Supplementary Lecture I: Spacetime Diagrams and Causality

This set of notes accompanied the first in a series of "fun" lectures about relativity given during the Fall 2005 Physics H7C course at UC Berkeley. Its focus is on using spacetime diagrams to understand the causal structure of Minkowski space and as a tool to solve problems in special relativity.

1 The Structure of Spacetime

Special relativity is a revolutionary theory that completely changes our understanding of space and time. Before special relativity, space and time were thought of as two unrelated things: space as a three-dimensional Euclidean coordinate grid describing position, and time as an unrelated parameter used to describe changes in position. After special relativity, space and time were inextricable combined into a new four-dimensional structure known as *Minkowski spacetime*[*]. The differences between Euclidean space and Minkowski spacetime form the core of many of the seemingly strange consequences of special relativity.

Mathematically, we can express the relationship between space and time by the *Lorentz transformations*,

$$ct' = \gamma(ct - \beta x) \tag{1a}$$

$$x' = \gamma(x - \beta ct) \tag{1b}$$

$$y' = y \tag{1c}$$

$$z' = z, \tag{1d}$$

[*]Minkowski spacetime, otherwise known as Minkowski space or just spacetime, is named after German mathematician Hermann Minkowski. He was the first to realize that the work of Lorentz and Einstein could be best understood in the context of a non-Euclidean space, and in doing so he laid much of the mathematical groundwork on which special and general relativity both rest.

which tell us how to transform between the position and time coordinates of two Lorentz frames[†] moving with relative velocity $v = \beta c$ in the x-direction. Before saying any more about these transformations, we will pause to make it clear exactly what we mean by a *frame* as opposed to an *observer*. A frame is synonymous with a system of coordinates, so that when we speak of a frame S as opposed to a frame S' we are really speaking of a set of coordinates $(ct,\ x,\ y,\ z)$ as opposed to a set $(ct',\ x',\ y',\ z')$. This is the sense in which the Lorentz transformations are used to convert between two different Lorentz frames. Clearly, then, a frame is something which covers all of spacetime.

An observer, on the other hand, is a person (or a rocket, or a particle, or any other material object) moving through spacetime. In principle, an observer need not be restricted to a particular Lorentz frame, although we often do so for convenience. If an observer does stay in the same Lorentz frame, it is tempting to speak of the observer and the frame interchangeably. However, this is a bad habit to get into, because it encourages us to describe what happens in spacetime by what a particular observer "sees," which is not what enters into the Lorentz transformations. As stated above, the Lorentz transformations convert between coordinate systems, which span *all* of spacetime, as opposed to observers, which exist at *particular points* in spacetime. Therefore, what a single human being flying through spacetime physically sees has very little to do with the underlying reality of the events occurring in that spacetime as described by the Lorentz transformations[‡].

Given the Lorentz transformations, we can define a quantity that is invariant, that is, a quantity that has the same value in any Lorentz frame; we call it the *spacetime*

[†]In special relativity, inertial frames are often referred to as Lorentz frames.

[‡]If we insist on thinking of frames as having to do with people, we can imagine an infinite array of people, each with a synchronized clock, spaced uniformly throughout space. Each of these people observes only what is happening in an infinitesimally small region of space immediately around them, and once they are done observing, they all get together to reconstruct what happened throughout spacetime. This is equivalent to speaking of a frame as a coordinate system, but it is needlessly personified.

interval between two points in spacetime. It is given by[§]

$$\begin{aligned}(\Delta s)^2 &= (c\Delta t)^2 - (\Delta x)^2 - (\Delta y)^2 - (\Delta z)^2 \\ &= c^2(t_2 - t_1)^2 - (x_2 - x_1)^2 - (y_2 - y_1)^2 - (z_2 - z_1)^2. \end{aligned} \tag{2}$$

To see that this quantity is indeed invariant, we can Lorentz transform it[¶],

$$\begin{aligned}(\Delta s')^2 &= (c\Delta t')^2 - (\Delta x')^2 - (\Delta y')^2 - (\Delta z')^2 \\ &= (\gamma(c\Delta t - \beta\Delta x))^2 - (\gamma(\Delta x - \beta c\Delta t))^2 - (\Delta y)^2 - (\Delta z)^2 \\ &= \gamma^2((c\Delta t)^2 + (\beta\Delta x)^2 - 2\beta c\Delta x\Delta t - (\Delta x)^2 - (\beta c\Delta t)^2 + 2\beta c\Delta x\Delta t) \\ &\quad -(\Delta y)^2 - (\Delta z)^2 \\ &= \gamma^2(1-\beta^2)((c\Delta t)^2 - (\Delta x)^2) - (\Delta y)^2 - (\Delta z)^2 \\ &= (c\Delta t)^2 - (\Delta x)^2 - (\Delta y)^2 - (\Delta z)^2 \\ &= (\Delta s)^2. \end{aligned} \tag{3}$$

Hence, $(\Delta s)^2$ is a Lorentz invariant.

1.1 Spacetime Diagrams

Because our day-to-day experiences are so far removed from the physics described by the above equations, much of the intuition we have about distances, times, velocities, and the like do not work in the realm of special relativity. We need to develop new tools to help us understand this new world. One of the most powerful of these is called a *spacetime diagram*.

Figure 1 is an example of a spacetime diagram; it represents all of spacetime from the point of view of a given Lorentz frame by plotting time and one spatial dimension on orthogonal axes[‖]. If we were to plot a point P on this diagram, it would have a well-defined set of coordinates given by projecting that point onto the axes in the usual way. Such points are referred to as *events*; they represent physical events (two

[§]Note that we could define the invariant interval to be the negative of what we present here without changing any of the physics involved. This choice is convention, and is in fact not the convention used by the entire physics community. Particle physicists tend to write the interval the way we are writing it, while general relativists tend to use the opposite convention.

[¶]Here we use the useful identity $\gamma^2(1-\beta^2) = 1$.

[‖]Because the surface of a piece of paper is 2-dimensional, we do not attempt to draw the y- or z-axes in this diagram. We assume that all motion occurs in the x-direction, so that we can ignore the other two spatial dimensions.

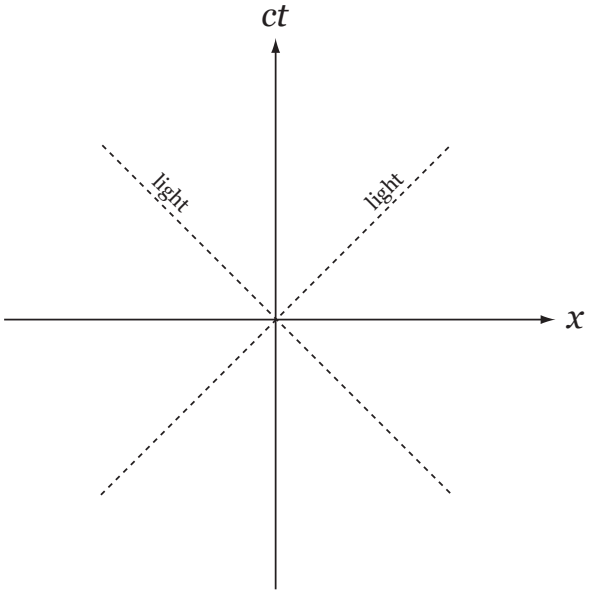

Figure 1: A spacetime diagram.

rockets colliding, a bolt of lightning striking, a star exploding, etc) that occurred at a particular position and time in a given frame.

A particle moving through spacetime traces out a curve on a spacetime diagram called a *worldline*. The worldline is simply the trajectory of the particle, but we do not use the word trajectory because it suggests only motion through space. For example, if we were asked to describe the trajectory of a rock sitting on the ground, we would likely say that it has no trajectory because it is not moving. It *does* have a worldline, however, because while its position is not changing, it is moving forward in time. Its worldline on a spacetime diagram is simply a vertical line. This is a good time to point out that we have chosen to plot position on the horizontal axis and time on the vertical axis, which is opposite the normal convention of plotting x vs. t, and that we have chosen to plot ct on the vertical axis instead of just t so that both axes carry the same units. This means that the slope of an object's worldline is given by the *reciprocal* of its velocity expressed as a fraction of c. In other words, the slope of a worldline is the reciprocal of β. Therefore, the slope of the worldline of a particle at rest is infinite, while the slope of a photon is 1.

Suppose we were to draw the trajectories of photons intersecting the origin; we produce the dotted lines in Figure 1. Imagine rotating this diagram about the ct-axis,

so that instead of intersecting just the x-axis they intersect the entire xy-plane. The light rays then form two cones. The upper cone is called the *future light cone* and the lower cone is called the *past light cone*[**].

These names are reasonable. Imagine an observer, Albert, starting out for a journey through spacetime from the origin of the spacetime diagram. Because all objects must travel at speeds less than or equal to the speed of light, the only points Albert can get to are those that lie within the future light cone. In this sense, these points define Albert's *future* because they are the set of points accessible to him from the origin. Similarly, the set of points that lie within the past light cone defines Albert's *past* because these points are the the only ones from which he might have arrived at the origin. We call the rest of spacetime the *present*, the set of points completely inaccessible to Albert[††]. These regions of spacetime are labeled in Figure 2. Note that these definitions of Albert's past, present, and future are only valid when Albert is at the origin. As Albert moves through spacetime, the set of points that are his past, present, or future changes depending on his location, as though he carries along his own set of light cones that constantly repartition these three regions. This ensures that Albert's worldline *never* has a slope less than 1.

The concepts of past, present, and future also have a more mathematical interpretation. Consider the points A, B, and C in Figure 2; A lies in the future of the origin, C in the present, and B on the future light cone. For A, $ct > x$, which is true of any point in the future. This means that the spacetime interval, given by Eq. (2), is positive for all points in the future. However, for point C, and any other point in the present, $x > ct$, which means that the spacetime interval is negative for these points. On the lightcone, $ct = x$, so the spacetime interval is identically zero for point B and any other point on the cone. These facts give rise to a new set of terminology. Given any two points separated in spacetime, the separation is said to be *timelike* if $(\Delta s)^2 > 0$, *spacelike* if $(\Delta s)^2 < 0$, and *lightlike* or *null* if $(\Delta s)^2 = 0$.

[**]Of course, because space has three spatial dimensions, the past and future light cones are really 4D hyper-cones. This is obviously impossible for us to picture, so we will stick to our simplified 2D diagrams.

[††]Compare these definitions of past, present, and future to those in Euclidean space combined with absolute time. From Albert's point of view at the origin, the future is all points with $t > 0$, the past is all points with $t < 0$, and the present is all points with $t = 0$, all without regard to position. This happens because in non-relativistic mechanics there is no speed limit, so that Albert, starting at the origin, could get to any point in space in finite time if he traveled fast enough.

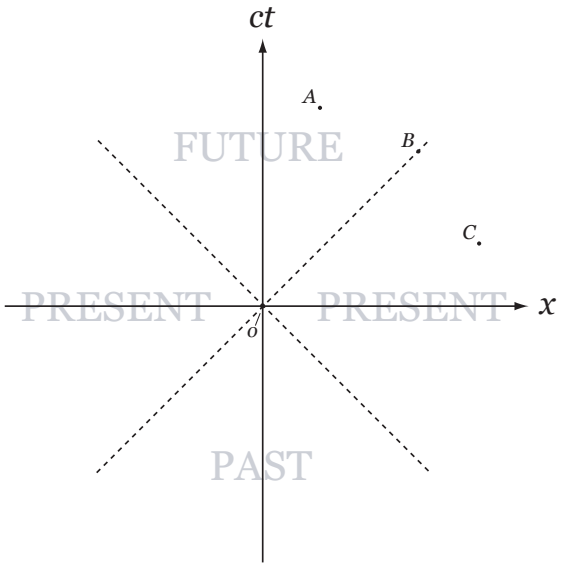

Figure 2: Three regions of spacetime.

1.1.1 The Geometry of Spacetime

Before moving on to more specific topics, we would like to say a few words about the geometry of Minkowski space as represented by spacetime diagrams. When mapmakers make maps of the Earth, they must always sacrifice some aspect of the Earth's true geometry in order to make a flat map out of a round surface. They may choose to correctly reproduce the relative sizes of the continents or to preserve the angles between lines of latitude and longitude, but they can never accurately reproduce all aspects of the round Earth. This is OK as long as people using the maps understand this distortion. In analogy to this, spacetime diagrams do not accurately reproduce the geometry of Minkowski space. By representing time and space equivalently on an intrinsically Euclidean plane, spacetime diagrams distort distances so that lines that appear longer on the diagram actually represent shorter spacetime separations.

Suppose we drew a right triangle on a spacetime diagram, like the one in Figure 3, such that one leg of the triangle has length $c\Delta t$ and the other has length Δx. We can interpret the vertices of this triangle as events: events A and C occurred at the same time but in different places while events B and C occurred at the same place but at different times. We can interpret the hypotenuse of the triangle as the separation Δs between events A and B; it is pictorially what we meant by the interval defined in

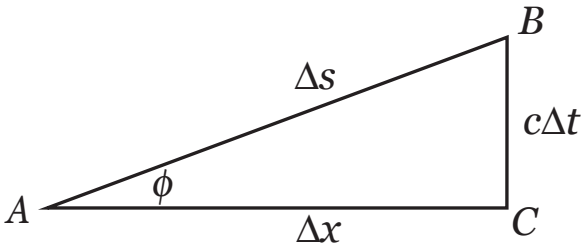

Figure 3: A right triangle from a spacetime diagram.

Eq. (2).

How can we calculate Δs? The natural approach, given our diagram, is to use the Pythagorean theorem:

$$(\Delta s)^2 = (c\Delta t)^2 + (\Delta x)^2. \tag{4}$$

However, this contradicts Eq. (2), which says that

$$(\Delta s)^2 = (c\Delta t)^2 - (\Delta x)^2. \tag{5}$$

We interpret this strange result in analogy to the example of the flat map of the Earth: although the line segment connecting events A and B in our triangle looks longer than the line segment connecting events B and C, it in fact represents a shorter spacetime interval. Just as representing the surface of a sphere on a flat piece of paper distorts geometrical features, representing Minkowski space on a flat piece of paper also distorts its geometry[‡‡]. Therefore, Minkowski space is fundamentally different from Euclidean space[✠], and our spacetime diagrams misrepresent some aspects of the true geometry of spacetime.

We can be a bit more concrete about what is causing this misrepresentation by thinking about trigonometry. Given the angle ϕ defined by Figure 3 (and forgetting

[‡‡]As an extreme example of this, imagine if $c\Delta t = \Delta x$ in our triangle; if that were the case, Δs, the "length" of the hypotenuse, would be zero!

[✠]There is a way to make the interval in Minkowski space and the interval in Euclidean space look the same: replace ct with ict in Eq. (4). This may seem like a tempting solution, but really it replaces one inconvenience, representing Minkowski space on a Euclidean surface, with a much worse one, plotting imaginary time. Since imaginary time is much harder to conceptualize than distorted distances, we will stick to plotting real time while keeping in mind the misrepresentation of distance it causes.

for the moment that we know there is something strange happening with geometry), we find that

$$\sin\phi = \frac{c\Delta t}{\Delta s} \tag{6a}$$

$$\cos\phi = \frac{\Delta x}{\Delta s}. \tag{6b}$$

If we substitute these expressions into the trigonometric identity

$$\sin^2\phi + \cos^2\phi = 1, \tag{7}$$

and simplify, we produce Eq. (4). This should not be too surprising since the Pythagorean theorem is equivalent to Eq. (7). However, we know it is the wrong expression for the interval Δs.

It turns out that there is a way to fix this: replace usual (circular) trig functions with *hyperbolic* trig functions by defining

$$\cosh\alpha = \frac{c\Delta t}{\Delta s} \tag{8a}$$

$$\sinh\alpha = \frac{\Delta x}{\Delta s}, \tag{8b}$$

where α is the *hyperbolic angle*[♉]. Then, using the identity

$$\cosh^2\alpha - \sinh^2\alpha = 1, \tag{9}$$

and simplifying, we produce Eq. (5), which is the correct expression for the interval Δs. Thus, we have discovered something interesting about the structure of Minkowski space: it is based on hyperbolas rather than circles.

One last note before we move on: we mentioned earlier the useful identity

$$\gamma^2 - (\gamma\beta)^2 = 1, \tag{10}$$

which looks similar in structure to Eq. (9). Let's define a quantity θ such that

$$\cosh\theta = \gamma \tag{11a}$$

$$\sinh\theta = \gamma\beta. \tag{11b}$$

[♉]The hyperbolic angle has an interesting geometrical interpretation that is beyond the scope of this set of notes. The Wikipedia article on "hyperbolic function" would be a good place to start learning more about this subject.

The quantity θ is called the *rapidity*, and it has an important interpretation in special relativity. If we calculate its hyperbolic tangent, we find

$$\begin{aligned} \tanh\theta &= \frac{\sinh\theta}{\cosh\theta} \\ &= \frac{\gamma\beta}{\gamma} \\ &= \beta. \end{aligned} \tag{12}$$

In other words, θ is closely related to the frame velocity β. In fact, in certain circumstances it is more useful to use θ in calculation than β, particularly when adding the velocities of several frames together. We will see an example of this in the next set of notes in this series.

1.2 Preserving Causality

One of the fundamental concepts on which all of physics is based is *causality*. Physicists rarely talk about it explicitly, especially at the undergraduate level, but it should be clear after a little thought that without the ability to definitively say "A caused B," physics as we know it would be meaningless.

The statement "A caused B" necessarily implies that A took place before B. However, we know that in special relativity it is sometimes possible to change the order in which two events seem to take place by switching to a different inertial frame. If an observer in one frame claims that A caused B while another in a different frame claims that B took place before A, we have a problem: an observer seeing an effect before its cause. Do we have to abandon the concept of causality if we are to accept special relativity? The answer is, thankfully, no, and the reason for that answer is the odd geometry of Minkowski space.

Imagine a point in spacetime with coordinates x and ct. We calculate the interval between that point and the origin to be some value I. Now imagine Lorentz transforming that point so that it has different values of x and ct. Because the interval is invariant, it will still have the value I. Therefore, the locus of points for constant I is given by

$$I = (ct)^2 - x^2, \tag{13}$$

where, as usual, we are ignoring y and z. This is the equation for a hyperbola, and it is plotted for both positive and negative I in Figure 4. For positive I, we see that the hyperbola lies entirely in the future or in the past, while for negative I it lies in the

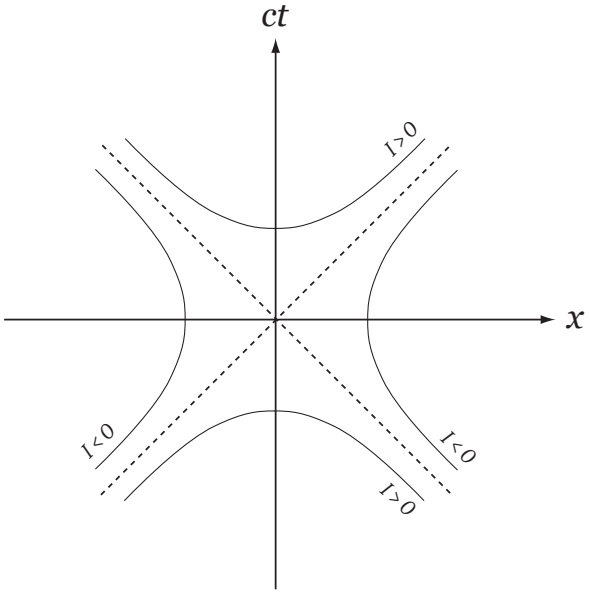

Figure 4: Hyperbolas for I greater and less than 0.

present. If we imagine rotating these pictures about the ct-axis, we produce Figure 5. We see that positive I forms a *hyperboloid of two sheets* while negative I forms a *hyperboloid of one sheet*.

What does this have to do with causality? Suppose we have two points in spacetime, the origin O and some other point P, which are timelike separated such that $t_P > 0$. Because they are timelike separated, we know that $(\Delta s)^2 = I > 0$. Therefore, point P must lie on the hyperbola contained in O's future. Since Lorentz transformations leave I invariant, after a Lorentz transformation P *must lie on the same hyperbola* that it started out on. Therefore, there is no way to move P out of the future of O! If an event at O caused an event at P, causality is preserved because the event at O *always* occurs before the event at P in *any Lorentz frame*. Thus, we conclude that events *can* be causally connected in special relativity so long as the interval between them is timelike, because this guarantees that all observers, no matter their inertial frame, will agree on the temporal order of the events

We could perform a similar analysis if the point P is spacelike separated from the origin. Unsurprisingly, we would find that such a point cannot be causally connected to the origin, because different observes will disagree about the order in which the events at O and P took place (just look at the hyperboloid for $I < 0$ to see that

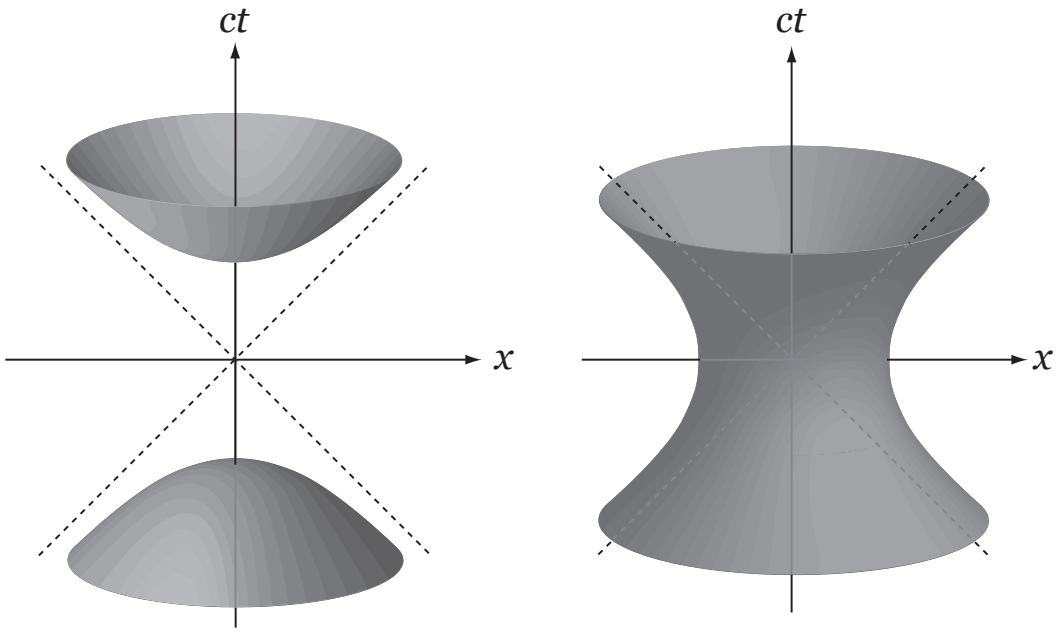

Figure 5: Hyperbolas rotated. The image on the left is the $I > 0$ hyperbolas, while the image on the right is the $I < 0$ hyperbola.

this is true). This is OK, though, because there is no way to send a signal between two spacelike separated points, so there is no practical way that anything at one such point could cause something to happen at the other anyway.

So what is the connection between these arguments and the geometry of Minkowski space? Suppose spacetime were Euclidean, with the interval given in Eq. (4). In that case, the locus of points of constant interval would be a circle around the origin, and the analysis given here would have shown that if P started out in O's future, it could be transformed into O's future, present, or past, destroying causality completely. Causality is rescued by the fact that spacetime is Minkowski in nature.

1.3 Two Frames on One Diagram

Now that we understand how to represent one Lorentz frame on a spacetime diagram, we will construct a diagram that represents two at once. We would like to know how to superimpose the axis of the second frame with the axis of the first.

First of all, we note the important facts that the x-axis defines the set of points in spacetime that have $ct = 0$ and the ct-axis defines the set of points in spacetime that have $x = 0$. Let's look at the first of the Lorentz transformations,

$$ct' = \gamma(ct - \beta x). \tag{14}$$

In analogy to the facts above, we should be able to construct the x'-axis by setting ct' equal to zero. If we do that, we find

$$ct = \beta x, \tag{15}$$

which tells us that the x'-axis can be drawn as a line in the x-ct plane that passes through O and has slope β. Similarly, we can construct the ct'-axis by setting x' equal to zero in

$$x' = \gamma(x - \beta ct), \tag{16}$$

producing

$$ct = \frac{x}{\beta}. \tag{17}$$

Therefore, the ct'-axis makes the same angle with respect to the ct-axis as the x'-axis made with respect to the x-axis; the slope of the ct'-axis is $\frac{1}{\beta}$♣. The primed and

♣This is a reasonable result since an observer at rest in the primed frame would make a slanted path in the unprimed frame, with the slope of that path equaling the reciprocal of the observer's β.

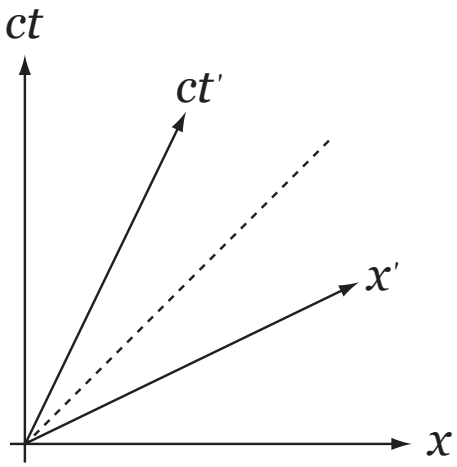

Figure 6: Two inertial frames on one spacetime diagram.

unprimed axes are shown in Figure 6. Note that in the primed coordinates, the light cone still has slope 1, which means that the speed of light is the same in both reference frames, as it should be. Note also that we are setting the origins of the primed and unprimed coordinates equal in our diagram; this is OK because spacetime is invariant under translations of the origin, so we always have the freedom to put the origins on top of each other.

By using a spacetime diagram representing two inertial frames simultaneously, most kinematics problems in special relativity reduce to the task of finding the coordinates of relevant events in the two frames. By plotting these events on a spacetime diagram and determining their coordinates in both frames with the Lorentz transformations, it is often much easier to solve problems in an understandable way than it is by doing algebra alone. We will see some examples of this below.

2 Problem-Solving with Spacetime Diagrams

Now that we know what spacetime diagrams are, let's use them derive time dilation and length contraction and to understand simultaneity.

2.1 Simultaneity

Suppose we are standing in front of a set of train tracks such that the tracks are oriented along our x-axis. A train car comes by at speed β. Just as the center of the train car is right in front of us, two lightning bolts strike. One bolt strikes the front of the train car and the other the back of the train car; both strikes are simultaneous in our frame. Are they simultaneous in the train car's frame?

In order to analyze this situation, we need to understand what the problem is asking. There are two events of relevance, the two lightning strikes; we will call them events A and B. By saying that these events look simultaneous in our reference frame, we mean that they have the same time coordinate. Clearly, however, they must have different space coordinates. Let's say that event A has coordinates $x = -d$ and $ct = 0$ and event B has coordinates $x = d$ and $ct = 0$.

To find what how these events "look" in the train car's frame, we must transform the coordinates of these events. Before blindly calculating, however, let's see what all this looks like on a spacetime diagram, like the one in Figure 7. We see pictured both sets of axes, as well as two slanted lines representing the front and back of the train car; the events A and B are labeled. Clearly, these events have the same ct-coordinate. However, they do NOT have the same ct'-coordinate. Starting from event A, we follow a line parallel to the x'-axis until we intercept the ct'-axis; this is the ct' coordinate of A, and it is positive. If we follow the same procedure for event B, we find that it has a negative ct' coordinate. This means that in the primed frame, event B happens before event A.

Let's confirm this with algebra. For event A, we have

$$\begin{aligned} ct' &= \gamma(ct - \beta x) \\ &= \gamma(0 - \beta(-d)) \\ &= \gamma\beta d, \end{aligned} \tag{18}$$

while for event B we have

$$\begin{aligned} ct' &= \gamma(ct - \beta x) \\ &= \gamma(0 - \beta(d)) \\ &= -\gamma\beta d. \end{aligned} \tag{19}$$

We see that the ct' coordinate of event B is indeed greater than that of event A. Therefore, events simultaneous in one Lorentz frame are NOT simultaneous in a different Lorentz frame.

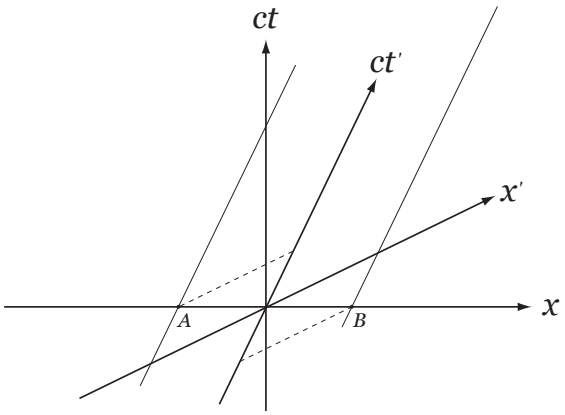

Figure 7: The spacetime diagram for the simultaneity problem.

2.2 Length Contraction

To study length contraction, let's imagine a rod at rest in a Lorentz frame, with one end at $x = 0$ and the other at $x = L_0$. By definition, L_0 is the *proper length* of the rod, because proper length is the length of an object as measured in the object's rest frame. What length is observed in a Lorentz frame moving with respect to the rest frame of the rod?

First, we draw a spacetime diagram, like the one in Figure 8. The vertical line intersecting the x-axis at L_0 represents the path of the rightmost end of the rod, while the ct-axis represents the leftmost end. The event A is the rightmost point of the rod intersecting the x-axis, while the event B is the rightmost point intersecting with the x'-axis. The length of the rod in the unprimed frame is represented by the distance between O and A, while its length in the primed frame is represented by the distance between O and B. From the looks of the diagram, the rod is longer in the moving frame. However, we know that this is incorrect, because length contraction is supposed to work the other way. How do we resolve this?

Let's look at the algebra. Event B has $ct' = 0$ and $x = L_0$. We need to calculate x'. First let's find ct:

$$ct' = \gamma(ct - \beta x) \quad \Rightarrow \quad ct = \beta L_0. \tag{20}$$

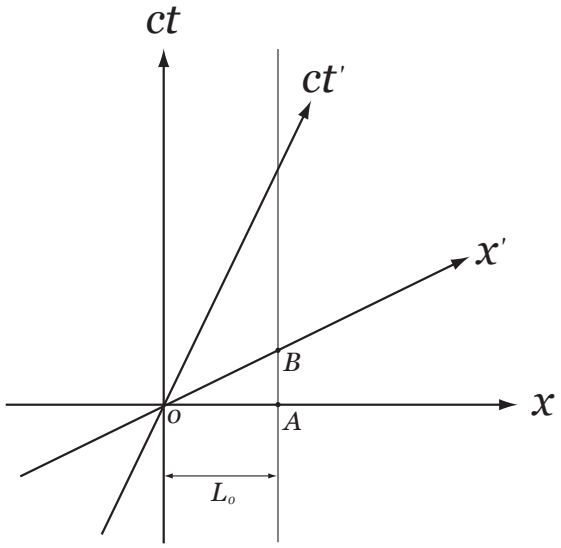

Figure 8: The spacetime diagram for length contraction.

Now we can calculate x':

$$\begin{aligned} x' &= \gamma(x - \beta ct) \\ &= \gamma(L_0 - \beta^2 L_0) \\ &= \gamma(1 - \beta^2)L_0 \\ &= \frac{L_0}{\gamma}. \end{aligned} \tag{21}$$

Since γ is always greater than 1, we see that the length as measured in the moving frame is shorter than the length measured in the rod's rest frame, which is the result we expected. Something funny is happening with geometry: the rod is shorter in the moving frame even though it looks like the opposite is true in the diagram. This is yet another example of the distortion of distances inherent in spacetime diagrams.

2.3 Time Dilation

Now let's study time dilation, which turns out to be very similar to length contraction. Imagine a clock at rest in the primed frame; it sits at $x' = 0$. It's a very simple clock: it emits a flash of light every τ units of time. Suppose it emits a flash at ct' equals 0 and again at ct' equals $c\tau$. How much time has elapsed between flashes in the unprimed frame?

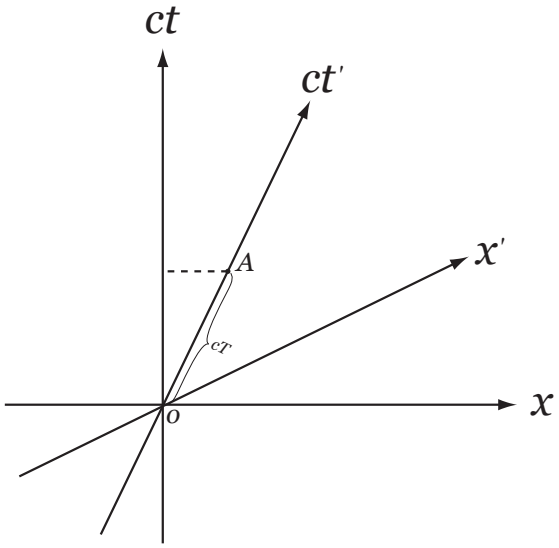

Figure 9: The spacetime diagram for time dilation.

This situation is depicted in Figure 9. The two flashes of light are our events, and they occur at points O and A. The dotted line from A to the ct-axis indicates how much time appears to have passed in the unprimed frame. It seems like less time has passed in the unprimed frame, which is opposite of what we expect, but we are again rescued by the hyperbolic geometry of spacetime.

Setting $x' = 0$ and $ct' = c\tau$, we find

$$\begin{aligned} ct &= \gamma(ct' + \beta x') \\ &= \gamma(c\tau). \end{aligned} \tag{22}$$

Thus the unprimed observer experiences the passage of a time $\gamma\tau$ while he sees the moving clock register the passage of a time τ, where τ is by definition the *proper time* of the clock, the time that elapsed in the clock's own rest frame. In other words, the unprimed observer sees the moving clock run slowly.

3 Proper Time and the Twin Paradox

The final topic that we will discuss is the twin paradox. Suppose we have a set of twins, Fred and George, for the sake of definiteness. Fred stays home on Earth, while George sets off in a rocket ship. George travels directly away from the Earth at

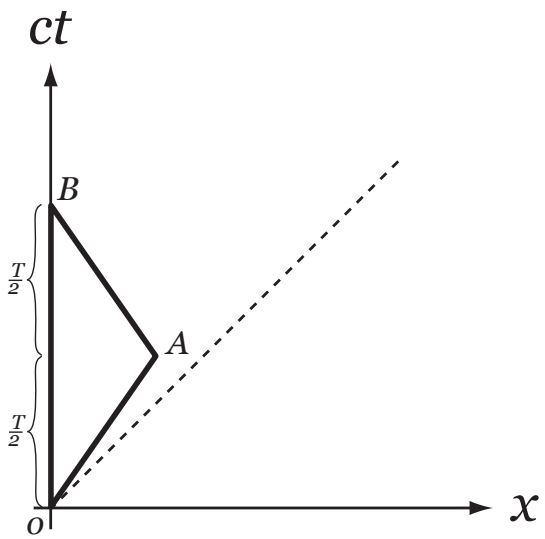

Figure 10: The spacetime diagram for the twin paradox.

constant speed, turns around, and returns home at that same speed; we will neglect the short periods of acceleration necessary to change his velocity. From Fred's point of view, George spends a time $T/2$ on his outward journey, and the same time on his return journey. This situation is depicted in Figure 10.

We would like to know which twin is older upon their reunion; in other words, we would like to know which of the two twins has experienced more proper time between their separation and reunion. For Fred, who remained on Earth, the amount of proper time experienced is T. The question before us is to calculate the proper time experienced by George.

To do this, we turn once again to the spacetime interval. We know that this interval is invariant under Lorentz transformations, but we have not discussed the physical interpretation of this invariant quantity. In order to do so, let's restrict our attention to the case in which the interval is timelike, so that it is positive. We can then take it's square root without worrying about generating imaginary numbers. Let's also think about infinitesimal length and time separations between two events (dx and dt) instead of finite separations (Δx and Δt). Then we can define a quantity

ds such that

$$\begin{aligned} ds &= \sqrt{c^2(dt)^2 - (dx)^2} \\ &= c\,dt\sqrt{1 - \frac{1}{c^2}\left(\frac{dx}{dt}\right)^2} \\ &= c\,dt\sqrt{1-\beta^2} \\ &= c\frac{dt}{\gamma}. \end{aligned} \tag{23}$$

When γ is constant (on our diagram, when the path under consideration is a straight line), this integrates to

$$s = \frac{ct}{\gamma}. \tag{24}$$

Comparing this to our expression for time dilation, Eq. (22), we see that s (up to a factor of c) is nothing but the proper time for an object traveling at γ such that a coordinate time t has passed in in unprimed frame. Therefore, when the interval is timelike, we can interpret it as the proper time between two events♠

$$(\Delta s)^2 = (c\Delta\tau)^2, \tag{25}$$

and we find that

$$\tau = \frac{t}{\gamma}. \tag{26}$$

Let's apply this to Fred and George. For Fred, $\beta = 0$ and therefore $\gamma = 1$. If the twins separate at $t = 0$ and reunite at $t = T$, then Fred has experienced a proper time of $\tau_F = T$. However, George has been traveling with nonzero but constant β. Therefore, he experiences a proper time of $\tau_G = \frac{T}{\gamma}$. We find that

$$\tau_F > \tau_G, \tag{27}$$

as expected.

3.1 Final Remarks on Proper Time

We leave these notes with two important remarks about proper time. First, let's think about how to minimize or maximize the proper time between two events. We all know that the straight line between two points in Euclidean space is always the

♠If the interval is spacelike, we can interpret it as the *proper length* or *rest length* between the two events. In that case, $(\Delta s)^2 = -(\Delta\Sigma)^2$, where Σ is the proper length.

path of minimum distance. It turns out that the straight line connecting two timelike-separated events in spacetime is always the path of *maximum* proper time; this is yet another result of the peculiar geometry of Minkowski space. Consider the case of Fred and George. Fred followed the straight line path through spacetime between the events marked by the twins' departure and arrival, whereas George followed a different path. Given the form of Eq. (26), we see that George's proper time must always be less than Fred's, no matter his speed, because his γ is greater than one. Hence, Fred followed the path of maximum proper time between those two events. For two arbitrary timelike separated events, there is always a straight-line path connecting them that corresponds to the worldline of an inertial observer. By transforming into the frame of that observer, we reproduce the situation of Fred and George because that observer is now at rest. Therefore, the straight line path always maximizes the proper time between two events.

So what is the minimum possible proper time between two events? We see from Eq. (26) that as γ increases, τ decreases. The limiting case is the one where γ equals infinity, which is the case for a ray of light traveling at c. Hence, the proper time elapsed for light is *always zero*; colloquially, we say that time does not pass for a ray of light. This means that we can always connect two events by a path of proper length equal to 0 by connecting them by a series of lines of slope 1.

The second important insight has to do with accelerations. The way that introductory special relativity is usually taught makes it seem like SR completely breaks down if an object accelerates. After all, the frames we consider are always inertial, so what place could accelerations possibly have in this theory? The answer is that while *frames* must always be inertial, no one ever said that *observers* all had to be; this is, in fact, why I made such a big distinction between frames and observers at the beginning of these notes. Given a Lorentz frame, an observer can move in an accelerated fashion with respect to the the frame; if drawn on a spacetime diagram, his worldline would be a smooth curve with slope always greater than 1.

How do we reconcile accelerated observers with our standard ideas about Lorentz frames? Take Eq. (24). We arrived at it by assuming that γ was constant. Suppose, however, that γ were changing, which is the same as supposing that the observer it represents is accelerating. We could write the proper time of such an observer as

$$\tau = \int_{t_1}^{t_2} \frac{dt}{\gamma(t)}. \tag{28}$$

This is nice mathematically, but we need to make sure it has a sensible physical

interpretation. In fact, the interpretation is quite straightforward: we say that the accelerating observer is in a *locally inertial frame* with a well-defined velocity at any given point in time. The specific local frame changes over time, but for infinitesimal separations in time the frame itself only changes infinitesimally. This allows us to integrate $d\tau$ (for example) over the observer's entire path, arriving at a quantity whose only possible interpretation is as the observer's proper time. This sort of analysis works with any quantity that is a function of path through spacetime, so special relativity can deal with accelerating observers just fine.