Lecture 1

The Principle of Relativity

1.1 RELATIVITY AND SPACETIME

> *Space* and *time* are commonly regarded as the *forms* of existence in the real world, *matter* as its *substance*. A definite portion of matter occupies a definite part of space at a definite moment of time. It is in the composite idea of *motion* that these three fundamental conceptions enter into intimate relationship.
>
> —Hermann Weyl, *Space-Time-Matter* (1921)

> Every body continues in its state of rest, or of uniform motion in a right line, unless it is compelled to change that state by forces impressed upon it.
>
> —Isaac Newton, *Mathematical Principles of Natural Philosophy* (1687)

Weyl's words open the Introduction to his textbook [1], which was among the first to discuss Einstein's theory of general relativity in depth. They encapsulate the general attitude of physicists toward the fundamental concepts of space, time and matter that underlie their discipline. However, Weyl's formulation is not free of controversy. The question of whether there is any such thing as a "definite part of space" that is *not* occupied by any "definite portion of matter," for example, has sparked some intense metaphysical debates over the years [2]. It is not our goal to explore this issue here. As Weyl does, we will adopt the standard point of view from Newtonian mechanics, that regions of space do exist, whether occupied by matter or not, and that bodies move by vacating one region for another. The reason we quote Weyl is that he does a particularly good job emphasizing a point that sometimes goes overlooked: the ideas of time and space, and of things existing in space, are bound together in the idea of motion.

The Copernican revolution in cosmology shifted the hypothetical center of the universe from the Earth to the Sun. Carrying this idea to its logical extreme, physicists came to believe long ago that there actually is no preferred point of space that one could reasonably call the center of the universe. Any two such points are basically the same. They may be occupied at a certain moment by different bodies, and thus may be distinguished by their contents, but their intrinsic properties, as points of space, do not differentiate them. Likewise, the fundamental laws of physics are held to be the same in all circumstances and, in particular, at all times. Accordingly, physicists generally believe that there is no fundamentally privileged

moment of time. Again, different things may actually be happening at different moments, but they are not *intrinsically* distinguishable. In stark contrast, however, there *are* privileged states of motion in Newtonian mechanics. These are the ***inertial motions*** described in Newton's first law of mechanics, also quoted above.

The concept of an inertial state of motion arose from a long series of abstractions and idealizations, which led eventually to the closely related idea of a ***test particle***. A test particle is supposed to be infinitely small, non-spinning, electrically neutral, and to have infinitesimally small mass. The first three properties decouple a test particle from tidal strains, gravitational torques and electromagnetic forces. The last one ensures that any gravitational radiation generated by the particle as it moves may safely be neglected. We will return later to describe these gravitational effects in more detail. For now, we note only that their absence for test particles means that such particles follow the most natural possible trajectory through space in Newtonian physics; they experience no forces. Newton asserts that any deviation of the motion of a real body from the test particle's idealized standard of natural motion must be explained by carefully describing all of the forces acting on the real body that do not affect the test particle.

Obviously, Newton's third law will have to be updated in relativistic physics. It turns out, however, that we need not abandon it all together. Rather, we must decouple its assertion that inertial states of motion *exist* from its assertion of what those motions actually *are*. That is, in relativistic physics, we can still imagine a test particle starting at any position in space with any initial velocity, and its subsequent dynamics will trace out an inertial motion. However, we cannot assert that one inertial motion differs from another by relative motion along a straight line at uniform velocity. This leads to

The Principle of Relativity Through any point of space, at any moment of time, there is exactly one inertial motion for each initial velocity a test particle might have at that point. The fundamental laws of physics do not distinguish these inertial motions from one another; they are fundamentally interchangeable.

Again, different inertial motions may be distinguished by reference to other phenomena — e.g., a particle may be (momentarily) at rest *relative to the sun* — but are intrinsically indistinguishable. The principle of relativity survives in this form in all known (classical) theories of physics. However, it leaves wide open the question of which trajectories through space are actually inertial. One of the great insights Einstein had in formulating general relativity is that this question must be answered *experimentally*, and that these experiments for a given initial position and velocity really probe the gravitational field in a surrounding region of space and time. We return to this point below.

The inertial trajectories of Newtonian mechanics describe preferred curves in space, the straight lines, but they are not identical with them. Indeed, one can accelerate back and forth along a straight line in a non-inertial motion that over time

passes through exactly the same set of spatial points as an inertial motion. What matters is *how* one moves along the preferred curve in time. This leads directly to the idea of spacetime, the four-dimensional fusion of the spatial coordinates (x, y, z) with the temporal coordinate t. Although Newtonian inertial motions are *not* merely curves in space, they are *exactly* curves in spacetime:

$$t \mapsto (t, x, y, z) = (t, v^x\, t, v^y\, t, v^z\, t). \tag{1.1}$$

Thus, the inertial motions underlying Newtonian physics are mathematically described as preferred structures not of space or time separately, but of spacetime. This recognition is fundamentally responsible for the vital role that spacetime plays in relativistic physics.

The above is rather an anachronistic point of view. Physicists before the advent of special relativity didn't spend much time thinking about spacetime as a unified whole. This is because, although inertial motions are certainly properties of that unified entity, Newtonian time is universal. Let us expand on this point in modern language. An ***inertial observer*** is an observer following an inertial trajectory, an inertial ***world-line***, through spacetime. A point of spacetime is called an ***event***. Using a set of clocks stationed at all points of space, and synchronized at some initial moment of time, any inertial observer can construct a ***spatial slice*** of all events that have the same time t as a given point on his or her own world-line. In Newtonian physics, all inertial observers agree as to what those spatial ***surfaces of simultaneity*** are. Thus, spacetime in Newtonian physics is naturally ***foliated*** by preferred spatial slices, one for each time t, which are picked out by Newton's universal time variable. There is consequently no pressing need to think in terms of spacetime when studying dynamics in Newtonian physics because space is a construct common to all inertial observers.

1.2 SPECIAL RELATIVITY

> It is known that Maxwell's electrodynamics — as usually understood at the present time — when applied to moving bodies, leads to asymmetries which do not appear to be inherent in the phenomena.
>
> —Albert Einstein, *On the electrodynamics of moving bodies* (1905)

The modern theory of relativity follows almost entirely from the one very simple observation due to Einstein. The principle of relativity described above asserts that no experiment can distinguish one inertial motion from another in Newtonian mechanics. Einstein suggested that no optical experiment could distinguish those motions either.

Maxwell's equations show that light is an electromagnetic wave phenomenon, and they make an unambiguous prediction that it propagates in vacuum at a fixed speed c. However, the equations are famously *not* preserved under the ordinary

Galilean transformation

$$t' = t \quad \text{and} \quad \vec{r}\,' = \vec{r} - \vec{v}\,t. \tag{1.2}$$ {GalBst}

of spacetime coordinate systems associated with different inertial observers. Thus, if one inertial observer finds an electromagnetic field satisfying Maxwell's equations, another in relative motion will not. This fact led to a debate over how to reconcile electromagnetism with the underpinnings of mechanics. Its broad outlines are probably familiar to most students, so we will survey the issues only briefly here.

The main historical response to the Maxwell equations' lack of Galilean covariance was to assert that the principle of relativity articulated in the previous section is, in fact, *false*. This point of view is roughly equivalent to the ***ether theory***, which holds that some fluid or mechanical substance, the ***ether***, permeates the universe, and that electromagnetic fields correspond to local disturbances in that medium. Undisturbed, the ether was supposed to move inertially, and that particular inertial motion was taken to define a universally preferred state of rest. Such theories were brought into doubt somewhat when detailed experiments, particularly those of Michelson and Morley, were unable to detect any preferred inertial motion using light interferometers. This result led theorists, principally Fitzgerald and Lorentz [3], to wonder whether motion through the ether could affect the geometry of material bodies. In particular, they imagined that a foreshortening of one of the interferometer arms used by Michelson and Morley could explain their negative result. This possibility is philosophically troubling since it seems to assert that, while a preferred state of inertial motion does exist, Nature conspires to hide it from experiment. But it proved fruitful nonetheless as it led Lorentz [4] to discover what we now call the Lorentz transformations

$$t' = \frac{t' - \vec{v}\cdot\vec{r}/c^2}{\sqrt{1 - v^2/c^2}} \quad \text{and} \quad \vec{r}\,' = \vec{r} - \hat{v}\,\hat{v}\cdot\vec{r} + \hat{v}\,\frac{\hat{v}\cdot\vec{r} - v\,t}{\sqrt{1 - v^2/c^2}} \tag{1.3}$$ {LorBst}

some time before Einstein formulated his special theory of relativity. Here, $\hat{v}$ denotes the unit vector along the direction of relative motion between two inertial observers. To first order in the relative speed v, of course, this transformation is identical to (1.2). For relative speeds v close to c, however, the two transformations are quite distinct. Lorentz's have the property that if a given electromagnetic field obeys the Maxwell equations for one inertial observer in his coordinate system, and others' coordinate systems on spacetime are related to his by (1.3), then all will find that the field obeys Maxwell's equations. Thus, the principle of relativity is restored.

Today, with benefit of hindsight, Lorentz's "materialist" approach to the problem underlying special relativity may initially seem like one of the many false starts in the history of theoretical physics. It is not. Bell takes up this approach in a stimulating paper [5] using more modern language then Lorentz himself. Starting from the electric and magnetic fields generated by charges in uniform motion, Bell shows that (classical) atoms, and therefore material bodies built from them, naturally contract by the Lorentz factor. He also shows that the orbits slow in such atoms,

thereby slowing atomic clocks, again by just the right factor. The result is that a moving inertial observer using instruments composed of (classical) atoms would construct spacetime coordinates differing from those of a resting observer by *exactly* the Lorentz transformations (1.3). This argument therefore leads to the same result as Einstein's. But the two do not contradict one another, both trying to explain the same phenomenon from different roots. Quite the opposite, they reinforce one another. Bell's argument can be read to show that there was never any need for an ether in the first place: If we insist that our observers use *real* clocks and *real* rulers to locate events, then (1.2) is replaced with (1.3) and, by Lorentz's work, Maxwell's equations take the same form for all inertial observers. There is no need for a universally preferred rest frame.

Einstein intuited that the key result of the "materialist" scheme is that, as a result of the subtleties of real clocks and rulers, all inertial observers measure a light ray to move at speed c. He then proceeded simply to postulate that as a foundation of his theory:

The Principle of a Constant Speed of Light Light propagates through empty space at speed c, regardless of the motion of its source.

Different inertial observers will of course see the source of a given light ray moving with different velocities. Einstein's principle states that, nonetheless, they all agree as to the speed of the ray itself. This seems at first to be fundamentally at odds with Newtonian physics but, just as with the apparent non-covariance of the Maxwell equations, this impression results from from an inappropriate application of (1.2) rather than (1.3).

We now have stated both of the principles that Einstein postulates in his celebrated first paper [6] on special relativity. The first, the principle of relativity, is identical to that in Newtonian physics, although we have relaxed it here with an eye toward the general theory. The second, the principle of a constant speed of light, follows immediately from the first when we insist that Maxwell's equations be covariant among all inertial observers. These postulates therefore hardly seem radical at all; they merely reassert one of the oldest and deepest principles in all of physics, relativity. It is the implications of this familiar principle extended to electromagnetism that radically contradict Newton's time-honored assumptions about space and time.

While Einstein's postulates above are the only ones he explicitly states, there are actually several others. These tacit assumptions are what makes the special theory of relativity "special." This being a course of general relativity, naturally we should make those assumptions explicit here:

Assumption of Euclidean Space Every inertial observer finds space at any moment of time to be a three-dimensional Euclidean continuum.

Assumption of Rectilinear Light Motion Every inertial observer sees a light

ray propagate in vacuum *along a straight line* at the fixed speed c demanded above.

Assumption of Rectilinear Relative Motion Any two inertial observers move relative to one another at uniform speed in a straight line through Euclidean space.

Assumption of Homogeneous Time Every inertial observer sees every other inertial observer's clock run at a uniform rate.

Assumption of Homogeneous Space The relationship between the spatial coordinates associated to a fixed spacetime event by two inertial observers is linear, with perhaps an additional time-dependence.

Let us comment briefly on each of these. Einstein's first assumption was completely natural in his day. The idea that space itself could actually possess any of the recently-discovered non-Euclidean geometries was not deemed practically important by most physicists. Einstein's second and third assumptions basically restore Newton's assumption that inertial motions take place along straight lines relative to one another. It was not until he formulated the principle of equivalence that he considered relaxing this. Einstein's fourth and fifth assumptions are natural when one takes spacetime to be a vector space, as he does. Note that the possible time-dependence in the transformation of spatial coordinates arises already in the Galilean transformation (1.2).

1.3 THE RELATIVITY OF SIMULTANEITY

> Henceforth space by itself, and time by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality.
>
> —Hermann Minkowski, *Space and time* (1908)

We have seen above the the principle of relativity, which states that there exists a preferred class of indistinguishable inertial motions, identifies a physically interesting structure not on space or time, but on spacetime. We have also seen that extending this principle to electromagnetism demands that all observers will agree whether or not a given curve in space time moves at the speed of light. This contradicts the familiar velocity addition law derived from the Galilean transformations (1.2). Einstein resolved this tension by stepping back and examining how inertial observers actually induce coordinates on spacetime *experimentally* by making measurements. In this sense, he was attacking the problem in much the same way as Fitzgerald and Lorentz, but of course his aproach was much more direct and simple.

Einstein's original scheme [6] to construct inertial coordinate systems on a spacetime obeying his constant-c postulate used systems of synchronized clocks and standard rulers. We will take a slightly different approach here, one based on ***radio-***

location. The idea is that an inertial observer O carrying a single clock and can locate events elsewhere in spacetime by bouncing light rays off them and measuring both the time t_S when the signal is sent and the time t_R when it returns. The procedure is sketched in Figure 1.1(a). Because the light ray spends half its time of travel getting to E, and the other half returning, naturally t_E is the average of t_S and t_R. Likewise, the distance to E is half the total traveled by the light ray. If O is also equipped to measure the spatial direction $\hat{r}$ along which the initial light ray was sent, then we can define

$$t_E = \frac{t_R + t_S}{2} \quad \text{and} \quad \vec{r}_E = c\,\frac{t_R - t_S}{2}\,\hat{r}. \tag{1.4}$$ {Einco}

This expression gives the time and place of any event E in terms of local measurements on the world-line of the observer O. Our assumptions above that space is homogeneous and Euclidean guarantee that every E corresponds to a unique set of coordinate values $(t_E, \vec{r}_E)$, as measured by O. That is, for each observer O, (1.4) defines a global ***inertial coordinate system*** on spacetime.

Different inertial observers will of course define different coordinates on spacetime, and it is natural to ask how these relate to one another. We answer this question in full in the next Lecture, where we derive the Lorentz transformations (1.3) in detail. Here, we focus on a more specific issue, which goes to the heart of Minkowski's quote above. In special relativity, different inertial observers generally do not agree as to whether or not two events occur simultaneously. Thus, Einstein's theory does not enjoy the benefit of Newton's universal time, flowing "equably without relation to anything external" [2] underneath the tangible universe. Rather, in modern terms, different inertial observers define completely different surfaces of simultaneity in spacetime. That is, each has his or her own notion of space. The idea of spacetime is therefore indispensable in special relativity. One cannot simply revert to describing events happening in a fixed space over the course of time, as one does in Newtonian physics, because physically-equivalent inertial observers will not share a common description of how those events unfold over time.

Figure 1.1(b) shows a two-dimensional example of different inertial observers, O and O', assigning coordinates to events A and B in spacetime. Because each sees light propagate with the same speed, they *will agree* on the trajectories of light rays through spacetime. That is, they will agree as to whether a given curve in spacetime defines a series of points in their respective spaces that moves always at speed c. So, for example, the ray that O' sends out to A from S' merely continues the one sent by O to A from S_A. This diagram is made from the point of view of O. His time runs up the vertical axis, and his spatial slices are horizontal. However, since the clock that O' carries runs at a uniform rate relative to his, the midpoint between S' and R' according to her clock will be the same point, N', that O will find with his clock. Thus, O'will decide that A and B are both simultaneous with N'. Meanwhile, O will judge A to be simultaneous with N_A and B with N_B. Thus, A and B are simultaneous for O', but not for O. In fact, we can trace out the surface

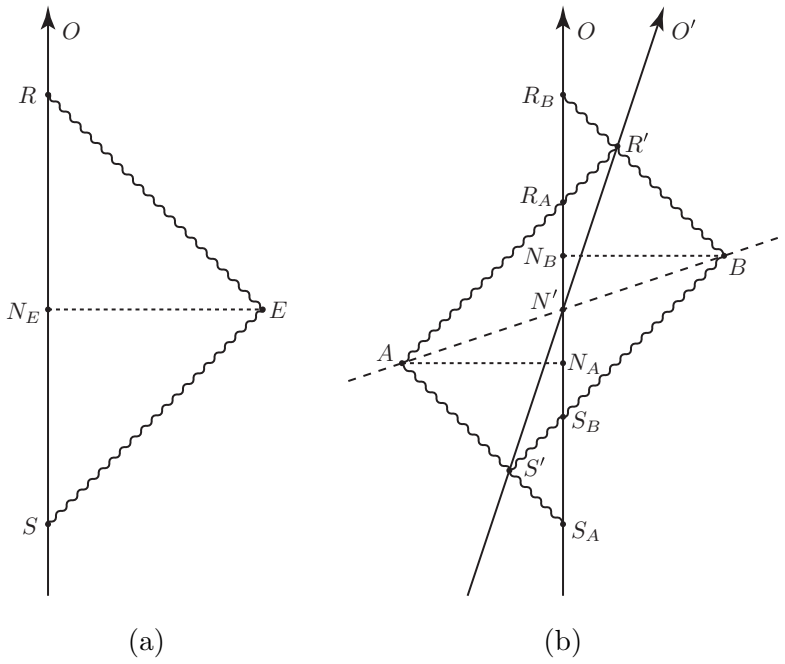

Figure 1.1: (a) An inertial observer O radio-locates an event E in spacetime. He sends a light signal from the event S on his own world-line that returns from E at R. He judges E to have occurred simultaneously with the event N_E on his world-line midway between those two. He also judges the distance to E to be half the total time of travel for the light rays. (b) A second inertial observer O' radio-locates a pair of events events A and B. Since she sends signals to both from the same event S', and receives responses at the same event R', she will judge A and B both to be simultaneous with the event N' on her own world-line. The original observer O, however, judges A to be simultaneous with N_A and B to be simultaneous with N_B. The dashed line indicates the surface of simultaneity in spacetime for the observer O' at time $t'_{N'}$. {simultaneity}

of all events in spacetime that O' will judge to be simultaneous with N'. In this two-dimensional case, this is the dashed line shown in the figure. It corresponds exactly to a line of constant t' in (1.3), and is certainly not one of the horizontal lines of constant t in this diagram. In this sense, O and O' have completely different definitions of "space" within spacetime. They are different surfaces, and space in special relativity is not universally agreed by all inertial observers. It is relative.

Exercise 1.5: Show that two inertial observers at rest relative to one another define the same surfaces of simultaneity in spacetime.

1.4 THE MINKOWSKI INTERVAL

The previous section has helped emphasize the importance of structures in spacetime on which all inertial observers agree. In that case, it was the curves through spacetime traced out by light rays. Here, we focus on a more general construct, the ***Minkowski interval***

$$\|E\|^2 := -c^2\,t^2 + x^2 + y^2 + z^2. \tag{1.6}$$ {Mint}

Here, E is an event in spacetime, and (t, x, y, z) are the inertial coordinates assigned to that event by an inertial observer O. Most students in a general relativity course have probably encountered this object at some point in their careers, and know that it is important because all inertial observers will find the same value of $\|E\|^2$ for a fixed event E. That is, if a second inertial observer O' assigns inertial coordinates (t', x', y', z') to E in then, even though the values of those individual coordinates will differ from the (t, x, y, z) measured by O,

$$-c^2\,t^2 + x^2 + y^2 + z^2 = \|E\|^2 = -c^2\,t'^2 + x'^2 + y'^2 + z'^2 \tag{1.7}$$ {intInv}

for all E. Thus, the interval is a spacetime structure common to all inertial observers and, because such observers are physically preferred, shows us an important, natural structure on Minkowski spacetime.

In this section, we *derive* the familiar Minkowski metric from the postulates and assumptions of special relativity laid out in the previous section. That is, we will argue that the above combination of inertial coordinates arises *automatically* when we *assume* that different inertial observers move relative to one another at uniform velocity along straight lines through a homogeneous, Euclidean space. This has profound implications for incorporating gravity into relativistic physics. These will be discussed below.

Figure 1.2 shows two inertial observers, O, whose inertial coordinate axes are shown, and O'. Without loss of generality, we choose the x-axis to lie along the ordinary three-velocity $\vec{v}$ of O' relative to O. Both observers will radio-locate an event E as described in the previous section. Again without loss of generality, we choose the y-axis such that E lies in the xy-plane. Thus, the Figure is generic.

The events S and R where O sends a signal and receives one back, respectively, are marked, as are the analogous events S' and R' for O'. These four events are defined by the intersection of the past and future light cones, which are shown, with the world-lines of the two observers. The key question concerns the relationship between the times t_S and t_R measured by O using his clock for his radio-location events, and the analogous times $t'_{S'}$ and $t'_{R'}$ measured by O' using her clock for hers. These relationships will induce a further one between the inertial coordinates assigned to this event by the two observers. Note that the sending and receiving times involve not only different clocks, but also completely different pairs of events in spacetime.

We proceed by finding the times $t_{S'}$ and $t_{R'}$ measured by O for the other observer's send and receive events. To do this, we work in his reference frame, and find the intersections of the world-line of O' with the light cone from E. Mathematically, we must solve

$$|\vec{r}_E - \vec{v}\,t|^2 = c^2\,(t_E - t)^2 \tag{1.8}$$

for the times t in question. Here, we have used the Euclidean norm that O observes on his spatial slice, which we assumed in the previous section. Expanding this equation gives a quadratic formula

$$t^2 - 2\,\frac{c^2\,t_E - \vec{v}\cdot\vec{r}_E}{c^2 - v^2}\,t + \frac{c^2\,t_E^2 - |\vec{r}_E|^2}{c^2 - v^2} = 0 = (t - t_{S'})\,(t - t_{R'}), \tag{1.9}$$ {tSR'q}

whose roots are precisely the two times $t_{S'}$ and $t_{R'}$, as measured by O, that we seek. Note that the lowest-order term in the quadratic is just the product of these two roots.

We now invoke the homogeneity of time assumed above. If our two observers both set their clocks to zero at the origin where their world-lines intersect, then

$$t_{S'} = \gamma\,t'_{S'} \quad \text{and} \quad t_{R'} = \gamma\,t'_{R'}. \tag{1.10}$$ {gamSR'}

This is because each inertial observer must measure the other's clock to run at a uniform rate. This particular choice of how to define the coefficient γ describing that relative rate is conventional. With this definition, we find

$$\gamma^2\,t'_{S'}\,t'_{R'} = t_{S'}\,t_{R'} = \frac{t_E^2 - |\vec{r}_E|^2/c^2}{1 - v^2/c^2} = \frac{t_S\,t_R}{1 - v^2/c^2}. \tag{1.11}$$ {intSR'}

The final equality here relates the times t_S and t_R measured by O using his clock for his send and receive events, to the times $t'_{S'}$ and $t'_{R'}$ measured by O' using her clock for her send and receive events, as desired. (This final step follows, for example, if we set $v = 0$ in the previous equality.) We note in passing that the coefficient of t in (1.9) is the sum of the roots of the polynomial, whence

$$\gamma\,t'_E := \gamma\,\frac{t'_{S'} + t'_{R'}}{2} = \frac{t_{S'} + t_{R'}}{2} = \frac{t_E - \vec{v}\cdot\vec{r}_E/c^2}{1 - v^2/c^2}. \tag{1.12}$$ {tR+tS'}

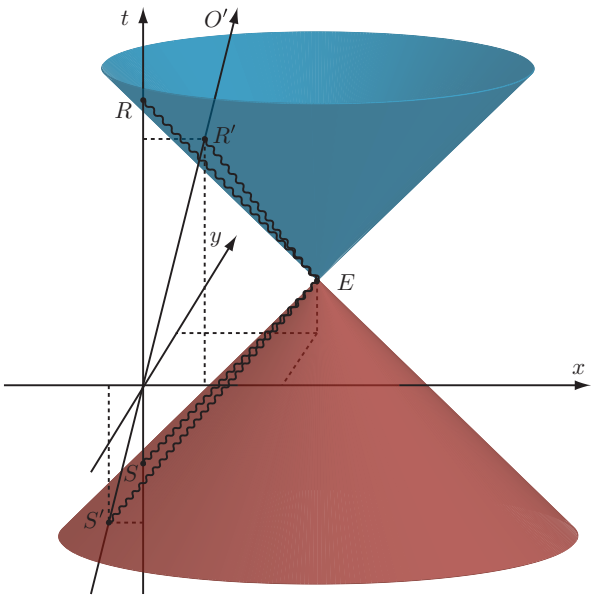

Figure 1.2: The inertial observer O' moves with uniform velocity relative to the inertial observer O, whose inertial coordinate system is shown. The unique events S' and R' that O' uses to radio-locate an event E are given by the intersection of that observer's world-line with the past and future light cones from that event, respectively. The analogous events S and R that O uses to radio-locate E are also shown. {interval}

We will return to this result below, when we work out the Lorentz transformations relating the inertial coordinate systems defined by our two observers.

Now suppose we ran through the same calculation again, but with the roles of O and O' reversed. Since O moves with three-velocity $-\vec{v}$ relative to O, and in particular with the same speed, we would find

$$(\gamma')^2\, t_S\, t_R = \frac{t'_{S'}\, t'_{R'}}{1 - v^2/c^2}, \tag{1.13}$$ {intSR}

where γ' is a constant analogous to γ in (1.10). Now we invoke the isotropy of space assumed in the previous section. There is no preferred direction in space, and therefore the factor γ in (1.10) can depend only on the relative *speed* of our two observers, and not on the *direction* of relative motion. It follows immediately that

$$\gamma' = \gamma. \tag{1.14}$$ {gamgam}

Inserting this result into (1.13), and then into (1.11), we find

$$\left(1 - v^2/c^2\right)^2 \gamma^4\, t_S\, t_R = t_S\, t_R \tag{1.15}$$

for all events E. We can meet this condition for all E provided that

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}}. \tag{1.16}$$ {gamdef}

Using this value in (1.13) gives

$$t'_{S'}\, t'_{R'} = t_S\, t_R \tag{1.17}$$ {invSR}

This is a remarkable result. Certainly each different inertial observer will use a different pair of events, S and R, to send and receive signals and radio-locate a given event E in spacetime. In addition, their different clocks will register different times for these each other's sending and receiving events (so, for instance, $t_R \neq t'_R$). However, our result says that *the product of each inertial observer's sending and receiving times as measured on his or her own clock is always exactly the same.* This product can vary if we change the event E, but not if we change the inertial observer O.

As we have discussed above, inertial observers are physically privileged. Therefore, any quantity on which they all agree is physically interesting. Thus, we define

$$\|E\|^2 := -c^2\, t_S\, t_R = -c^2\, (t_E)^2 + |\vec{r}_E|^2, \tag{1.18}$$ {intDef}

where we have again used the last equality in (1.11) to get the second expression here. This is an invariant quantity associated with each event E in spacetime, the Minkowski interval from (1.6). It measures a sort of distance from the origin to an event E in spacetime. It is often postulated as a foundation of special relativity,

but we have not taken that approach here. Rather, we have tried to show that (1.6) arises as a direct result of our assumptions that inertial observers (a) see space as a Euclidean continuum, (b) measure each other's clocks to run at uniform rates relative to their own, and (c) move at uniform velocity along straight lines relative to one another. If any of these assumptions were to break down, then clearly the spacetime interval would have to differ from Minkowski's. This is exactly what happens in the presence of a non-trivial gravitational field, as we will show below. This is basically why gravity becomes a field theory of the spacetime metric in relativistic physics.

1.5 MINKOWSKI SPACETIME

Let us return now to the question of how different inertial coordinate systems on spacetime relate to one another in special relativity. We have already found in (1.12) the relationship between the time coordinates assigned to an event E by different inertial observers:

$$t'_E = \gamma\left(t_E - \vec{v}\cdot\vec{r}_E/c^2\right). \tag{1.19}$$ {tLor}

We have used (1.16) here to simplify the expression slightly. This gives the first half of (1.3). To get the second half, we must calculate

$$\|E\|^2 = \left|\vec{r}'_E\right|^2 - \left(c\,t'_E\right)^2 \quad\Rightarrow\quad \left|\vec{r}'_E\right|^2 = \left|\vec{r}_E\right|^2 - \left(c\,t_E\right)^2 + \gamma^2\left(c\,t_E - \vec{v}\cdot\vec{r}_E/c\right)^2 \tag{1.20}$$ {sLor2'}

where we have used (1.19) and the definition (1.18) of the interval. We simplify this expression by recalling the definition (1.16) of γ and rearranging terms to find

$$\begin{aligned}\left|\vec{r}'_E\right|^2 &= \left|\vec{r}_E - \hat{v}\,\hat{v}\cdot\vec{r}_E\right|^2 + (\hat{v}\cdot\vec{r}_E)^2 + \gamma^2\left[v^2\,t_E^2 - 2\,v\,(\hat{v}\cdot\vec{r}_E)\,t_E + v^2\,(\hat{v}\cdot\vec{r}_E)^2/c^2\right]\\ &= \left|\vec{r}_E - \hat{v}\,\hat{v}\cdot\vec{r}_E\right|^2 + \gamma^2\left(\hat{v}\cdot\vec{r}_E - v\,t_E\right)^2.\end{aligned} \tag{1.21}$$ {sLor2}

This expression has the following geometric interpretation. The first term on the right is the norm of the component of $\vec{r}_E$ orthogonal to the direction of relative motion, encoded by the unit vector $\hat{v}$. The second term mixes the component $\hat{v}\cdot\vec{r}_E$ of the spatial coordinates *along* the direction with the time variable t_E, much as in the non-relativistic expression (1.2). If we note that our observers will agree on their direction of relative motion through spacetime, and invoke the homogeneity of space assumed in the previous Lecture, then we can recover the essentially linear transformation of spatial coordinates in (1.3).

It is conventional in special relativity to collect the temporal and spatial coordinates of an event E into a four-dimensional column matrix called a ***world-vector***. Different inertial observers will of course use different world-vectors to describe the same event E. These will be related by

$$\begin{pmatrix} t' \\ \vec{r}' \end{pmatrix} = \begin{pmatrix} \gamma & -\gamma c^{-2}\,\vec{v}\,\cdot \\ -\gamma\,\vec{v} & I\,\cdot\, + (\gamma-1)\,\hat{v}\hat{v}\cdot \end{pmatrix}\begin{pmatrix} t \\ \vec{r} \end{pmatrix}. \tag{1.22}$$ {LorMat}

Here, we have dropped the subscript E that we have previously used to label the coordinates and used the ordinary algebra of 2×2 matrices to collect the transformation of (1.3) into a single equation. One can, if one likes, think of the spatial part of a world vector as a set of three rows containing the Cartesian components x, y and z of the spatial vector $\vec{r}$. In this language, the two terms in the lower right corner of the transformation matrix are

$$I := \begin{pmatrix} 1 & & \\ & 1 & \\ & & 1 \end{pmatrix} \quad \text{and} \quad \hat{v}\hat{v} := \begin{pmatrix} \hat{v}^x\,\hat{v}^x & \hat{v}^x\,\hat{v}^y & \hat{v}^x\,\hat{v}^z \\ \hat{v}^y\,\hat{v}^x & \hat{v}^y\,\hat{v}^y & \hat{v}^y\,\hat{v}^z \\ \hat{v}^z\,\hat{v}^x & \hat{v}^z\,\hat{v}^y & \hat{v}^z\,\hat{v}^z \end{pmatrix}. \tag{1.23}$$ {IvvDef}

The first, of course, is the ***identity matrix***, while the second is called the ***dyad product*** of the unit vector $\hat{v}$ with itself. In the latter case, note that the action of the dyad product on a vector is exactly what the notation would lead you to expect:

$$\hat{v}\hat{v} \cdot \vec{r} := \hat{v}\,(\hat{v} \cdot \vec{r}). \tag{1.24}$$

However, it is perfectly acceptable, and sometimes preferable, to think in terms of the two-dimensional notation in (1.22); it is really the *algebra* of the matrix multiplication that matters.

The world-vectors of (1.22), with their linear transformation law, suggest strongly that spacetime in special relativity is a vector space. It is. Every inertial observer in special relativity can define the vector sum and scalar multiples of events using his or her inertial coordinate system. That is, if an event A has coordinates $(t_A, \vec{r}_A)$ and an event B has coordinates $(t_B, \vec{r}_B)$, then for scalars α and β we *define*

$$C = \alpha\,A + \beta\,B \tag{1.25}$$

to be the unique event in spacetime with coordinates

$$\begin{pmatrix} t_C \\ \vec{r}_C \end{pmatrix} := \alpha \begin{pmatrix} t_A \\ \vec{r}_A \end{pmatrix} + \beta \begin{pmatrix} t_B \\ \vec{r}_B \end{pmatrix}, \tag{1.26}$$

where the right side denotes the usual linear combination of column vectors. This process works because every event has a unique set of inertial coordinate values for a given system in special relativity. Thus, each inertial observer can *induce* the structure of a vector space — the ability to form linear combinations — on spacetime using his or her inertial coordinates. The key point, of course, is that, because the relationship between inertial coordinate systems in special relativity is *linear*, the vector structures induced by various inertial observers are *one and the same*. That is, all inertial observers, *as long as they share a common origin of coordinates*, will agree as to which particular event in spacetime results from a certain linear combination of other events. As we have remarked many times previously, such mathematical structures on which all inertial observers agree are physically

interesting and useful. Note that even the vector-space structure of spacetime arises as a result of the special assumptions we have made regarding space and time in the previous Lecture. More generally, spacetime should not be expected to form a vector space.

Minkowski spacetime is closely related to the vector space described here. Although the vector structure does not depend on the inertial state of motion of an observer, it does depend on where one takes the origin of coordinates. Thus, Minkowski spacetime is not *technically* a vector space. Like Euclidean space in Newtonian physics, it really has the structure of an ***affine space***. One can form differences of points, and the result is a true vector in an associated vector space. One can also form weighted averages of points in an affine space, but more general linear combinations do not exist. A good geometric example of an affine space is a plane in $\mathbb{R}^3$ that does *not* pass through the origin. The associated vector space is the parallel plane through the origin. We will generally ignore this technical subtlety, and assume that a particular origin has been chosen in our calculations.

We have introduced the term "world-vector" above to describe the *column vectors* formed by the coordinates assigned to particular events by an inertial observer. We must distinguish these objects carefully from the *abstract vectors* that make up Minkowski spacetime. We will call such an abstract object a ***four-vector***, and will denote them with symbols like x. Thus, an inertial observer's coordinate system can be viewed as a map

$$\mathsf{x} \xrightarrow[O]{} \begin{pmatrix} t_x \\ \vec{x} \end{pmatrix} \tag{1.27}$$ {fvwv0}

from abstract four-vectors to world-vector columns. Here, t_x and $\vec{x}$ are respectively the temporal and spatial inertial coordinates assigned by an observer O to the event in Minkowski spacetime specified by the four-vector x. This pedantic definition is necessary for us since one of the most important lessons of *general* relativity is that analyses of physical phenomena should not depend on the spacetime coordinates used to describe them. Thus, even in the special theory, we must distinguish real objects, spacetime events or four-vectors, from the coordinate-dependent objects, world-vectors, used to describe them.

The Minkowski interval (1.18) defines a ***quadratic form*** of Minkowski spacetime. That is, $\|\mathsf{x}\|^2$ is a function mapping four-vectors x to numbers, and its value scales by α^2 when we scale x to $\alpha\mathsf{x}$. In addition, this particular quadratic form obeys the ***parallelogram identity***

$$\|\mathsf{x}+\mathsf{y}\|^2 + \|\mathsf{x}-\mathsf{y}\|^2 = 2\|\mathsf{x}\|^2 + 2\|\mathsf{y}\|^2. \tag{1.28}$$ {ParId}

Whenever a quadratic form satisfies this condition, there exists an associated ***bilinear form*** defined by

$$\langle \mathsf{x}, \mathsf{y} \rangle := \tfrac{1}{4}\big(\|\mathsf{x}+\mathsf{y}\|^2 - \|\mathsf{x}-\mathsf{y}\|^2\big). \tag{1.29}$$ {aBilF}

A bilinear form maps a pair of vectors to a number, and is linear in each of its two arguments. The most familiar example of this sort of relation is that between the norm and inner product on ordinary, Euclidean space. In the case of the Minkowski interval, the corresponding bilinear form is the ***Minkowski product***

$$\mathsf{x} \cdot \mathsf{y} = -c^2\, t_x\, t_y + \vec{x} \cdot \vec{y} \tag{1.30}$$ {Mprod}

of four-vectors. This expression uses the inertial coordinates assigned to the events x and y by an arbitrary inertial observer O, and the second term on the right is the ordinary dot product of spatial vectors. As with the interval, the product of four-vectors is independent of the observer O whose coordinates we use to calculate it. It is an invariant.

Exercise 1.31: Confirm that using the interval (1.18) as the quadratic form in (1.29) does yield the Minkowski product (1.30).

The Minkowski product, and indeed any bilinear form, defines a particularly simple example of a ***tensor***, which generally is a function that takes several different vectors as arguments and produces numbers as a result. In particular, a tensor must be linear in each of its vector arguments, and of course the Minkowski product is. The tensor defined by the Minkowski product is often denoted $\boldsymbol{\eta}$, and we set

$$\boldsymbol{\eta}(\mathsf{x}, \mathsf{y}) := \mathsf{x} \cdot \mathsf{y}. \tag{1.32}$$

This tensor is called the ***Minkowski metric***. Note that, by the discussion above, this metric arises on Minkowski spacetime *precisely* because of the assumptions made in special relativity about the relationship between different inertial observers.

Just as any inertial observer O can associate a particular column vector with a given abstract four-vector using coordinates, he or she can associate a matrix with the abstract tensor $\boldsymbol{\eta}$. The matrix is quite simple:

$$\boldsymbol{\eta} \underset{O}{\longrightarrow} \begin{pmatrix} -c^2 & \\ & I\,\cdot \end{pmatrix}. \tag{1.33}$$

It has the same form for all inertial observers. Using this expression, we can easily compute the invariant Minkowski product between four-vectors whose coordinates, relative to some observer, are known:

$$\boldsymbol{\eta}(\mathsf{x}, \mathsf{y}) = \begin{pmatrix} t_x & \vec{x} \cdot \end{pmatrix} \begin{pmatrix} -c^2 & \\ & I\,\cdot \end{pmatrix} \begin{pmatrix} t_y \\ \vec{y} \end{pmatrix} = -c^2\, t_x\, t_y + \vec{x} \cdot \vec{y}. \tag{1.34}$$

Such matrix expressions are *not* fundamental, of course, but are enormously useful for practical calculations. This fact is probably familiar.

1.6 SUMMARY

This Lecture has been a long and unusual discussion of special relativity. Rather than just leave these issues, it may be a good idea to try to collect the essential results and arguments in one place in order to emphasize more clearly the key points. There really are only two.

First, the roots of modern relativity theory exist already underneath Newtonian mechanics. When we solve Newton's dynamical equations to compute the acceleration of a body, we must ask: relative to what is that body accelerating? In Newton's case, the answer is that there is a preferred Euclidean space and we compute the acceleration through that space. However, Newton's space is not something tangible. We can't touch it or move it. The question therefore becomes: how do we *observe* its effect? The answer lies in the idea of inertial motions. These motions trace out the curves that particles follow when they are *not* influenced by outside forces. These are the straight lines that define Newton's Euclidean space, and it is relative to these straight lines that we aim to compute trajectories. The principle of relativity, from its beginning with Galileo all the way through general relativity, has always been the same: It is impossible to distinguish one of these inertial motions from another in any absolute sense. The only difference between Galilean and Einsteinian relativity is that, in the latter case, we apply the principle of relativity to electromagnetism as well as to mechanics. This is what gives us Einstein's principle of a constant speed of light, and thence the indispensable idea of spacetime.

Second, special relativity makes several additional *assumptions* regarding the relationship between inertial observers in spacetime. The principles discussed above hold true for all spacetimes, including those encountered in general relativity. However, the additional assumptions of special relativity give rise to much of the rich structure of Minkowski spacetime, particularly its vector nature and metric. In the presence of gravity, as we will see below, inertial observers still exist, but do not move relative to one another along the straight lines assumed in special relativity. It therefore should come as no surprise that spacetime in general is *not* a vector space and, while it does have a metric since obviously we can compute distances and so forth, that metric is not Minkowski's flat one. Roughly speaking, spacetimes describing true gravitational fields are *curved* because the relative motions of inertial observers in those spacetimes are curved by gravity. This is the essential insight offered by Einstein's equivalence principle, discussed below.