *Moon:* Luc Viatour, CC-BY-SA 3.0. **102** *Heliotrope:* ca. 1878, public domain. **102** *Triangulation survey:* Otto Lueger, 1904, public domain. **106** *Triangle in a space with negative curvature:* Wikipedia user Kieff, public domain. **111** *Eclipse:* Eddington's original 1919 photo, public domain. **123** *Torsion pendulum:* University of Washington Eot-Wash group, `http://www.npl.washington.edu/eotwash/publications/pdf/lowfrontier2.pdf`. **131** *Coin:* Kurt Wirth, public-domain product of the Swiss govern-

ment. **134** *Bill Unruh:* Wikipedia user Childrenofthedragon, public domain. **151** *Accretion disk:* Public-domain product of NASA and ESA. **170** *Cosmic microwave background image:* NASA/WMAP Science Team, public domain. **173** *Graph of pulsar's period:* Weisberg and Taylor, `http://arxiv.org/abs/astro-ph/0211217`.

Index

Euclidean geometry (page 12):

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.[7]

Ordered geometry (page 13):

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 Causality: Any three distinct events A, B, and C lying on the same line can be sorted out in order (and by statement 3, this order is unique).

Affine geometry (page 26):
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 (page 183):

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.

[7] This is a form known as Playfair's axiom, rather than the version of the postulate originally given by Euclid.

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:* Observers in different inertial frames agree on the time-ordering of events.

L5 *No simultaneity:* The experimental evidence in section 1.2 shows that observers in different inertial frames do not agree on the simultaneity of events.

Statements of the equivalence principle:

Accelerations and gravitational fields are equivalent. There is no experiment that can distinguish one from the other (page 17).

It is always possible to define a *local* Lorentz frame in a particular neighborhood of spacetime (page 19).

There is no way to associate a preferred tensor field with spacetime (page 89).