Relativity and FTL Travel
Edition 5.2

Jason W. Hinson[1]

Last Modified: December 9, 2025

[1] mailto:hinson@physicguy.com

Part 0

Introduction to the FAQ

Note: This document is also available on the Web[1]

0.1 What is this FAQ About, and Who Should Read It

The primary purpose of this FAQ is to discuss, in straightforward and simple terms, the relationship between relativity and faster-than-light (FTL) travel. Part I introduces the aspects of special relativity needed to understand the FTL Travel discussion. One of its sections includes an introduction to space-time diagrams, which are used to illustrate several key points later on. If you're not familiar with these diagrams, this will be an especially important section to read.

Parts II and III are what I call "optional reading." You don't have to read them to follow the FTL discussion, but they may be helpful. Part II provides additional information about special relativity and explains two "solvable paradoxes" associated with the theory, while Part III introduces and explores several major concepts in general relativity, doing so at considerable length.

Part IV is the section that directly discusses the question of FTL travel. It addresses two basic problems that arise. Although most science-fiction concepts for FTL travel deal with the first problem (the light-speed barrier), many simply ignore the second (which leads to unsolvable paradoxes). Part IV also examines various conceptual forms of FTL travel (each of which circumvents the "first problem") and introduces special provisions for avoiding the "second problem." Finally, because this FAQ was originally written for the rec.arts.startrek.tech Usenet newsgroup, I apply the FTL discussion to show how warp travel, as depicted in Trek, can allow FTL motion while avoiding both problems. As far as I can tell, this provides the best explanation for everything shown in the series.

I hope you can learn something from reading this, or at least strengthen your understanding of what you already know. Your comments and criticisms are welcome, especially if they suggest improvements I can make in future revisions. If you ever notice missing or truncated sections in any of the available formats, feel free to let me know, and I can provide the material directly.

0.2 Edition Information

This is edition 5.2 of the Relativity and FTL Travel FAQ. The main text has not significantly changed since version 5.1, though I did fix a few issues and made improvements here and there. This edition is published in three versions (or formats): HTML, plain text, and PDF. See Section 0.3: The Versions (below) to learn more about these versions.

As usual, I cannot guarantee the FAQ doesn't contains mistakes here and there I have yet to catch. Therefore, as always, if you see any mistakes or if you think that any changes should be made, please let me know.

Edition 5.2 was completed on December 9, 2025. Any later modifications involve smaller changes and/or corrections.

Here is some information about previous editions: No detailed information was kept concerning changes in editions 1 through 3, but they were all single-part documents concerned mainly with giving the reader a

[1] http://www.physicsguy.com/ftl/

quick introduction special relativity and explaining how FTL travel seemed impossible because of it. They also included only one “special provision” for getting around all the problems of relativity (that provision being the use of a special frame of reference).

In edition 4.0b, the FAQ was split into five parts (an introduction and four parts to the FAQ itself). In addition to one part that introduced special relativity and another that discussed FTL travel, two completely new parts were added (one which looked further into special relativity and one which introduced general relativity). In this edition, the FAQ was also made available in an HTML version (though all diagrams and equations were still in ASCII).

In edition 4.1b (completed September 8, 1995) I added another “special prevision” in the FTL section.

With edition 5.0b1 of this FAQ (completed on July 11, 1997), in addition to a text version and an HTML version with only ASCII graphics, I also made it available in a graphics-rich HTML version and a LaTeX version! In 5.0b1, I also made changes in various chapters (expanding some material and moving some into new sections) to help improve explanations and readability. In general, the basic information included has not changed, but the FAQ was hopefully made even more understandable to its readers.

Edition 5.1 (completed on September 6, 1999) added a couple of subsections to the discussion of relativistic energy and momentum, and a few other minor corrections were made over the previous edition.

0.3 The Versions

As of edition 5.2, the FAQ is available in three formats. The HTML version includes cross-reference links and vector-based graphics for equations and figures. The FAQ is also provided in its original text-only format—similar to the version once posted to the old rec.arts.startrek.tech Usenet newsgroup. Like the HTML version, the text-only version is divided into an introduction and four parts, but it presents figures and equations using ASCII graphics. Finally, if you prefer to have the entire FAQ as a single document, you can download it as a PDF file. The PDF version is prepared and published using LaTeX. If you are interested in the LaTeX source material used to create it, feel free to contact me and I can send it to you.

As a historical note, when I published edition 5.0b1, the FAQ was available in several different formats: a set of plain-text files with ASCII graphics; HTML pages that used ASCII graphics (to support users with limited internet speeds); HTML pages with GIF-based diagrams and equations; a LaTeX package; and a PostScript file (an older print-oriented format that served as a precursor to PDF).

For those curious about LaTeX, it is a high-quality document preparation system that uses markup coding to produce professional scientific and technical documents and to handle complex mathematical expressions. You can read more at the LaTeX Project website[2]. I also want to give special thanks to Ricardo Aler Mur for his help in converting the original text into a LaTeX version. He provided an excellent foundation for the final LaTeX edition, and without his work I might never taken the time to do the conversion.

[2] https://www.latex-project.org/about/

Contents

Part I

Special Relativity

This is Part I of the “Relativity and FTL Travel” FAQ. It contains basic information about the theory of special relativity. In the FTL discussion (Part IV of this FAQ), it is assumed that the reader understands the concepts discussed below, while it is not assumed that the reader has read Parts II and IV of this FAQ as they are “optional reading”. Therefore, if the reader is unfamiliar with special relativity in general (and especially if the reader is unfamiliar with space-time diagrams) then he or she should read this part of the FAQ to understand the FTL discussion in Part IV.

For more information about this FAQ (including copyright information and a table of contents for all parts of the FAQ), see the Introduction to the FAQ portion.

Chapter 1

An Introduction to Special Relativity

The main goal of this introduction is to make relativity and its consequences feasible to those who have not seen them before. It should also reinforce such ideas for those who are already somewhat familiar with them. This introduction will not really follow the traditional way in which relativity came about, but it will try to explain the concepts through an easy to follow perspective. After discussing the basic terminology, the introduction will discuss points in the pre-Einstein view of relativity. It will then give some reasoning for why Einstein's view is plausible. This will lead to a discussion of some of the consequences this theory has, odd as they may seem. Finally, I want to mention some experimental evidence that supports the theory.

1.1 Relativity Terminology

As we begin our discussion, I want to first introduce the reader to some terms which will be used. The first term to consider is the obvious one, "relativity". Why is this field of study called relativity? Well, it involves considering how an event or series of events would look to one observer given that you know how it looks to another observer who may be moving with respect to the first. This is called "transforming" the observation from one frame to another, and relativity tells us how to do that. Thus, we are concerned with the way something seems to one observer **relative to** how it seems to another. Certain measurements or calculations will be the same regardless of your frame of reference. They are "frame independent" or "absolute" or "invariant" in nature. Other aspects of our universe depend greatly on your frame of reference, and they are thus "frame dependent" or "relative" in their nature. Relativity is thus study of the relative nature of things in our universe.

In that last paragraph, I use the term "frame of reference," and I should take a moment to explain what it is I am talking about. By "frame of reference" I sort of mean the "point of view" of a particular observer. Essentially, your frame of reference is what decides your relative "view" of things, so observers in different reference frames will have different relative "views". In special relativity, moving with respect to another observer is what makes your frame of reference different from his. Note too that frames of reference are relative, so that what we are really concerned with is what one frame of reference is like with respect to another frame of reference. Thus, we would say that your frame of reference relative to another frame depends on your velocity in that other frame of reference.

Now it is very easy for a newcomer to relativity to get mislead by this concept of frame of reference. The sticky phrase in the above explanation is "relative 'view' of things." You see, whenever I talk about when something occurs in some frame of reference, I **do not** necessarily mean what the observer in that frame would actually see with their eyes. This is because the observer only sees the event after the light from the event reaches him. To figure out when the event actually occurred for that observer, one must account for the "signal delay". For example, an observer may see an event today, but if the event occurred on some star ten light-years away (the distance light would travel in 10 years) in this observer's frame, then we must realize that the event actually occurred ten years ago in this observer's frame of reference (because then light from the event would just be reaching the observer today). I mention this because it is sometimes tempting for newcomers of relativity to conclude that its odd effects (like time dilation–which we will discuss later in this chapter) are only illusions created by the fact that light from an event may reach one observer before it reaches another. However, here I am clearly stating that when we talk about when an event occurs in a frame of reference, we are talking about when it **actually occurred** in that frame after all light signal delays are taken into account.

Similarly, if I say that event A and event B occur simultaneously in some frame of reference, I do not mean that an observer in that frame would necessarily see them occur at the same time, but rather that they actually happened at the same time. For example, if two explosions really happened at the same time in our frame of reference, and one occurred on the moon while the other occurred on the sun, then we would see the one from on the moon first (because it is closer). However, we must take into account the time it takes the light to get to us. We must note that it would take longer for the light from the explosion on the sun to get to us, and we can then understand why we saw the explosion on the moon first. Then, with the proper calculations, we could conclude that the explosions actually happened at the same time in our frame. It will be important to remember that this is what we mean as we talk about when and where events occur in different frames of reference (especially in Chapter 2).

Now, with these terms and considerations in mind, we can go on to reason as to why the theory of relativity exists as it does today.

1.2 Reasoning for its Existence

Before Einstein, there was Newton, and Newtonian physics had its own concept of relativity; however, it was incomplete. Remember that relativity involves figuring out what an observation would seem like to one observer once you knew what it looked like to another observer who is moving with respect to the first. Before Einstein, this transformation from one frame to another was not completely correct, but it seemed so in the realm of small speeds.

Here is an example of the Newtonian idea of transforming from one frame of reference to another. Consider two observers, you and me, for example. Let's say I am on a train (in some enclosed, see-through car–if you want to visualize the situation) that passes you at 30 miles per hour. I throw a ball in the direction the train is moving such that the ball moves at 10 mph in MY point of view. Now consider a mark on the train tracks which starts out ahead of the train. As I am holding the ball (before I throw it), you will see it moving along at the same speed I am moving (the speed of the train). When I throw the ball, you will see that the ball is able to reach the mark on the track before I do. So to you, the ball is moving even faster than I (and the train). Obviously, it seems as if the speed of the ball with respect to you is just the speed of the ball with respect to me plus the speed of me with respect to you. So, the speed of the ball with respect to you = 10 mph + 30 mph = 40 mph.

This was the first, simple idea for transforming velocities from one frame of reference to another. It tries to explain a bit about observations of one observer relative to another observer's observations. In other words, this was part of the first concept of relativity, but it is incomplete.

Now I introduce you to an important postulate that leads to the concept of relativity that we have today. I believe it will seem quite reasonable. I state it as it appears in a physics text by Serway: "the laws of physics are the same in every inertial frame of reference." (Note that by "inertial frame of reference" we basically mean a frame of reference which is not accelerating.) What the postulate means is that if two observers are moving at a constant speed with respect to one another, and one observes any physical laws for a given situation in their frame of reference, then the other observer must also agree that those physical laws apply to that situation.

As an example, consider the conservation of momentum (which I will briefly explain here). Say that there are two balls coming straight at one another. They collide and go off in opposite directions. Conservation of momentum says that if you add up the total momentum (which for small velocities is given by the mass of the ball times its velocity) of both the balls before the collision and after the collision, then the two should be identical. Now, let this experiment be performed on a train where one ball is moving in the same direction as the train, and the other is moving in the opposite direction. An outside observer would say that the initial and final velocities of the balls are one thing, while an observer on the train would say they were something different. However, **BOTH** observers must agree that the total momentum is conserved. One will say that momentum was conserved because the total momentum before **AND** after the collision were both some number, A; while the other will say that momentum was conserved because the total momentum before **AND** after were both some other number, B. They will disagree on what the actual numbers are, but they will agree that the law holds. We should be able to apply this postulate to any physical law. If not, (i.e., if physical laws were different for different frames of reference) then we could change the laws of physics just by traveling in a particular reference frame.

A very interesting result occurs when you apply this postulate to the laws of electrodynamics (the area of physics which deals with electricity and magnetism). What one finds is that in order for the laws of electrodynamics to be the same in all inertial reference frames, it must be true that the speed of electromagnetic waves (such as light) is the same for all inertial observers. Perhaps the easiest way to explain why this is so is to discuss two constants used in basic electrodynamics. They are denoted as ϵ_0 and μ_0. ϵ_0 is used in the basic equation which describes the attraction or repulsion between two electrically charged particles while μ_0 is used in the basic equation which describes the magnetic force on a charged particle. According to electrodynamics, these two constants are properties of the universe, and if any observer in any frame of reference does an electro-magnetic experiment to measure those constants, he or she must always come up with the same answers. However, it is also a property of electrodynamics that the speed (c) of an electro-magnetic wave (such as light) can be expressed in terms of those two constants: $c = 1/\sqrt{\mu_0\epsilon_0}$. If ϵ_0 and μ_0 are constants for all inertial observers, then so is c.

Thus, requiring the laws of electrodynamics to be the same for all inertial observers suggests that the speed of light should be the same for all inertial observers. Simply stating that may not make you think that there is anything that interesting about it, but it has amazing and far-reaching consequences. Consider letting a beam of light take the place of the ball in our earlier example (the one where I was on a train throwing a ball, and you were outside the train). If the train is moving at half the velocity of light (c), and I say that the light beam is traveling at the speed c with respect to me, wouldn't you expect the light beam to look as if it were traveling one and a half that speed with respect to you? Well, because of the postulate above, this is not the case, and the old ideas of relativity in Newton's day fail to explain the situation. All observers must agree that the speed of any light beam is c, regardless of their frame of reference. Thus, even though I measure the speed of the light beam to be c with respect to me, and you see me traveling past you and one half that speed, still, you must also agree that the light is traveling at the speed c with respect to you. This obviously seems odd at first glance, but time dilation and length contraction are what account for the peculiarity.

1.3 Time Dilation and Length Contraction Effects

Now, I give an example of how time dilation can help explain a peculiarity that arises from the above concept. Again we consider a case where I am on a train and you are outside the train, but let's give the train a speed of $0.6c$ with respect to you. (Note that c is generally used to denote the speed of light which is 300,000,000 meters per second. We can also write this as 3E8 m/s where "3E8" means 3 times 10 to the eighth). Now I (on the train) shine a small burst or pulse of light so that (to me) the light goes straight up, hits a mirror at the top of the train, and bounces back to the floor of the train where some instrument detects it. Now, in your point of view (outside the train), that pulse of light does not travel straight up and straight down, but makes an up-side-down "V" shape because of the motion of the train. This is not just some "illusion", but rather it is truly the way the light travels **relative to you**, and thus this is truly the way the situation must be considered in your frame of reference. Below is a diagram of what occurs in your frame: [!ht]

Let's say that the trip up takes 10 seconds in your frame of reference. The distance the train travels during that time is given by its velocity ($0.6c$) multiplied by that time of 10 seconds:

$$(0.6 \times 3\text{E}8\, m/s)10\, s = 18\text{E}8\, m \tag{1:1}$$

The distance that the light pulse travels on the way up (the slanted line to the left) must be given by its speed with respect to you (which **must** be c given our previous discussion) multiplied by the time of 10 seconds:

$$3\text{E}8\, m/s \times 10s = 30\text{E}8\, m \tag{1:2}$$

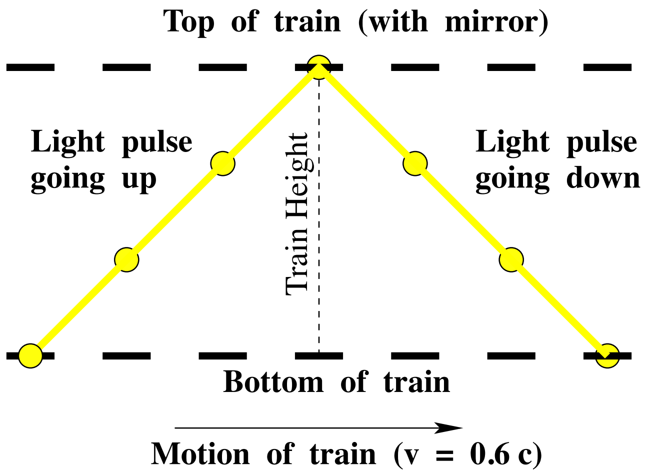

Diagram 1-1:

Since the left side of the above figure is a right triangle, and we know the length of its hypotenuse (the path of the light pulse) and one of its sides (the distance the train traveled), we can now solve for the height of the train using the Pythagorean theorem. That theorem states that for a right triangle the length of the hypotenuse squared is equal to the length of one of the sides squared plus the length of the other side squared. We can thus write the following:

$$\begin{aligned} &\text{Height}^2 + (18\text{E}8\,m)^2 = (30\text{E}8\,m)^2 \\ &\text{so} \\ &\text{Height} = \sqrt{(30\text{E}8\,m)^2 - (18\text{E}8\,m)^2} = 24\text{E}8\,m \end{aligned} \tag{1:3}$$

(It is a tall train because we said that it took the light 10 seconds to reach the top, but this IS just a thought experiment.) Now we consider my frame of reference (on the train). In my frame, the light is truly traveling straight up and straight back down to me. This is truly the way the light travels in my frame of reference, and so that's the way we must analyze the situation relative to me. Again, according to our previous discussion, the light **must** travel at 3E8 m/s as measured by me as well. Further the height of the train doesn't change because relativity doesn't affect lengths perpendicular to the direction of motion. Therefore, we can calculate how long it takes for the light to reach the top of the train in my frame of reference. That is given by the distance (the height of the train) divided by the speed of the light pulse (c):

$$\frac{24\text{E}8\,m}{3\text{E}8\,m/s} = 8\,\text{seconds}, \tag{1:4}$$

and there you have it. To you the event takes 10 seconds, while according to me it must take only 8 seconds. We measure time in different ways.

You see, to you the distance the light travels is longer than the height of the train (see the diagram). So, the only way I (on the train) could say that the light traveled the height of the train while you say that the **SAME** light travels a longer distance is if we either (1) have different ideas for the speed of the light because we are in different frames of reference, or (2) we have different ideas for the time it takes the light to travel because we are in different frames of reference. Now, in Newton's days, they would believe that the former were true. The light would be no different from, say, a ball, and observers in different frames of reference can observer different speeds for a ball (remember our first "train" example in this introduction). However, with the principles of Einstein's relativity, we find that the speed of light is unlike other speeds in that it must always be the same regardless of your frame of reference. Thus, the second explanation must be the case, and in your frame of reference, my clock (on the fast moving train) is going slower than yours.

As I mentioned in the last part of the previous section, length contraction is another consequence of relativity. Consider the same two travelers in our previous example, and let each of them hold a meter stick horizontally (so that the length of the stick is oriented in the direction of motion of the train). To the outside observer (you), the meter stick of the traveler on the train (me) will look as if it is shorter than a meter. One can actually derive this given the time dilation effect (which we have already derived), but I wont go through that explanation for the sake of time.

Now, **don't be fooled**! One of the first concepts which can get into the mind of a newcomer to relativity involves a statement like, "if you are moving, your clock slows down." However, the question of which clock is **really** running slowly (yours or mine) has **no** absolute answer! It is important to remember that all inertial motion is relative. That is, there is no such thing as absolute inertial motion. You cannot say that it is the train that is absolutely moving and that you are the one who is actually sitting still.

Have you ever had the experience of sitting in a car, noticing that you seemed to be moving backwards, and then realizing that it was the car beside which was "actually" moving forward. Well, the only reason you say that "actually" the other car was moving forward is because you are considering the ground to be stationary, and it was the other car who was moving with respect to the ground rather than your car. Before you looked at the ground (or surrounding scenery) you had no way of knowing which of you was "really" moving. Now, if you did this in space (with space ships instead of cars), and there were no other objects around to reference to, and neither space ship was accelerating (they were moving at a constant speed with respect to one another) then what would be the difference in saying that your space ship was the one that was moving or saying that it was the other space ship that was moving? As long as neither of you is undergoing an acceleration (which would mean you were not in an inertial frame of reference) there is no absolute answer to the question of which one of you is moving and which of you is sitting still. You are moving with respect to him, but then again, he is moving with respect to you. All motion is relative, and all inertial frames are equivalent.

So what does that mean for us in this "train" example. Well, from my point of view on the train, I am the one who is sitting still, while you zip past me at 0.6 c. Since I can apply the concepts of relativity just as you can (that's the postulate of relativity–all physical laws are the same for all inertial observers), and in my frame of reference you are the one who is in motion, that means that I will think that it is **your** clock that is running slowly and that **your** meter sticks are length contracted.

So, there is **NO** absolute answer to the question of which of our clocks is **really** running slower than the other and which of our meter sticks is **really** length contracted smaller than the other. The only way to answer this question is relative to whose frame of reference you are considering. In my frame of reference your clock is running slower than mine, but in your frame of reference my clock is running slower than yours. This lends itself over to what seem to be paradoxes such as "the twin paradox" (doesn't it seem like a paradox that we each believe that the other person's clock is running slower than our own?). Understanding these paradoxes can be a key to really grasping some major concepts of special relativity. The explanation of these paradoxes will be given for the interested reader in Part II of this FAQ.

1.4 Introducing Gamma (γ)

Now, the closer one gets to the speed of light with respect to an observer, the slower ones clock ticks and the shorter ones meter stick will be in the frame of reference of that observer. The factor which determines the amount of length contraction and time dilation is called gamma (denoted γ).

Gamma (γ) for an object moving with speed v in your frame of reference is defined as

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}} \tag{1:5}$$

For our train (for which $v = 0.6c$ in your frame of reference), γ is 1.25 in your frame. Lengths will be contracted and time dilated (as seen by you–the outside observer) by a factor of $\frac{1}{\gamma} = 0.8$. That is what we demonstrated in our example by showing that the difference in measured times was 10 seconds for you (off the train) and 8 seconds for me (on the train) in your point of view. Gamma is obviously an important number in relativity, and it will appear as we discuss other consequences of the theory (including the effects of special relativity on energy and momentum considerations).

1.5 Energy and Momentum Considerations

Another consequence of relativity is a relationship between mass, energy, and momentum. Note that velocity involves the question of how far you go and how long it takes. Obviously, if relativity affects the way observers view lengths and times relative to one another, one could expect that any Newtonian concepts involving velocity might need to be re-thought. For example, because of relativity we can no longer simply add velocities to transform from one frame to another as we did with the ball and the train earlier. (However, for small velocities like we see every day, the differences which comes in because of relativity are much to small for us to notice).

Further, consider momentum (which in Newtonian mechanics is defined as mass times velocity). With relativity, this value is no longer conserved in different reference frames when an interaction takes place. The quantity that is conserved is relativistic momentum which is defined as

$$p = \gamma\, mv \tag{1:6}$$

where gamma (γ) is defined in the previous section.

By further considering conservation of momentum and energy as viewed from two frames of reference, one can find that the following equation must be true for the total energy of an unbound particle:

$$E^2 = p^2c^2 + m^2c^4 \tag{1:7}$$

Where E is energy, m is mass, and p is the relativistic momentum as defined above.

Now, by manipulating the above equations, one can find another way to express the total energy as

$$E = \gamma\, mc^2 \tag{1:8}$$

Notice that even when an object is at rest ($\gamma = 1$) it still has an energy of

$$E = mc^2 \tag{1:9}$$

Many of you have seen something like this stated in context with the theory of relativity ("E equals m c squared"). It says that because of the relationship between space and time for different observers as discovered by special relativity, we must conclude that an object possesses an internal energy contained in its mass–mass itself contains energy, or, to put it more eloquently, mass is simply a convenient form of energy.

1.5.1 Rest Mass versus "Observed Mass"

It is important to note that the mass, m, in the above equations has a special definition which we will now discuss (by "mass", we generally mean the property of an object that indicates (1) how much force is needed to cause the object to have a certain acceleration and (2) how much gravitational pull you will feel from that object in Newtonian gravitation). First, note what happens to the relativistic momentum (Equation 1:6) of an object as its speed approaches c with respect to some observer. In that observer's frame of reference, its momentum becomes very large (because γ goes to infinity), especially compared to the old definition of momentum, $p = mv$. However, if we define a property called "observed mass" as being γm, then we see that the momentum can be written as

$$p = (\text{observed mass})\; v \tag{1:10}$$

We see that the momentum can be written exactly as it was in Newtonian physics, except that it seems the mass of the object as seen by an outside observer is larger than its "rest mass" (m). Further, if we take the relativistic equation for the energy of an object, Equation 1:8, we see it too can be written as

$$E = (\text{observed mass})\; c^2 \tag{1:11}$$

This is like the energy of an object at rest ($E = mc^2$) with the "observed mass" substituted in for the "rest mass."

Thus, one way to interpret relativity's effect on our view of momentum and energy is to say that because of relativity, an observer sees an object's mass increase as the object approaches the speed of light in that observer's frame of reference. The mass (m) in our equations is thus the mass as measured when the object is at rest in our frame of reference (the rest mass), not the "observed mass" we have defined.

However, this concept of observed mass doesn't really work for gravitational mass. In a relativistic setting, you can't figure out the gravitational effects of an object that is moving (in your frame) by simply figuring out what gravitational effects its mass would have at rest and replacing its mass with the observed mass in your frame of reference. For example, as the velocity of an object with respect to you approaches c, its "observed mass" approaches infinity. However, this does not mean that the object will eventually look like a black hole predicted by general relativity (as it would if the same object really did have a huge mass sitting at rest).

Also, let's look at kinetic energy in relation to mass. Kinetic energy is energy of motion–it's the total energy of a free object minus the amount of that energy that is internal to the object:

$$\begin{aligned} E_{kinetic} &= E_{total} - E_{internal} \\ &= \gamma mc^2 - mc^2 \\ &= (\gamma - 1)mc^2 \end{aligned} \tag{1:12}$$

As it turns out, when v is much smaller than c, the equation $\gamma - 1$ is approximately equal to $\frac{1}{2}\frac{v^2}{c^2}$ such that $E_{kinetic}$ is approximately $\frac{1}{2}mv^2$ (that's the Newtonian equation for kinetic energy which is approximately correct for non-relativistic speeds). But with relativistic velocities, the kinetic energy becomes much larger than we would have calculated it to be using the Newtonian equations. In that sense, there does seem to be some "extra energy" which could be considered as extra mass energy; however, you can't get the correct kinetic energy in relativity by simply plugging our expression for "observed mass" into the Newtonian equation for kinetic energy. The observed mass concept doesn't really work here, and we see that it's better to simply argue that the mass isn't really increasing, but rather the equations for energy and momentum are different than expressed by Newtonian physics.

So, "observed mass" has its uses, but physicists today rarely use the concept in practice. Rather, an object is said to have a rest mass (which truly is its inherent internal energy) as well as an energy due to its motion with respect to an observer (kinetic energy) which come together to produce its total energy, E.

1.5.2 The Energy and Momentum of a Photon (Where $m = 0$)

We should quickly note the case where the rest mass of an object is zero (such is the case for a photon–a particle of light). Given the equation for the energy in the form of Equation 1:8 ($E = \gamma mc^2$), one might at first glance think that the energy was zero when $m = 0$. However, note that massless particles like the photon travel at the speed of light. Since γ goes to infinity as the velocity of an object goes to c, the equation $E = \gamma mc^2$ involves one part which goes to zero (m) and one part which goes to infinity (γ). Thus, it is not obvious what the energy would be. However, if we use the energy equation in the form of Equation 1:7 ($E^2 = p^2c^2 + m^2c^4$), then we can see that when $m = 0$ then the energy is given by $E = pc$).

Now, a photon has a momentum (it can "slam" into particles and change their motion, for example) which is determined by its wavelength (λ) in the equation $p = h/\lambda$ (where $h = 6.626 \times 10^{-34}$Joules is called Planck's constant). A photon of wavelength λ thus has an energy given by $E = pc = hc/\lambda$, even though it has no rest mass.

1.6 Experimental Support for the Theory

These amazing consequences of relativity do have experimental foundations. For example, using atomic clocks and super-sonic jets, we have been able to confirm the effects of time dilation just as relativity predicts. Another experimental confirmation involves the creation of particles called muons by cosmic rays (from the sun) in the upper atmosphere. These muons then travel at very fast speeds towards the earth. In the rest frame of a muon, its life time is only about 2.2×10^{-6} seconds. Even if the muon could travel at the speed of light, it could still go only about 660 meters during its life time. Because of that, they should not be able to reach the surface of the Earth. However, it has been observed that large numbers of them do reach the Earth. From our point of view, time in the muon's frame of reference is running slowly, since the muons are traveling very fast with respect to us. So the 2.2×10^{-6} seconds are slowed down, and the muon has enough time to reach the earth.

We must also be able to explain the result from the muon's frame of reference. In its point of view, it does have only 2.2×10^{-6} seconds to live. However, the muon would say that it is the Earth which is speeding toward the muon. Therefore, the distance from the top of the atmosphere to the Earth's surface is length contracted. Thus, from the muon's point of view, it lives a very small amount of time, but it doesn't have that far to go. This is an interesting point of Relativity–the physical results (e.g. the muon reaches the Earth's surface) must be true for all observers; however, the explanation as to how it came about can be different for different frames of reference.

Another verification of special relativity is found all the time in particle physics. In particle physics, large accelerators push particles to speeds *very* close to the speed of light, and experimenters then cause those particles to strike other particles. The results of such collisions can be understood only if one uses the momentum and energy equations which were predicted by relativity (for example, one must take the total energy of the particle to be $E = \gamma mc^2$, which was predicted by relativity).

These are only a few examples that give credibility to the theory of relativity. Its predictions have turned out to be true in many cases, and to date, no evidence exists that would tend to undermine the theory in the areas where it applies.

In the above discussion of relativity's effects on space and time we have specifically mentioned length contraction and time dilation. However, there is a little more to it than that, and the next section attempts to explain this to some extent.

Chapter 2

Space-Time Diagrams

In this section we examine certain constructions known as space-time diagrams. After a short look at why we need to discuss these diagrams, I will explain what they are and what purpose they serve. Next we will construct a space-time diagram for a particular observer. Then, using the same techniques, we will construct a second diagram to represent the coordinate system for a second observer who is moving with respect to the first observer. This second diagram will show the second observer's frame of reference with respect to the first observer; however, we will also switch around the diagram to show what the first observer's frame of reference looks like with respect to the second observer. Finally, we will compare the concepts these two observers have of future and past, which will make it necessary to first discuss a diagram known as a light cone.

2.1 What are Space-Time Diagrams?

In the previous section we talked about the major consequences of special relativity, but now I want to concentrate more specifically on how relativity causes a transformation of space and time. Relativity causes a little more than can be understood by simple notions of length contraction and time dilation. It actually results in two different observers having two different space-time coordinate systems. The coordinates transform from one frame to the other through what is known as a Lorentz Transformation. Without getting deep into the math, much can be understood about such transforms by considering space-time diagrams.

2.2 Time as Another Dimension

One of the first points to make as we begin discussing space-time diagrams is that we are treating time as another dimension along with the three dimensions of space. Generally, people aren't used to thinking of time as just another dimension, but doing so allows us to truly understand how relativity works. So, how do we represent time as just another dimension?

Obviously we can't actually picture four dimensions all at once (three of space and one of time). Our minds are limited to picturing the three dimensions of space that we are used to dealing with. However, we can consider one or two dimensions of space and then use another dimension of space to represent time.

To see how this can work, consider Diagram 2-1. There you see a film strip on which each frame represents a moment in time. As you watch a film, you see each moment in time presented one right after another, and this gives the impression of seeing time pass. If we cut the film up into frames then we can stack the frames flat, evenly spaced, and one on top of the other (as shown in the diagram). Then each frame is a two dimensional representation of space and as you move through the third dimension you go up the stack, and each frame you pass represents another point in time. Thus, we have a three dimensional stack which represents two dimensions of space and the third dimension represents time. [!ht]

Note too that in the diagram the film shows a ball moving from one corner of the screen to the other. However, in the three dimensional stack, the ball now follows a three dimensional path through space-time. In four dimensional space-time, objects which we see moving in time through three dimensional space are

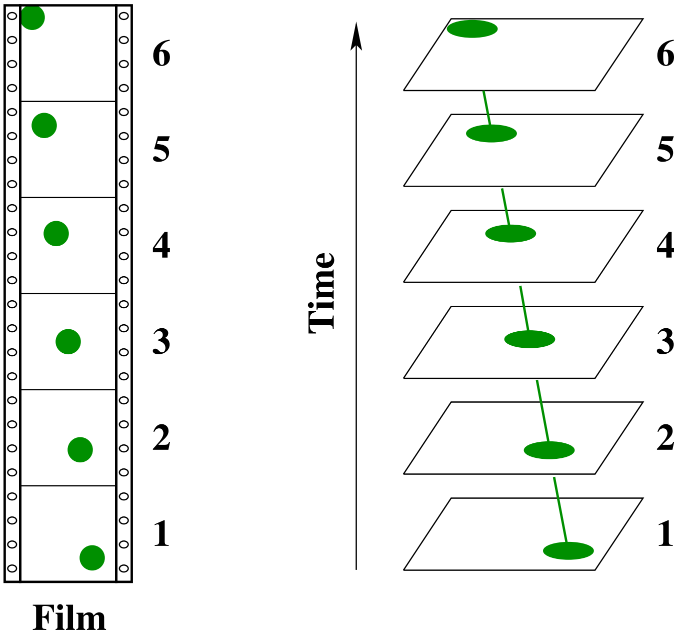

Diagram 2-1:

following a four-dimensional path through space-time. On space-time diagrams, paths you draw represent objects moving through space as time passes, but we'll see more about that later in the chapter.

Further, consider an event such as "the ball reaches the far corner of the screen." That is a single event–it occurs at one moment in time and at one particular place in space. On our diagram, it is a single point (it is a spot represented by the ball which is on the upper most frame in the stack). Any single event which occurs is represented by a single point on a space-time diagram.

And so, a space-time diagram gives us a means of representing events which occur at different locations and at different times. Every event is portrayed as a point somewhere on the space-time diagram.

Now, because of relativity, different observers which are moving relative to one another will have different coordinates for any given event. However, with space-time diagrams, we can picture these different coordinate systems on the same diagram, and this allows us to understand how they are related to one another.

2.3 Basic Information About the Diagrams we will Construct

In Diagram 2-1 we saw how one can use three dimensions to represent two dimensions of space and one of time, but for simplicity the diagrams we use will be two dimensional–one of space and one of time. We will consider the one dimension of space to be the x direction. So, the space-time diagram consists of a coordinate system with one axis to represent space (the x direction) and another to represent time. Where these two principal axes meet is the origin. This is simply a point in space that we have defined as $x = 0$ and a moment in time that we have defined as $t = 0$. In Diagram 2-2 (below) I have drawn these two axes and marked the origin with an o.

For certain reasons we want to define the units that we will use for distances and times in a very specific way. Let's define the unit for time to be the second. This means that moving one unit up the time axis will represent waiting one second of time. We then want to define the unit for distance to be a light second (the distance light travels in one second). So if you move one unit to the right on the x axis, you will be considering a point in space that is one light second away from your previous location. In Diagram 2-2, I have marked the locations of the different space and time units.

With these units, it is interesting to note how a beam of light is represented in our diagram. Consider a beam of light leaving the origin and traveling to the right. One second later, it will have traveled one light second away. Two seconds after it leaves it will have traveled two light seconds away, and so on. So a beam of light will always make a line at an angle of 45 degrees to the x and t axes. I have drawn such a light beam in Diagram 2-3. [!ht]

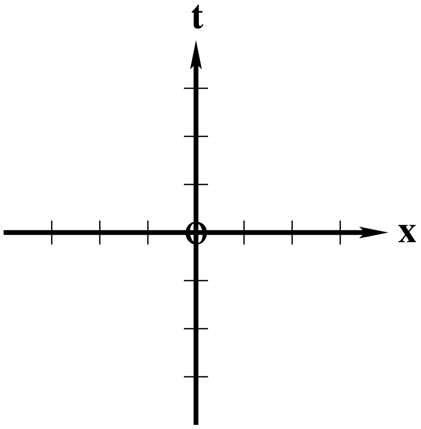

Diagram 2-2:

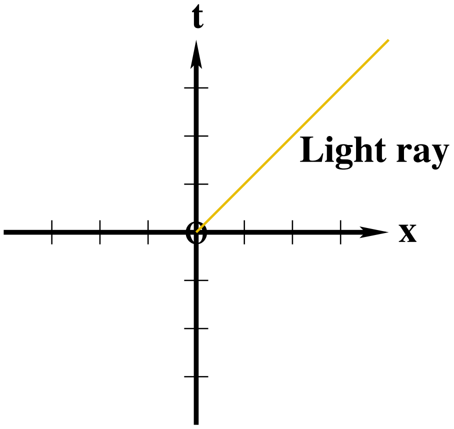

Diagram 2-3:

2.4 Constructing One for a “Stationary” Observer

At this point, we want to decide exactly how to represent events on this coordinate system for a particular observer. First note that it is convenient to think of any particular space-time diagram as being specifically drawn for one particular observer. For Diagram 2-2, that particular observer (let’s call him the O observer) is the one whose coordinate system has the vertical time axis and horizontal space axis shown in that diagram. Now, other frames of reference (which don’t follow those axes) can also be represented on this same diagram (as we will see). However, because we are used to seeing coordinate systems with horizontal and vertical axes, it is natural to think of this space-time diagram as being drawn specifically with the O observer in mind. In fact, we could say that in this space-time diagram, the O observer is considered to be “at rest”.

So if the O observer starts at the origin, then one second later he is still at $x = 0$ (because he isn’t moving in this coordinate system). Two seconds later he is still at $x = 0$, etc. If we look at the diagram, we see that this means he is always on the time axis in our representation. Similarly, any lines drawn parallel to the t axis (in this case, vertical lines) will represent lines of constant position. If a second observer is not moving with respect to the first, and this second observer starts at a position two light seconds away to the right of the first, then as time progresses he will stay on the vertical line that runs through $x = 2$.

Next we want to figure out how to represent lines of constant time. We might first find a point on our diagram that represents an event which occurs at the same time as, say, the origin ($t = 0$). To do this we will use a method that Einstein used. First we choose a point on the t axis which occurred prior to $t = 0$. Let’s use an example where this point occurs at $t = -3$ seconds. At that time we send out a beam of light in the positive x direction. If the beam bounces off of a distant mirror at $t = 0$ and heads back toward the t axis, then it will come back to the us at $t = 3$ seconds. We know this because (1) it will have traveled for three seconds away from us, (2) it will have the same distance to travel back to us in our frame of reference, and (3) according to relativity it must travel at the same speed, c, going **and** coming back. Thus, it must take three seconds to get back to us as well which means it reaches as at the time $t = 3$ seconds. So, if we send out a beam at $t = -3$ seconds and it returns at $t = 3$ seconds, then the event “it bounced off the mirror” occurred simultaneously with the time $t = 0$ at the origin.

To use this on our diagram, we first pick the two points on the t axis that mark $t = -3$ and $t = 3$ (let’s call these points A and B respectively). We then draw one light beam leaving from A in the positive x direction. Next we draw a light beam coming to B in the negative x direction. Where these two beams meet (let’s call this point C) marks the point where the original beam bounces off the mirror. Thus the event marked by C is simultaneous with $t = 0$ (the origin). A line drawn through C and o will thus be a line of constant time. All lines parallel to this line will also be lines of constant time. So any two events that lie along one of these lines truly occur at the same time in this frame of reference. I have drawn this procedure in Diagram 2-4, and you can see that the x axis is the line through both o and C which is a line of simultaneity (as one might have expected).

Note that the event marked by C is not seen by the O observer (who, remember, is represented by the t axes because he sits at $x = 0$) at the moment it happens ($t = 0$) but it is seen once light from C reaches the O observer (which is the point marked B). However, because of the way we did the experiment, we know that in this frame of reference, C truly did happen simultaneously with the origin, o. This just goes to illustrate, as discussed in Section 1.1, that when I say that two events happened simultaneously in some frame of reference, I am not talking about when they are *seen* by some observer in that frame. Rather, I am talking about when they actually occur in that frame of reference. On our diagrams, events are represented at their actual space-time locations relative to one another, and in a particular frame of reference that means that we show exactly when and where the event occurred (not "observed" but truly occurred) in that frame.

Now, by constructing a set of simultaneous time lines and constant position lines we will have a grid on our space-time diagram. Any event has a specific location on the grid which tells where and when it occurs in this frame of reference. In Diagram 2-5 I have drawn one of these grids and marked an event (@) that occurred 3 light seconds away to the left of the origin ($x = -3$) and 1 second before the origin ($t = -1$). [!ht]

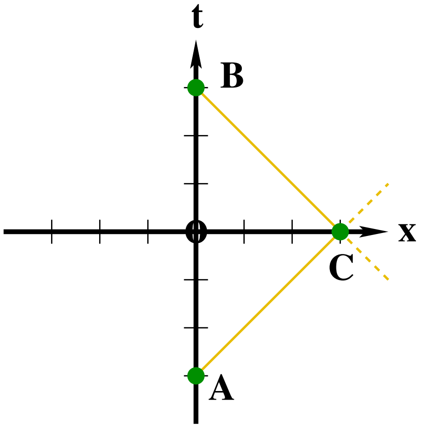

Diagram 2-4:

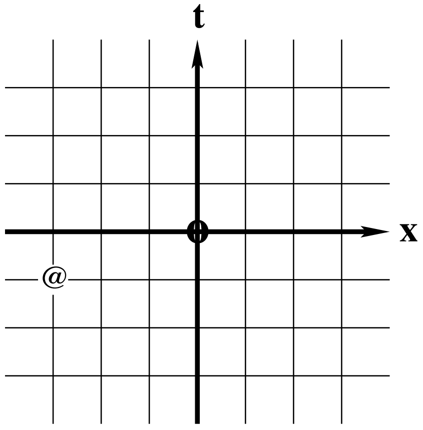

Diagram 2-5:

2.5 Constructing One for a "Moving" Observer

Now comes an important addition to our discussion of space-time diagrams. The coordinate system we have drawn will work fine for any observer who is not moving with respect to the O observer. Now we want to construct a coordinate system for an observer who IS traveling with respect to the O observer. The trajectories of two such observers have been drawn in Diagram 2-6 and Diagram 2-7. Notice that in our discussion we will usually consider moving observers who pass by the O observer at the time $t = 0$ and at the position $x = 0$. Thus, the origin will mark the event "the two observers pass by one another".

Now, the traveler in Diagram 2-6 is moving slower than the one in Diagram 2-7. You can see this because in a given amount of time (distance along the t axis), the Diagram 2-7 traveler has moved further away from the time axis than the Diagram 2-6 traveler. So the faster a traveler moves, the more slanted this line becomes. [!ht]

What does this line actually represent? Well, remember that the line marks the position of our observer at different times on our diagram. But, also, consider an object sitting right next to our moving observer. If a few seconds later the object is still sitting right next to him (practically on that line), then, in his point of view, the object has not moved. So, the line is a line of constant position for the moving observer. Nothing on that line is moving with respect to him. But that means that this line represents the same thing for the moving observer as the t axis represented for the O observer; and in fact, this line becomes the moving

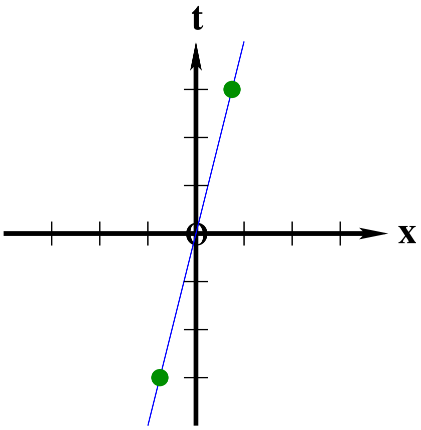

Diagram 2-6:

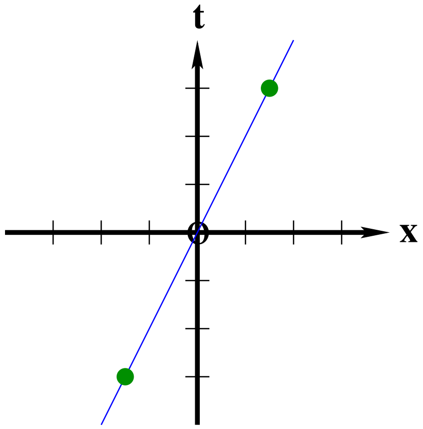

Diagram 2-7:

observer's new time axis. We will mark this new time axis as t' (t-prime). All lines parallel to this slanted line will also be lines of constant position for our moving observer.

Now, just as we did for the O observer, we want to construct lines of constant time for our traveling observer. To do this, we will use the same method that we did for the O observer. The moving observer will send out a light beam at some time $t' = -T$, and the beam will bounce off some mirror so that it returns to him at time $t' = +T$. Now remember, light travels at the same speed in any direction for **all** observers, so our traveling observer must conclude that the light beam took the same amount of time traveling out as it did coming back in his frame of reference. If in his frame the light left at $t' = -T$ and returned at $t' = +T$, then the point at which the beam bounces off the mirror must have occurred simultaneously with the origin, where $t' = t = 0$, in the frame of reference of our moving observer.

There is a very important point to note here. What if instead of light, we wanted to throw a ball at 0.5 c, have it bounce off some wall, and then return at the same speed (0.5 c). The problem with this is that to find a line of constant time for the moving observer, the ball must travel at 0.5 c **both ways** in the reference frame of the **moving** observer. But we have not yet defined the coordinate system for the moving observer, so we do not know what a ball moving at 0.5 c with respect to him will look like on our diagram. However, because of relativity, we know that the speed of light itself **cannot** change from one observer to the next. In that case, a beam of light traveling at c in the frame of the moving observer will also be traveling at c for the O observer. So, a line which makes a 45 degree angle with respect to the x and t axes will **always** represent a beam of light traveling at speed c for **any** observer in **any** frame of reference.

In Diagram 2-8, I have labeled a point A' on the t' axes which occurs some amount of time before $t' = 0$ and a point B' which occurs the same amount of time after $t' = 0$. I then drew the two light rays (remember, these are "45 degree angle" lines) as before–one leaving from A and going to the right, and one moving to the left and coming in to B. I then found the point where they would meet (C') which marks the point where the ray from A' would have had to bounce in order to get back to the moving observer at B'. Thus, C' and o occur at the same time in the frame of the moving observer. Notice that for the O observer, C' is above his line of simultaneity at o (the x axis). So while the moving O' observer says that C' occurs when the two observers pass (at the origin), the O observer says that C' occurs after the two observers have passed by one another. We will further discuss this difference in the concepts of future and past in Section 2.8.

In Diagram 2-9, I have drawn a line passing through C' and o. This line represents the same thing for our moving observer as the x axis did for the O observer. So we label this line x'. [!ht]

From the geometry involved in finding this x' axis, we can state a general rule for finding the x' axis for any moving observer. First recall that the t' axis is the line that represents the moving observer's position on the space-time diagram. The faster O' is moving with respect to O, the greater the angle between the t axis and the t' axis. So the t' axis is rotated away from the t axis at some angle (either clockwise or counterclockwise, depending on the direction O' is going–right or left). The x' axis is then a line rotated at the same angle away from the x axis, but in the *opposite* direction (counterclockwise or clockwise).

Now, x' is a line of constant time for O', and any line drawn parallel to x' is also a line of constant time.

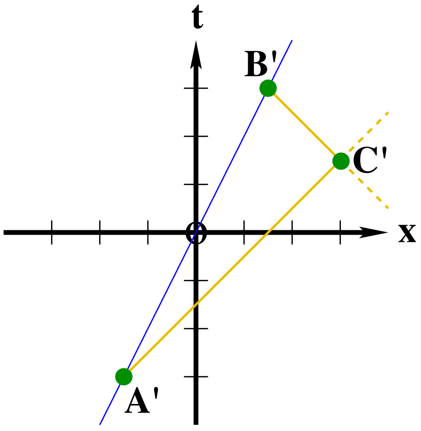

Diagram 2-8:

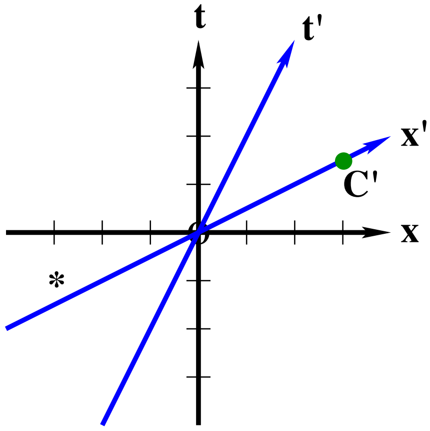

Diagram 2-9:

Such lines, along with the lines of constant position, form a grid of the space-time coordinates for the O' observer. I have tried my best to draw such a grid in Diagram 2-10. If you look at that diagram, you can see the skewed squares of the coordinate grid. You can see that if you pick a point on the space-time diagram, the two observers with their two different coordinate systems will disagree on when and where the event occurs. [!ht]

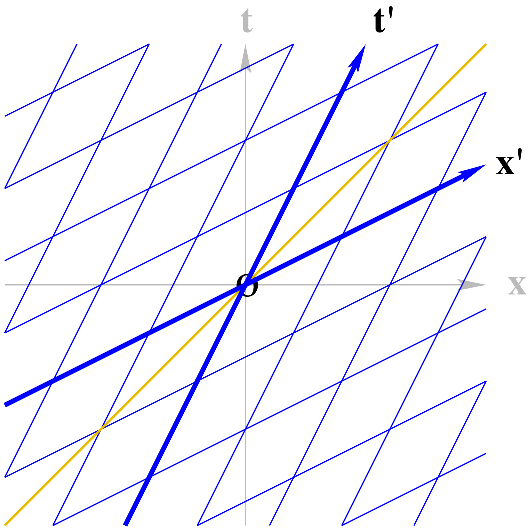

Diagram 2-10:

As a final note about this procedure, think back to what really made these two coordinate systems look differently. Well, the only thing we assumed in creating these systems is that the speed of light is the same for all observers. In fact, this is the only reason that the two coordinate systems look the way they do.

2.6 A Quick Comparison of the two Observers

For a moment, I want to go back and compare the two observers in Diagram 2-8. Consider how the O observer would explain the experiment done by the O' observer. First note that in the coordinate system used by the O observer, the point marked C' is above the x axis. This means that in the O observer's frame of reference, C' happens after the origin (when the two observers pass by one another). However, we concluded that for O' the C' event happens at the same time as the two observers are passing one another.

What does that mean?

Look at the parts of the experiment O' did (including the actions of O' and the events A', B', and C') as they appear in the O observer's frame. In that frame, O' sends out a light signal when his own clock reads $t' = -T$, but note also that he is moving along with that signal (according to O). The distance between them changes slowly at the beginning according to O because O' is moving along with the signal in the same direction. Then, according to O, the two observers pass by one another. Next, the C' event happens and the light bounces back toward the two observers. In the frame of the O observer, the O' observer is now racing towards the light beam, and so the distance between them is changing very quickly. Finally, the light beam reaches O' as his clock is ticking $t' = +T$.

So, we see that in the O frame of reference, because O' is moving along with the light before C' and is moving towards the light after C' that means C' has to happen after the "half way point" (when the two observers pass one another).

However, relativity says that O' cannot agree with that analysis. In the frame of O', it is the O observer who is moving. Further, O' cannot agree that the distance between him and the light is changing slowly before C' and quickly after C'. Why can't he agree? Well, because then he would measure the speed of the light in his frame of reference and find it to be different going away from him than it is coming back to him. As discussed in Section 1.2, relativity dictates that for **any** inertial observer, when he measures the speed of light he **must** find the speed to be c–**always**, and in **all** directions. If O' has to find that the light is traveling at the same speed going and coming back, then O' also has to conclude that in his frame C' really, truly happens at the same time as the origin (when and where the two observers pass one another). O' thus has a different coordinate system than O, and he measure space and time differently.

And so, in one frame of reference C' really, truly happens after the two observers pass one another, but in another frame of reference C' really, truly happens and the same time the two observer's pass. We find that the notion of simultaneity is relative, and we will discuss this further in just a bit.

Next, though, I want to address a possibility you might be thinking right now. That is, why can't it simply be that O' is just wrong in interpreting things as he does and that O is correct. One might want to claim that the reason O' is confused is that he is moving while O is not. But next we will see that we can interchange the two observers, and it becomes obvious that there is no absolute way to claim that one of them is the "correct" observer.

2.7 Interchanging "Stationary" and "Moving"

In our understanding of space-time diagrams, we need to incorporate the idea that all reference frames that are not accelerating are considered equivalent and that all motion is relative. By this I mean that O was considered as the stationary observer only because we defined him as such. Remember? We said that this it is natural to think of the diagram being drawn specifically for the observer whose coordinate system is drawn with vertical and horizontal axes. We then said that we can think of that observer (O) to be considered "at rest" in this diagram. Then, when I called O' the moving observer, I meant that he was moving with respect to O.

However, we should just as easily be able to define O' as the stationary observer. Then, to him, O is moving away from him to the left. Then, we should be able to draw the t' and x' axes as the vertical and horizontal lines, while the t and x axes become the rotated lines. I have done this in Diagram 2-11. By examining this diagram, you can confirm that it makes sense to you in light of our discussion thus far. (For example, picture grabbing the x' and t' axes in Diagram 2-9 and rotating them around the origin until they are horizontal and vertical lines. If x and t follow your rotation, then you can see how they would end up as they are drawn in Diagram 2-11.) [!ht]

I have also included in Diagram 2-11 the experiment that O' did in which he decided how to draw the x' axis, and you can see that it now looks just like the experiment O did when his x and t axes were the horizontal and vertical lines. Further in Diagram 2-11 you can see that the experiment done by the O observer now looks like the one which has incorrectly concluded that C occurs at the same time the two observers are passing one another.

Thus, you can see that we can completely interchange the concept of which observer is moving and which is sitting still, and as a result we must conclude that neither frame of reference is any "better" than the

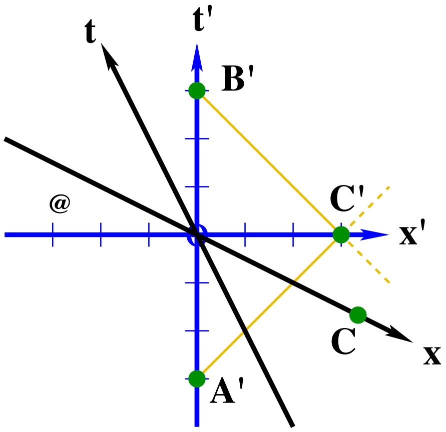

Diagram 2-11:

other. When O concludes that C occurs simultaneous with o, he is **really**, **truly** correct for his frame of reference. Also, when O' concludes that it is C' which occurs simultaneous with o, he is also **really**, **truly** correct for his frame of reference. The notion of simultaneity is not absolute, but **really**, **truly** depends on your frame of reference. To understand why this doesn't cause contradictions, we go to the next section in which we discuss the notion of future and past with relativity in mind.

2.8 "Future", "Past", and the Light Cone

For the later FTL discussions, it will be important to understand the way different observers have different notions concerning the future and the past. This difference comes about because of the way the different coordinate systems of the two observers compare to one another.

First, let me note that with what we have discussed we cannot make a complete comparison of the two observers' coordinate systems. You see, we have not seen how the lengths which represents one unit of space and time in the reference frame of O compare with the lengths representing the same units in O'. This will be covered in the Part II: More on Special Relativity (which is "optional" for those of you just interested in the faster than light discussions). We can, however, compare the observers' notions of future and past.

Back on Diagram 2-9, in addition to the O and O' space and time axes, I also marked a particular event with a star, "*". Recall that for O, any event on the x axis occurs at the same time as the origin (the place and time that the two observers pass each other). Since the marked event appears under the x axis, then O must find that the event occurs before the observers pass each other in his frame. Also recall that for O', those events on the x' axis are the ones that occur at the same time the observers are passing. Since the marked event appears above the x' axis, O' must find that the event occurs after the observers pass each other in his frame. So, when and where events occur with respect to other events is completely dependent on ones frame of reference. Note that this is not a question of when the events are seen to happen in different frames of reference, but it is a question of when they *really do* happen in the different frames (recall our discussion of reference frames in Section 1.1). So, how can this make sense? How can one event be both in the future for one observer and in the past for another observer. To better understand why this situation doesn't contradict itself, we need to look at one other construction typically shown on a space-time diagram.

In Diagram 2-12 I have drawn two light rays, one which travels in the +x direction and another which travels in the -x direction. At some negative time, the two rays were headed towards $x = 0$. At $t = 0$, the two rays finally get to $x = 0$ and cross paths (at the origin). As time progresses, the two then speed away from $x = 0$. This construction is known as a light cone. [!ht]

A light cone divides a space-time diagram into two major sections: the area inside the cone and the area outside the cone (as shown in Diagram 2-12). (Let me mention here that I will specifically call the cone I have drawn "a light cone centered at the origin", because that is where the two beams meet.) Now, consider an observer who has been sitting at $x = 0$ (like our O observer) and is receiving and sending signals at the

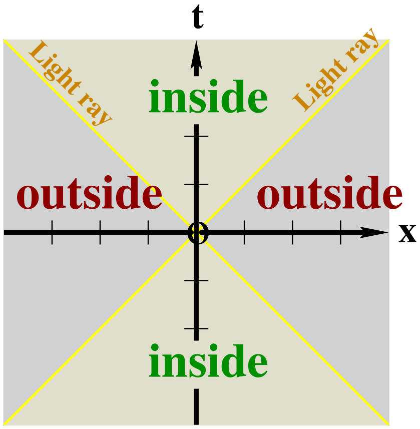

Diagram 2-12:

moment marked by $x = 0, t = 0$ (at the origin). Obviously, if he sends out a signal, it proceeds away from $x = 0$ into the future, and the event marked by someone receiving the signal would be above the x axis (in his future). Also, if he is receiving signals at $t = 0$, then the event marked by someone sending the signal would have to be under the x axis (in his past). Now, if it is impossible for anything to travel faster than light, then the only events occurring before $t = 0$ that the observer can know about at the moment are those that are inside the light cone. Also, the only future events (those occurring after $t = 0$) that he can influence are, again, those inside the light cone.

Now, one of the most important things to note about a light cone is that its position is the same for all observers (because the speed of light is the same for all observers). For example, picture taking the skewed coordinate system of the moving observer and superimposing it on the light cone I have drawn (note: a diagram which shows this view will be given in Part II: More on Special Relativity). If you were to move one unit “down” the x’ axis (a distance that represents one light second for our moving observer), and you move one unit “up” the t' axes (one second for our moving observer), then the point you end up at should lie somewhere on the light cone. In effect, a light cone will always look the same on our diagram regardless of which observer is drawing the cone.

This fact has great importance. Consider different observers who are all passing by one another at some point in space and time. In general, they will disagree with each other on when and where different events had and will occur. However, if you draw a light cone centered at the point where they are passing each other, then they will ALL agree as to which events are inside the light cone and which events are outside the light cone. So, regardless of the coordinate system for any of these observers, the following facts remain: The only events that any of these observers can ever hope to influence are those which lie inside the upper half of the light cone. Similarly, the only events that any of these observers can know about as they pass by one another are those which lie inside the lower half of the cone. Since the light cone is the same for all the observers, then they all agree as to which events can be known about as they are passing and which can be influenced at some point after they pass.

Now let’s apply this to the observers and event in Diagram 2-9. As you can see, the marked event is indeed outside the light cone. Because of this, even though the event is in one observer’s past at the time in question ($t = t' = 0$), he cannot know about the event at the time. Also, even though the event is in the other observer’s future at the time, he can never have an effect on the event after. In essence, the event (when it happens, where it happens, how it happens, etc.) is of absolutely no consequence for these two observers at the time in question. As it turns out, anytime you find two observers who are passing by one another and an event which one observer’s coordinate system places in the past and the other observer’s coordinate system places in the future, then the event will always be outside of the light cone centered at the point where the observers pass.

But doesn’t this relativistic picture of the universe still present an ambiguity in the concepts of past and future? Perhaps philosophically it does, but not physically. You see, the only time you can see these ambiguities is when you are looking at the whole space-time picture at once. If you were one of the observers who is actually viewing space and time, then as the other observer passes by you, your whole picture of space and time can only be constructed from events that are inside the lower half of the light cone. If you wait for a while, then eventually you can get all of the information from all of the events that were happening around the time you were passing the other observer. From this information, you can draw the whole space-time diagram, and then you can see the ambiguity. But by that time, the ambiguity that you are considering no longer exists. So the ambiguity can never actually play a part in any physical situation. Finally, remember that this is only true if nothing can travel faster than the speed of light.

Well, that concludes our introduction to special relativity and space-time diagrams. The next section deals with these concepts with more detail; however, if the reader wishes to skip to the FTL discussion, the information provided in the above sections should be enough to follow that discussion.

Part II

More on Special Relativity

This is Part II of the "Relativity and FTL Travel" FAQ. It is an "optional reading" part of the FAQ in that the FTL discussion in Part II does not assume that the reader has read the information discussed below. If your only interest in this FAQ is the consideration of FTL travel with relativity in mind, then you may only want to read Part I: Special Relativity and Part IV: Faster Than Light Travel–Concepts and Their "Problems".

In this part, we look more deeply into some points of special relativity. By completing our discussion on the space time diagram as well as explaining some of the paradoxes involved with SR, it should give the reader a better understanding of the theory.

For more information about this FAQ (including copyright information and a table of contents for all parts of the FAQ), see the Introduction to the FAQ portion.

Chapter 3

Completing the Space-Time Diagram Discussion

Here we will complete the discussion of the space-time diagrams which we began in the previous chapter. We will do this by completely comparing the coordinates our observers have for a particular event. To make that comparison we will need to see how the lengths which represent one unit of space and time in the reference frame of O compare with the lengths representing the same units in O'. The easiest way for us to do this is to use information we have already seen–in particular, we use the fact that a clock moving with respect to an observer seems to be running slowly to that observer and a pole moving with respect to that observer seems to be shorter to that observer by a factor of γ. (Note: this was explained in Chapter 1. of this FAQ.) Understanding how to use this in the space-time diagram in order to completely construct the two observers' coordinate systems should give some solid insight into time dilation and length contraction in special relativity.

3.1 Comparing Time for O and O'

So, how do we show time dilation on our space-time diagram. Well, the key to this can be found by expressing time dilation in the following way: In the O observer's frame of reference, let the tick t_1 of his clock be simultaneous with the tick t'_1 of the O' observer's clock. Also, let the tick t_2 of his (the O observer's) clock be simultaneous with the t'_2 tick of the O' observer's clock. Then, we would find that

$$t'_2 - t'_1 = \frac{t_2 - t_1}{\gamma} \tag{3:1}$$

where gamma (γ) (as defined in Section 1.4) would be calculated using the relative velocity of O and O'. What Equation 3:1 says is that in the O observer's frame of reference, the difference in the ticks of the O' observer's clock is smaller than the difference in the O observer's own ticks by a factor of γ. Thus, we see that in the frame of O, the O' observer's clock is running slowly.

As an example, from here on we will consider the case where the relative velocity is $0.6c$ such that $\gamma = 1.25$. Using an example like this will make the procedure easier to understand for the reader; however, remember that we could redo this whole process with any speed (calculating a new γ factor, drawing a different speed for the observers, drawing appropriate lines of simultaneity, etc.).

Now, what if we let the t_1 tick be the "zero" tick. That means that at the origin, when both of our observers are right next to one another, $t_1 = t'_1 = 0$. So, both of the observers agree (because there is no separation between them in space at the origin) that t_1 and t'_1 are simultaneous, and happen at $t = t' = 0$. However, after some time, there will be a tick (t_2) on the O observer's clock. In the frame of reference of O, that tick is simultaneous with the tick t'_2 of the O' observer's clock. Since $t_1 = t'_1 = 0$, and we are using $\gamma = 1.25$, we know (from Equation 3:1) that

$$\begin{aligned} t'_2 - 0 &= \tfrac{t_2 - 0}{1.25} \\ \text{so:} & \\ t'_2 &= 0.8 t_2 \end{aligned} \tag{3:2}$$

So, this says that in the frame of the O observer, the tick t_2 of his clock is simultaneous with the tick $0.8t_2$ on the O' observer's clock. If we draw a line of simultaneity in the O observer's frame of reference such that it goes through the tick t_2 of his clock, then it must also go through the tick $0.8t_2$ of the O' observer's clock. If we let $t_2 = 1$ second, then we get what is shown in Diagram 3-1. The distance from the origin, o, to the first mark along t in that diagram is defined to be 1 second for our O observer. Meanwhile, the distance from o to the "*" symbol along t' in that diagram is 0.8 second **for the O' observer**. So, we begin to see that we can relate distances in time along the axes of the different observers. [!ht]

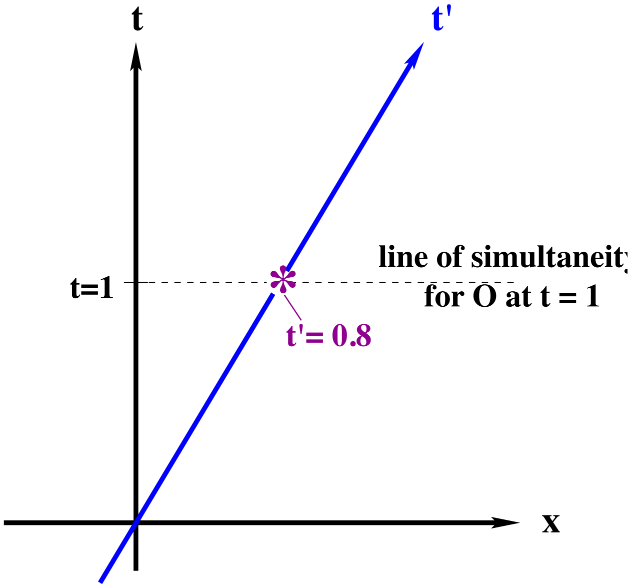

Diagram 3-1:

This puts us on our way to understanding how, for example, different lengths along t and t' relate to particular times on the clocks of the two observers. Our next step to understanding this better will be to look at the situation from the O' observer's frame of reference.

We have found what tick of the O' observer's clock is simultaneous with the $t = 1$ tick of the O observer's clock in the O observer's frame of reference. However, say we want to decide what t' tick is simultaneous with the O observer's $t = 1$ tick in the O' observer's frame of reference (remember, the line of simultaneity in Diagram 3-1 is only valid for the O observer's frame). To figure this out, we need to draw a line of simultaneity in the O' observer's frame of reference which passes through the event "the O observer's clock ticks 1". When we do this, we want to note where that line passes the t' axis, because that mark points out the tick on the O' observer's clock which is simultaneous with O observer's $t = 1$ tick in the O' observer's frame of reference. I have drawn this line in Diagram 3-2, but I have also left everything that was in Diagram 3-1. [!ht]

Now, note that I marked the "%" symbol in that diagram (where the line of simultaneity for O'–which goes through $t = 1$–crosses the t' axes) as the event $t' = 1.25$. But how did I know that? Well, because in the frame of reference of O', it is the O observer who is moving at $0.6c$, and thus it is the O observer who's clocks are running slowly by a factor of 1.25. So, in the frame of O', the event "$t = 1$ at the O observer's position" must be simultaneous with the event "$t' = 1.25$ at the O' observer's position." That way, in the O' observer's frame, it will be the O observer's clock which is running slowly by a factor of 1.25. In addition, I could use the length from the origin to "*" (which I know is 0.8 seconds for the O' observer) to figure out how much time passes between the origin and the "%" symbol for O'. Either way, I find the same thing.

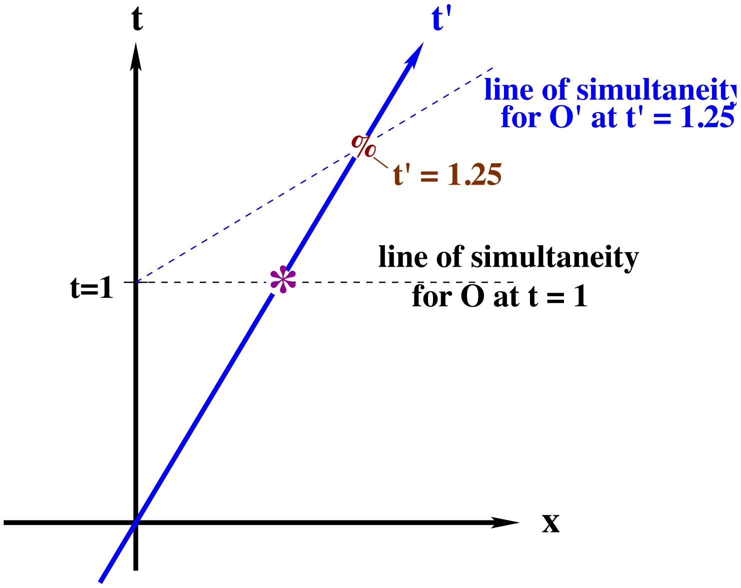

Diagram 3-2:

In Diagram 3-2, one can begin to see the power of using space-time diagrams to understand special relativity. Note that from one glance at that diagram not only can we see that in the O observer's frame of reference the O' observer's clock is running slow by a factor of 1.25 (i.e. the event "$t = 1$" is simultaneous with the event "$t' = 0.8$" in the O observer's frame) but we also see that in the O' observer's frame it is the O observer's clock which is running slow by a factor of 1.25 (i.e. the event "$t = 1$" is simultaneous with the event "$t' = 1.25$" in the O' observer's frame). Thus, we can see at once on this diagram that in each observer's own frame, the other observer's clock is running slow. This happens to be one of the first, key points to understanding the twin paradox (which will be discussed fully in the next section).

3.2 Comparing Space for O and O'

So, we have found a correlation between the lengths which represent certain times along the t axis for O and the lengths which represent certain times along the t' axis for O'. We did this by using (1) the idea of time dilation which was found earlier to be caused by the fact that light always travels at c for all inertial observers and (2) the lines of simultaneity for different observers which we learned how to draw by also using the fact that light always travels at c for all inertial observers. Similarly, we can find a correlation between lengths which represent certain distances along the x axis for O and the lengths which represent certain distances along the x' axis for O'. As an example, I have drawn a comparison of distances in Diagram 3-3 which will be explained below. [!ht]

Perhaps the best way to explain this diagram is as follows: Consider a rod being held by the O observer such that one end of the rod follows the t axis (and is thus always next to the O observer) while the other end follows the vertical line drawn at $x = 1$. The rod then is obviously stationary in the O observer's frame of reference. Second consider a rod being held by the O' observer such that one end follows the t' axis and the other end follows the line of constant position for O' which I have drawn.

Well, in the O observer's frame, his rod is obviously 1 light-second long. But notice that in his frame the ends of the O' observer's rod are next to the ends of the O observer's rod at $t = 0$. Thus, in the O observer's frame, the O' observer's rod is also 1 light-second long. But length contraction tells us that in the O observer's frame, the O' observer's rod is shorter than its "rest length" by a factor of 1.25. Thus, in the O' observer's frame (the frame in which his rod is at rest), his rod must actually be 1.25 light-seconds long. That is how I know that the line of constant position for O' I drew was for $x' = 1.25$.

Now, look at the distance along x' from the origin (o) to the point marked "#". That distance represents

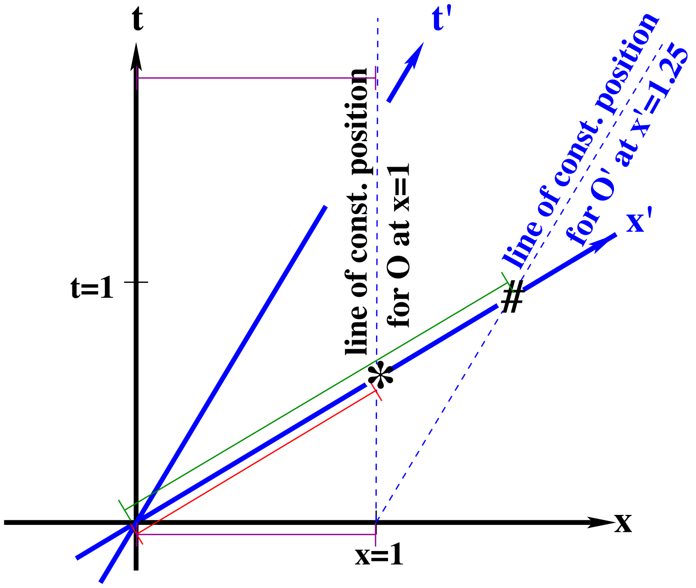

Diagram 3-3:

the length of the O' observer's rod from his own frame of reference (i.e. 1.25 light-seconds). Also, the distance along x' from the origin to the point marked "*" represents the length of the O observer's rod in the O' observer's frame of reference. That distance must be 0.8 because in the O' frame, it is O and his rod which are moving, and thus his rod seems length contracted by a factor of 1.25 from its length in the frame of reference in which it is at rest (the O frame). That number could have also been found by using the fact that the distance from o to "#" was 1.25 light-seconds.

Finally, we again note the power of the space-time diagram. At one glance of Diagram 3-3 we are able to see that in the O' observer's frame, his rod is 1.25 light-seconds long, while in the O observer's frame it is only 1 light-second long. At the same time we are able to see that in the O observer's frame, his rod is 1 light-second long, while in the O' observer's frame, it is only 0.8 light-seconds long. Thus, each observer believes that the other observer's rod is shorter than it is in the frame of reference in which the rod is at rest. They each believe that the other is experiencing length contraction, and with a space-time diagram, we are able to see how that is so.

3.3 Once Again: The Light Cone

Here I want to demonstrate how a light cone appears in the two coordinate systems. In Section 2.8 I mentioned that the light cone is drawn exactly the same for the two observers. Now that we understand how to draw the two coordinate systems completely (i.e. we can now draw "tick" marks on the x' and the t' axes as well as the x and t axes because of the discussion above) we can make a diagram which clearly shows this. To start, in Diagram 3-4 I have shown the results of our discussion above in that I have indicated where the tick marks would appear on the x' and t' axes. [!ht]

Next, in Diagram 3-5 I have drawn the x and t axes along with lines of simultaneity and lines of constant position (for O) at each tick mark. In addition, the upper half of a light cone centered at the origin is shown. As you see (and as we would expect), it passes through the points $x = 1$ light-second, $t = 1$ second; $x = 2$ light-seconds, $t = 2$ seconds; etc. [!ht]

Continuing with the diagrams, Diagram 3-6 shows the x' and t' axes along with lines of simultaneity and lines of constant position (for O') at each tick mark along those axes. Again, the upper half of a light cone centered at the origin is also shown. As you see (and as we would again expect), it passes through the points $x' = 1$ light-second, $t' = 1$ second; etc. Note that the light cone itself is drawn exactly the same as it is in Diagram 3-5. [!ht]

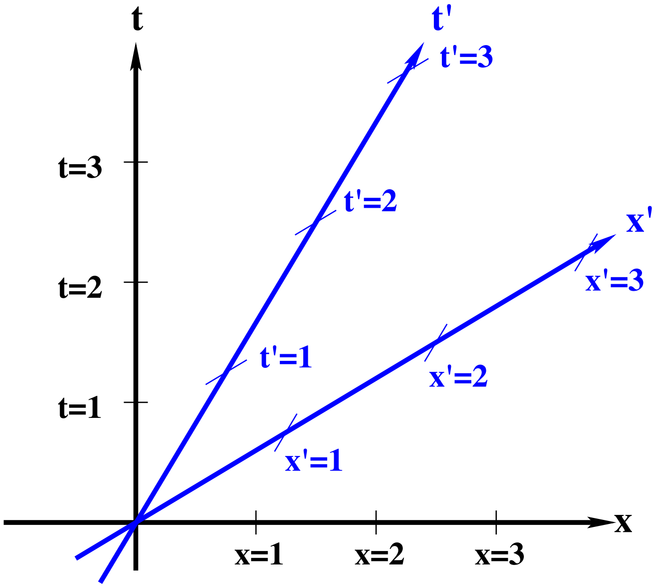

Diagram 3-4:

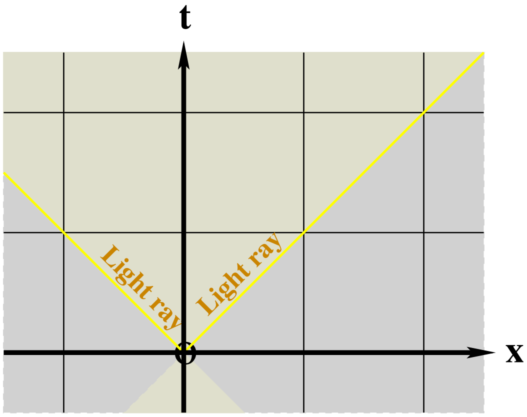

Diagram 3-5:

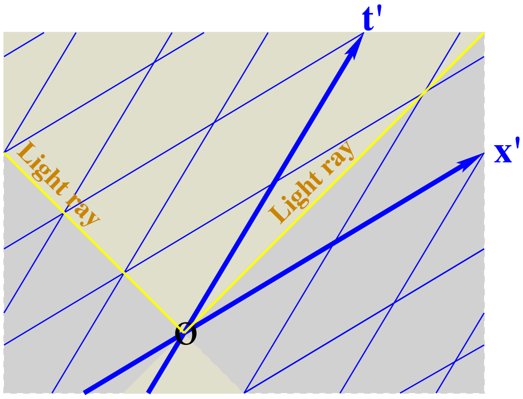

Diagram 3-6:

Finally, I want to superimpose Diagram 3-5 and Diagram 3-6 to some extent onto Diagram 3-7. It would be quite cluttered to put all the lines included in the two diagrams, but I want to include the lines which make up $x = 1$, $t = 1$, $x' = 1$, and $t' = 1$. These lines are thus drawn on Diagram 3-7, but they terminate where they meet the light cone which is also shown. You should begin to see the relationship between the two different frames of reference and the fact that the light cone itself is exactly the same in both coordinate systems. This is a direct result from the fact that every step we took in producing these diagrams used the assumption that the speed of light is the same in all inertial frames of reference. [!ht]

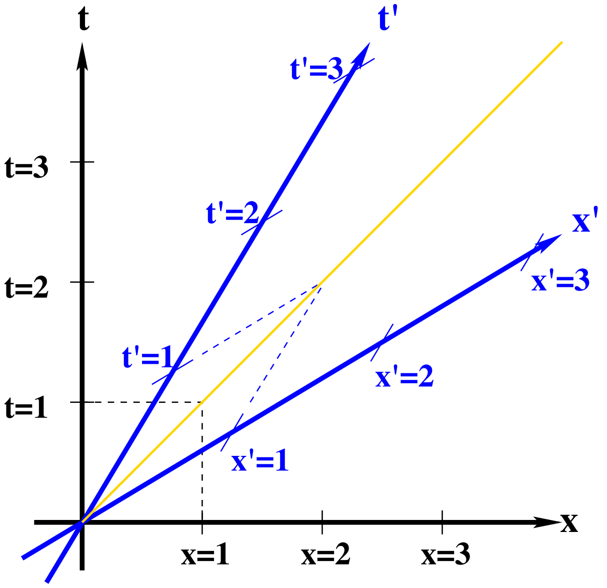

Diagram 3-7:

Though this concludes our discussion of space-time diagrams, we will continue to see them in the next section, because they can be vital tools for understanding paradoxes in special relativity.

Chapter 4

Paradoxes and Solutions

One misleading statement many people hear in connection with relativity is something like this: "Time moves slower for you as your speed increases." It is misleading because it implies some incorrect concepts. It implies that there is an **absolute** way to decide whether or not someone is truly at rest or moving (at a constant, non-zero velocity) when in reality this depends on your frame of reference. It implies that if you are moving at a constant velocity, then your clock is moving slower than some sort of "correct" clock which is truly not in motion. It also implies that you yourself might find your clock ticking slower than usual.

However, as I have mentioned earlier, motion is relative. There is no way to say that one object is truly at rest and another is truly moving at a constant velocity. You can only say that one object is moving at a constant velocity **relative to** another object. You can say that in the frame of reference of one observer (call him Joe) another observer (call her Jane) is moving at a constant velocity. Then, in Joe's frame of reference, Jane's clock is running slowly, and she is length contracted in the direction of her motion. However, in Jane's frame of reference, **Joe** is the one who is moving at a constant velocity relative to her. Because the laws of physics are the same for all inertial frames, we must be able to apply the same laws to Jane as we just applied to Joe. Thus, in Jane's frame, Joe's clock is the one which is running slowly, and Joe is length contracted in the direction of his motion.

This leads one to question whether or not relativity contradicts itself. If all motion is relative, we have concluded that each observer believes that the other observer's clock is running slowly, and each believes that the other observer is length contracted in the direction of motion. Isn't that a contradiction? For example, how can Jane's clock be running slower than Joe's **and** Joe's clock be running slower than Jane's? Well, these questions lead to various solvable paradoxes in special relativity.

As a note, the word "paradox" has a few different meanings, and when I use it here, I will have this meaning in mine: "a paradox is a statement that seems contradictory or absurd but that may in fact make sense." A "solvable paradox" is then a paradox that does in fact make sense when explained correctly, while an "unsolvable paradox" is a paradox for which the statement "may in fact make sense" doesn't hold (i.e. an unsolvable paradox is truly self-contradictory).

The paradoxes in special relativity are solvable, and below I will present two of these paradoxes along with their solutions.

4.1 The "Twin Paradox"

The twin paradox deals with the question of "who's clock is running slower?" The story goes as followers: Two twins (say Sam and Ed) are both on Earth when one of them (say Sam) decides to leave the Earth by very quickly accelerating to a speed close to the speed of light. We then consider the two frames of reference after Sam has reached a constant velocity. According to special relativity, in Ed's frame of reference, Sam's clock is running slowly, while in Sam's frame of reference, it is Ed's clock which is running slowly.

Now, as long as the two are apart, it is not to hard to argue that the question is strictly dependent on your point of view. By this I mean that we can argue that there is no correct answer to the questions–that who's clock is running slower depends completely on what frame of reference you are in. However, how would we continue this argument if we added the following to the story:

At some point after Sam begins his trip away from the Earth, one of the twins decides to go meet with the other twin. Either Ed decides to accelerate away from the Earth and catch up to Sam, or Sam decides to accelerate back towards the Earth to go back and meet with Ed. We then ask this question: when the two twins are standing next to one another again, which one is older?

With the above addition to the story, there must be a definite answer to the final question. So, how can we continue to say that the answer depends on your frame of reference? Well, as we will see, the final question does have a definite answer, but the question of how this came about **is** dependent on who you ask.

4.1.1 Viewing it with a Space-Time Diagram

So, now we will try to understand the twin paradox by using our old friend, the space-time diagram. To do this, we have to decide on some specifics. First, we will say that the relative motion of Sam and Ed is $0.6c$. So, after Sam has accelerated to a constant speed, he will be traveling at $0.6c$ with respect to the Ed. (Of course, in Sam's frame, it is Ed who is moving at a speed of $0.6c$ away from Sam.) Next, we need to decide who will be the one who eventually accelerates to go and meet with the other twin. In our case, we will look at the situation where Sam turns around to go back and meet with Ed. Finally, I should mention that the accelerations we will be using will be "instantaneous" accelerations. This means that they take no time to accomplish. In the real world, it would (of course) take time to accelerate, and while this would make the space-time diagrams look differently, the basic ideas we will discuss still hold.

Now we look at the space-time diagrams. In Diagrams 4-1 and 4-2 below, I have drawn the whole trip in two parts. In Diagram 4-1, you see Sam headed away from Ed, and in Diagram 4-2, you see Sam after he has turned around and is headed back to Ed. [!ht]

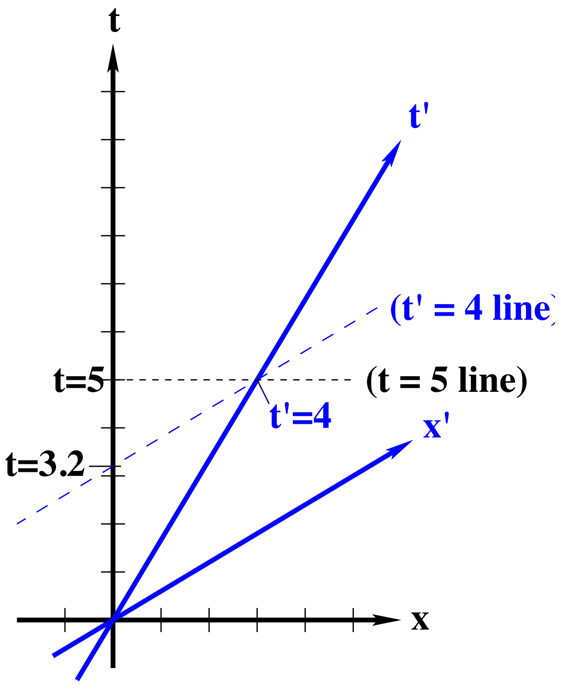

Diagram 4-1:

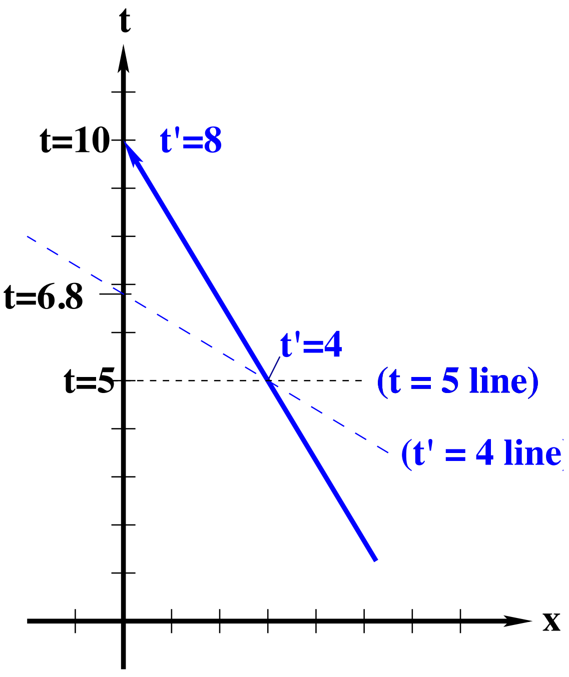

Diagram 4-2:

Now, to explain the diagrams: Ed (the twin on Earth) is represented by the x and t axes while x' and t' denote the coordinate system for Sam. Sam's motion through space-time is represented by the blue line marked t', as usual. Now, at the origin, Sam instantaneously accelerates to the speed of $0.6c$. He then proceeds away from Ed until Sam sees that his own clock read 4 years (just to pick some unit of time–which means that the distances would be in light-years). When Sam sees his own clock read 4, he turns around with an instantaneous acceleration. At that point, we switch to Diagram 4-2. In that diagram, Sam heads back to Ed.

4.1.2 Explaining the “First Part”

Now let’s concentrate on the first of the two diagrams. Just before Sam turns around, his clock reads 4 years. At that point I have drawn two lines of constant time (or lines of simultaneity)–one for each observer. The line parallel to the x axes is (of course) the line of simultaneity for Ed which passes through the event “Sam’s clock reads 4 years”. Note that this line of simultaneity for Ed also passes through the event “Ed’s clock reads 5 years”. Therefore, in Ed’s frame of reference, the events “Sam’s clock reads 4 years” and “Ed’s clock reads 5 years” are simultaneous events. This diagram thus explains how in Ed’s frame of reference, Sam’s clock is running slower than Ed’s by a factor of 0.8 (that’s one over γ when $v = 0.6c$).

However, the line of simultaneity we were looking at is not a line of simultaneity for Sam. Sam’s line of simultaneity which passes through the event “Sam’s clock reads 4 years” is the one marked “$t' = 4$ line”. This line also passes through the event “Ed’s clock reads 3.2 years”. Therefore, in Sam’s frame of reference the events “Sam’s clock reads 4 years” and “Ed’s clock reads 3.2 years” are the simultaneous events. This diagram thus explains how in Sam’s frame of reference, Ed’s clock is running slower than Sam’s by a factor of 0.8.

So, the idea that they each believe the other person’s clock is running slowly can be explained. We see that it is, indeed, a question of which frame of reference you are in, because different events are simultaneous in different frames. It is interesting to note that this situation only seems paradoxical in the first place because we are not use to the fact that simultaneity isn’t absolute. In everyday life, we get the idea that when two events happen at the same time, then that is an absolute fact. However, relativity shows us that this is not the case, and once we realize that, we can understand how each observer can believe the other observer’s clock is running slowly.

With this “first part” of the paradox solved, we must now move to the second part and ask this question: “how do we explain what happens when the twins come back together?”

4.1.3 Explaining the “Second Part”

In Diagram 4-2 Sam has seen his own clock read 4 years, and he then instantaneously accelerated to head back towards Ed. Right after the acceleration, Sam’s clock still basically reads 4 years. Note, however, that Sam’s frame of reference has changed. The inertial frame he was in before he turned around is different from his inertial frame after he turned around. I have thus drawn his new time line and a line of simultaneity (one which passes through the event “Sam’s clock reads 4 years”) for his new frame of reference.

Once again we will look at the simultaneous events in Ed’s frame and in Sam’s (new) frame. Since Ed hasn’t accelerated, he has remained an inertial observer, and his frame of reference hasn’t changed. Thus, in his frame the events “Ed’s clock reads 5 years” and “Sam’s clock reads 4 years” are still simultaneous. However, Sam is in a new frame of reference, and in this frame the events “Ed’s clock reads 6.8 years” and “Sam’s clock reads 4 years” are the simultaneous events.

So, each observer has his own explanation for the final outcome of the situation. For Ed, Sam’s clock is ticking slowly before the turn-around, nothing significant happens when Sam turns around, and Sam’s clock continues to tick slowly after the turn-around (because he is still moving at $0.6c$ with respect to Ed). That is how Ed explains why he has aged 10 years and Sam has only aged 8 years when they get back together at the end of Sam’s trip.

However, for Sam, the explanation is different. Before the turn-around, Sam is in a frame of reference in which Ed’s clock has been ticking slow, and it has ticked 3.2 years while Sam’s clock has ticked 4 years. After the turn-around, Sam is in a frame in which Ed’s clock (though it is still ticking slowly) has already ticked 6.8 years while Sam’s clock still reads only 4 years have passed. Note that since Ed’s clock is still running slowly in Sam’s new frame of reference, it will still only tick another 3.2 years (in Sam’s frame) during the last half of the trip, while Sam’s clock ticks another 4 years. However, since in Sam’s new frame, Ed’s clock has already ticked 6.8 years, the additional 3.2 years will make a total of 10 years of ticks for Ed’s clock. Meanwhile, Sam has seen his own clock tick a total of only 8 years.

And there you have it. Each observer agrees (as it must be) that when the two are back together again, Ed will have aged a total of 10 years while Sam has only aged a total of 8 years. They each have completely different ways of explaining how this happened, but in the end, they each agree on the final outcome.

4.1.4 Some Additional Notes

There are four specific things I want to make note of concerning the twin paradox as I have explained it.

First, we should note that the outcome of the above thought experiment (i.e. the fact that Sam ended up younger than Ed) is completely dependent on the fact that Sam turned around and headed back to Ed. If instead Ed had done the acceleration when he saw his own clock tick 4 years and had headed over to meet Sam, then Ed would be the one who had aged a total of 8 years while Sam had aged 10 years. Notice that the twin who undergoes the acceleration must actually have a physical force applied to him to cause that acceleration. During the acceleration he is no longer an inertial observer (this is why his frame of reference shifts while the other twin's frame does not shift). That differentiates his situation from the twin who does not accelerate, and that breaks the symmetry between the two observers. Unless one of them goes through an acceleration, their situations are completely symmetric, and there is no absolute answer to the question "which twin is younger?"

Second, I want to note something particular about the acceleration Sam went through. Look back at the lines of simultaneity drawn for Sam's frame before and after he accelerated. As we noted, the point where his "$t' = 4$" line of simultaneity cross the t axis (Ed's position) shifts upward when Sam turns around. Notice, however, that if Sam had taken a longer trip, then he would have done the acceleration when he was further from Ed. Then that "shift" would have been even larger, and after the acceleration, Sam's new frame of reference would be one in which Ed's clock had "jumped" ahead an even greater number of years. So, for Sam, the longer the trip he takes, the bigger the change will be when he switches his frame of reference, and that will make him an even greater number of years younger than Ed when they get back together. Of course, for Ed, the longer the trip is for Sam, the longer Sam's clock will be running slowly. So, Ed too agrees (with a different explanation) that Sam will be more years younger than Ed in the end if the trip is longer. As a final point on this, note that when Sam first accelerates to start his trip, he is right next to Ed, so the acceleration doesn't have much effect at all (as is true for his final acceleration at the end of the trip). That is why we basically ignored those accelerations.

Third, I want to note something about Sam's explanation of the events. Recall that when he changed frames of reference, his clock read 4 years while (in his new frame) Ed's clock read 6.8 years. One may think that Sam has thus changed to a frame where Ed's clock has been running faster; however, we know that in Sam's new frame, Ed is still moving with respect to Sam. Thus, in Sam's new frame Ed's clock has still been running slowly the whole time. To understand how this can be, consider a third observer (Tim) who has always been in the frame of reference which Sam has during the last part of the trip. Let's say that Tim was traveling along (going to Earth) when he saw Sam headed towards him, and to Tim's surprise, Sam turns around and joins Tim in Tim's frame of reference as the two come together. Thus, after Sam turns around, he and Tim are moving together, side by side. Now, Tim notices that right after Sam turns around, Sam's clock reads 4 years. Regardless of what Tim's clock reads, he can reset his clock to 4 years, and we can backtrack 4 years along Tim's path to identify the origin of Tim's frame (Sam's new frame). In Diagram 4-3 I have drawn (along with everything in Diagram 4-2) Tim's path, the origin (o') of Sam's new frame of reference, and a line of simultaneity for Tim's and Sam's frame at that origin. [!ht]

Notice that for Sam's new frame (the frame Tim has always been in) if $t' = 4$ when Sam turns around, then the event at Ed's position which is simultaneous with the origin in this frame (o') is the event "Ed's clock reads 3.6 years". And there you have it. In Sam's new frame, while it is true that Ed's clock is always been running slow, at the "beginning" for this frame (i.e. at its origin) Ed's clock started at 3.6 years. In this new frame of Sam's, Ed's clock had a "head start" (so to speak) when compared to Tim's clock. That is why Ed's clock already reads 6.8 years while Sam's clock reads only 4 years in Sam's new frame. In the end, we can describe the events in whatever frame of reference we wish, and though they may each have different explanation for what actually happens, they must all agree with the final outcome when the two twins come back together.

The final note I want to make is, again, about Sam's "view" of the events. When we say that before Sam's turn-around he is in a frame of reference in which Ed's clock reads 3.2 years, and after the turn-around Sam is in a frame of reference in which Ed's clock reads 6.8 years, one might be tempted to say that as Sam accelerates, Ed's clock speeds up in Sam's frame of reference. Of course, this doesn't change the way Ed sees his clock running, but it is only the way things occur in Sam's changing frame of reference. However, think about what would happen if Sam quickly changed his mind after the turn-around and immediately turned

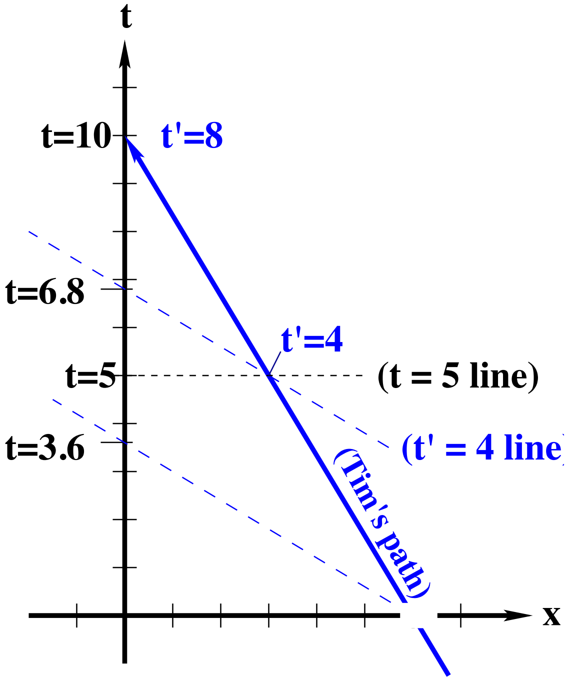

Diagram 4-3:

back around to his original heading. Then, in this new acceleration, Sam went from a frame where Ed’s clock read 6.8 years to a frame where Ed’s clock reads 3.2 years again. One would thus argue that Ed’s clock went backwards in Sam’s changing frame of reference. Again, this doesn’t have any real significance to the way Ed is reading his own clock, but we have to come to terms with the fact that Sam’s new acceleration caused Ed’s clock to go backwards in Sam’s changing frame. Perhaps the best way to think about this is simply to realize that Sam is not in an inertial frame since he is accelerating. Rather, Sam is simply changing into various inertial frames, and in each of these inertial frames, moving clocks do tick slowly, time does goes forward in all frames, etc. Either way you like to think about it, in the end, we can explain the outcomes as needed.

4.2 The “Car and Barn Paradox”

The “Car and Barn” paradox deals with the question of “whose lengths are shorter?” We have a barn whose front and back doors can be quickly open and closed. There is also a car which is just long enough so that if you try to fit it in the barn, and the barn doors close, they would close down on the front and back bumpers of the car. Now, an observer in the car (say, Carol) speeds the car towards the barn at a significant fraction of the speed of light. One might then argue the following: from the point of view of an observer sitting in the barn (say, Bob) the car will be length contracted, and at some point it will be completely inside the barn. Bob then reasons that he can close and open both barn doors while the car is completely inside the barn. However, Carol will argue that it is the Barn which moving with respect to here, and thus it the barn which is length contracted. So, she argues, if Bob tries to close both doors at the same time as the car goes through the barn, then the doors will smash into the car.

We thus want to ask whether or not the barn doors do smash into the car if Bob tries his idea, and how does each observer explain the outcome.

4.2.1 Viewing it with a Space-Time Diagram

As we did with the twin paradox, here we will look at a space-time diagram of the car and barn experiment in order to explain the paradox. We will draw our diagrams such the relative velocity of Carol and Bob is

0.6 c. In Diagram 4-4 I have drawn the situation keeping Bob's frame of reference in mind. To keep the diagram from getting too cluttered, a second diagram (Diagram 4-5) of the same situation will be used to mark points according to Carol's frame of reference. [!ht]

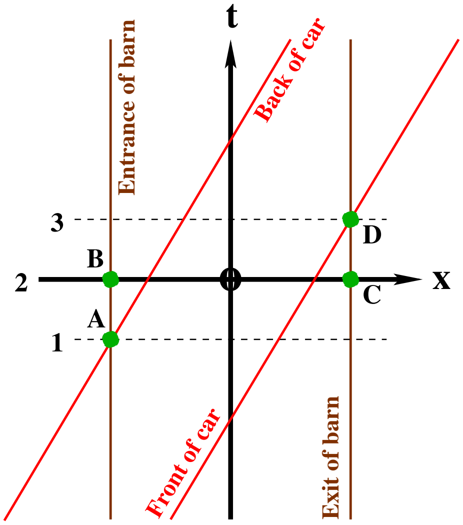

Diagram 4-4:

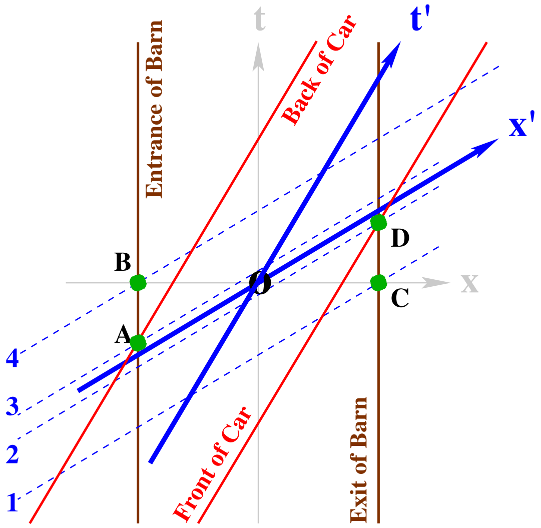

Diagram 4-5:

In the diagrams we have the following: The red lines indicate the paths of the front and back of the car (as marked) through space-time. The brown lines indicate the paths of the entrance and exit of the barn (as marked) through space-time. Hopefully it is apparent to the reader that the car travels from left to right (with respect to the barn) and passes through the barn. Also note that at the point where the entrance and exit of the barn cross the x axis (i.e. when the front and back of the barn are both at $t = 0$ in Bob's frame), both the front and back doors quickly close and open again. Those points are labeled B and C.

4.2.2 The explanation

We are interested in six different occurrences (though only 4 are shown in the diagrams). The ones not shown in the diagrams are, first, the front of the car enters the barn, and second, the back of the car exits the barn. These would appear much lower and much higher (respectively) in the diagram than is being shown here. The four events that we do note in the diagrams are (A) the back of the car enters the barn, (B) the entrance door of the barn closes and opens again, (C) the exit door of the barn closes and opens again, and (D) the front of the car exits the barn. In the diagrams, I have marked each of these events with the letters given and drawn lines of simultaneity (marked with dashes) for the observers.

In Diagram 4-4, we see that for Bob (whose lines of simultaneity are drawn in that diagram), (A) is the first event which happens, and everything that occurs simultaneous to (A) in Bob's frame of reference lies on the line marked with a 1. The next two events in Bob's frame are (B) and (C), which occur simultaneously. Everything which occurs simultaneous to these events lies on the line marked with a 2. Finally for Bob, (D) occurs, and everything which occurs simultaneous to it lies on the line marked with a 3. Note that for Bob, as the back of the car enters the barn–event (A)–the front of the car has yet to exit the barn. Also, when the doors close and open–events (B) and (C)–simultaneous in Bob's frame, the front and back of the car are inside the barn (the two red lines are inside the two brown lines along the line marked 2). Thus, in Bob's frame, the car is smaller than the barn, and it is inside the barn when the doors close and open. Finally, after both doors close and open, the front of the car exits the barn–event (D)–in Bob's frame.

However, in Diagram 4-5 we see simultaneous events marked from Carol's frame of reference. Again, the lines of simultaneity at each event are marked with dashes (but here they are drawn from Carol's frame and are blue). Now, we see that the "lowest" line of simultaneity on the diagram from Carol's frame of reference passes through the event (C), the exit door of the barn closes and opens. Thus, this event occurs first in Carol's frame. Everything occurring simultaneous with it in Carol's frame is on the line marked with a 1.

Next in Carol’s frame, event (D) occurs, followed by event (A), while event (B) occurs last. The events occurring simultaneous with these events are on the lines marked 2, 3, and 4, respectively. Thus, according to Carol’s frame, things happen as follows: First, while the front of the car is in the barn, but before the back of the car enters the barn, the exit door of the barn closes and opens. Next, the front of the car exits the barn. (Note that while the front of the car is then outside the exit of the barn at this point, the back of the car has yet to enter the barn in Carol’s frame–look along the x' axis, for example. So for Carol, the barn is smaller than the car.) Next, the back of the car enters the barn in Carol’s frame. Finally, after the front of the car has exited the barn and the back of the car has entered the barn, the entrance door of the barn closes and opens.

And there you have it. In the end, each observer must agree that the car gets through the barn without smashing into the doors. However, each frame of reference offers a different explanation for how this comes to be, because in each frame, different events are simultaneous with one another. In Bob’s frame, the car is in the barn all at once while the doors close and open simultaneously. However, in Carol’s frame, the doors do not close simultaneously, and the car is never completely in the barn.

So, I hope you have seen the power of space-time diagrams when it comes to explaining things in special relativity. When we simply say that moving clocks run slower and moving rulers length contract, we miss a real understanding of special relativity. That understanding comes from realizing that the actual coordinates in space and time for events are different for different observers who are moving with respect to one another. This relationship can be viewed with space-time diagrams, and the answers to many nagging questions in special relativity can be explained if one understands these diagrams.

Part III

A Bit About General Relativity

This is Part III of the “Relativity and FTL Travel” FAQ. It is an “optional reading” part of the FAQ in that the FTL discussion in Part IV does not assume that the reader has read the information discussed below. If your only interest in this FAQ is the consideration of FTL travel with relativity in mind, then you may only want to read Part I: Special Relativity and Part IV: Faster Than Light Travel–Concepts and Their “Problems”.

In this part, we take a look at general relativity. The discussion is rather lengthy, but I hope you will find it straight forward and easy to follow. The subject of GR is still new to this FAQ, and your comments on the usefulness, ease of reading, etc. for this part of the FAQ would be appreciated.

For more information about this FAQ (including copyright information and a table of contents for all parts of the FAQ), see the Introduction to the FAQ portion.

Chapter 5

Introduction to General Relativity

Thus far, we have confined our talks to the realm of what is known as Special Relativity (or SR). In this section I will introduce a few of the main concepts in General Relativity (or GR). The difference between the two is basically that GR deals with how relativity applies to gravitation. As it turns out, our concept of how gravity works must be changed because of relativity, and GR explains the new concept of gravity. It is called "General" relativity because if you look at General Relativity in the case where there is little or no gravity, you get Special Relativity (SR is a special case of GR).

Now, GR is a heavily mathematical theory, and while I will try to simply give the reader some understanding of the physical notions underlining the theory, some mathematics will inevitably come into play. I will, however, try to give simple, straight-forward explanations of where the math comes from and how it helps explain the theory. I will start by discussing why we might even think that gravity and relativity are related in the first place. This will lead us to change our concepts of space and time in the presence of gravity. To discuss this new concept of space-time, we will need to introduce the idea of mathematical constructs known as Tensors. The two tensors we will talk about in specific are called the Metric Tensor and the Stress-Energy Tensor. Once we have discussed these concepts, we will look at how it all comes together to produce the basic ideas behind the theory of general relativity. We will also consider a couple of examples to illustrate the use of the theory. Finally, we will mention some of the experimental evidence which supports general relativity.

5.1 Reasoning for its Existence

To start off our discussion, I want to indicate why one would reason that gravity and relativity are connected. While I could start with a somewhat unrealistic thought experiment to explain the first point I want to make, perhaps it will be better if I just tell you about actual experimental evidence to support the point. We thus start by considering an experiment in which a light beam is emitted from Earth and rises in the atmosphere to some point where the light is detected. When one performs this experiment, one finds that the energy of the light decreases as it rises.

So, what does this have to do with our view of relativity and gravity? Well, let's reason through the situation: First, we note that the energy of light is related to its frequency. (If you think of light as a wave with crests and troughs, and if you could make note of the crests and troughs as they passed you, then you could calculate the frequency of the wave as $1/dt$, where dt is the time between the point when one crest passes you and the point when the next crest passes you.) So, if the energy of the light decreases (and thus its frequency decreases), then dt (the time between crests) must increase. Let's then consider a frame of reference sitting stationary on the Earth. We will look at a space-time diagram in this frame which shows the paths that two crests would take as the light travels away from the Earth.

In Diagram 5-1 I have drawn indications of the paths the two crests might take. The diagram shows distance above the Earth as distance in the positive x direction, so as time goes on, the two crests rise (move in the positive x direction) and eventually meet a detector. Now, we don't know what the gravity of the Earth might do to the light. We thus want to generalize our diagram by allowing for the possibility that the paths of the crests might be influenced in some unknown way by gravity. So, I have drawn a haphazard path for the two crests marked with question marks. The actual paths don't matter for our argument, but what does matter is this: whatever gravity does to the light, it must act the same way on both crests. Therefore, the two haphazard paths are drawn the same way. [!ht]

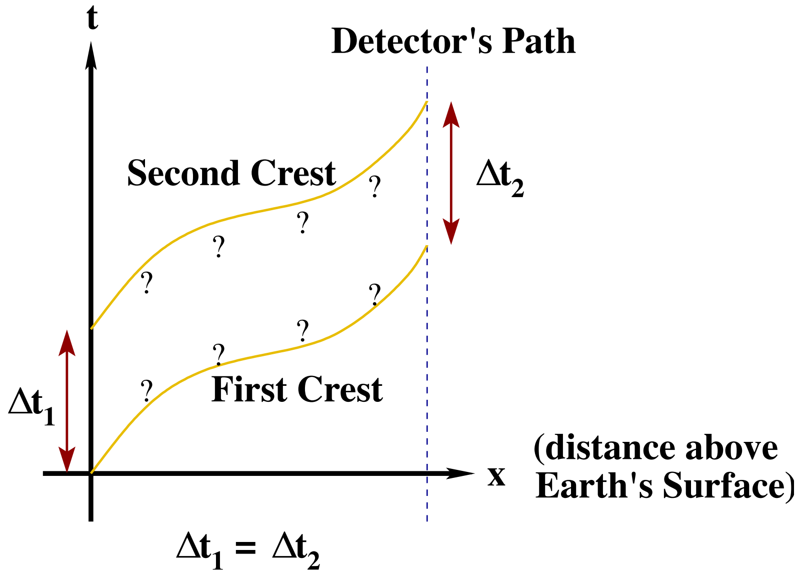

Diagram 5-1:

As we see in the diagram, because gravity acts the same way on both crests, the time between them when they leave the surface (Δt_1) is the same as the time between them when they are detected (Δt_2). Thus, our diagram does not predict that the energy of the light should change, but experimental evidence shows it does. According to special relativity, this frame of reference we have drawn is an inertial frame (that is, if we ignore the Earth's motion, this frame of reference is stationary–it's just inside a gravitational field). Thus our diagram (drawn for an inertial frame of reference) should explain the geometry of the situation, but does not. That indicates that SR must be changed in light of gravity. However, we have yet to show that SR must be completely thrown out.

What if there were another way to define an inertial frame such that its geometry would explain the above situation and other situations which occur in the presence of a gravitational field? That is what we will consider next.

5.2 The "New Inertial Frame"

Before starting this section, I want to mention something to the reader: in the end, when gravity is concerned, we will not be able to find a single inertial frame of reference which will correctly explain the geometry of all situations. This will be the actual death-blow to special relativity. In this section, it will start to look as if the situation is hopeful, and that by defining a proper inertial frame, SR will be saved. However, in the next section, we will see where this all falls apart, and I want the reader to realize this from the beginning.

Now, in the previous section we showed that a space-time diagram drawn for an inertial frame of reference doesn't explain the way things really are for a frame of reference sitting stationary on the Earth's surface. If such a frame cannot be called an inertial frame because of some effect of gravity, then perhaps there is another way to define an inertial frame of reference in the presence of gravity.

First, let's consider the properties of a frame which we know to be an inertial frame without gravity. Consider a space ship sitting far from any source of gravity. Here we will assume that the ship isn't accelerating–it's just sitting there in the middle of space. Diagram 5-2 shows such a space ship at different times. Also shown is an observer and a ball, both of which start out stationary in this frame of reference.

Both the observer and the ball are weightless along with the ship, and as time goes on neither move with-respect-to the sides of the ship. This is obviously what we would consider to be an ideal inertial frame of reference. [!ht]

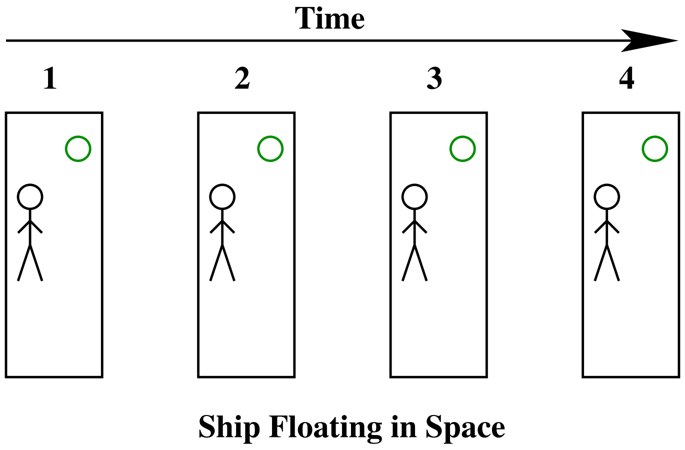

Diagram 5-2:

Next, consider the same ship, but let it be sitting stationary on the Earth. Diagram 5-3 shows such a ship at different times, and again there is an observer and a ball shown as well. Obviously, the observer and the ball in this case cannot remain stationary with respect to the ship–rather they must fall in the Earth's gravity and accelerate towards the Earth's surface. Note that because of the way gravity works, the observer and the ball and anything else in the ship will accelerate downward at the same rate regardless of their mass (as long as they are at relatively the same height above the Earth's surface, and neglecting air resistance). This distinguishes gravity from all other forces in nature. With the other three forces (electromagnetism, the strong nuclear force, and the weak nuclear force) the motion of an object in the presence of the force depends on the composition of the object. For example, electromagnetism doesn't act on neutral particles, but does act on charged ones. However, when we consider gravity, the path taken by an object which is released with a given velocity in a gravitational field does not depend on the composition of the object–not even its mass. So, both the ball and the observer in Diagram 5-3 accelerate at the same, constant rate towards the bottom of the ship. In step 3 on that diagram, the observer hits the bottom of the ship, and in step 4 the ball reaches the bottom as well. Obviously this situation isn't like the inertial frame of reference we described above, and the observer in these two situations could easily tell the difference between the two cases. [!ht]

Further, consider the same ship again, this time letting it accelerate at a constant rate in the middle of space. Diagram 5-4 shows such a ship at different times (again with an observer and a ball). Note that in the diagram, the observer and the ball start out at a constant speed (in steps 1, 2, and 3, both move one interval up during each step of time). However, the acceleration of the ship causes it to move further between steps 2 and 3 than it did between steps 1 and 2, and so on. Therefore, at step 3 the bottom of the ship meets with the observer, and the observer begins to be pushed by the ship, accelerating along with the it from then on. This would cause the observer to feel the force of the ship against him, “holding” him against the floor. In the final step, the ball meets with the bottom of the ship, and it too accelerates from then on because the ship is pushing against it. This case thus looks very much like the case just above where the ship was sitting on the Earth's surface–in both cases objects in the ship will seem to accelerate at the same, constant rate towards the bottom of the ship (regardless of their mass) and once there they will feel a force against them as they sit on the floor of the ship. The observer in each of these cases would find it hard to tell which of the two situations he was really in. [!ht]

Given all three examples above, it seems obvious that a frame sitting stationary on the Earth is much more like an accelerating frame than it is like an inertial frame. Seeing that, it now seems perfectly reasonable for us to find that an experiment performed on the surface of the Earth can't be explained by a diagram drawn for an inertial frame.

But, can we now find a frame of reference in the presence of gravity which **does** look like an inertial

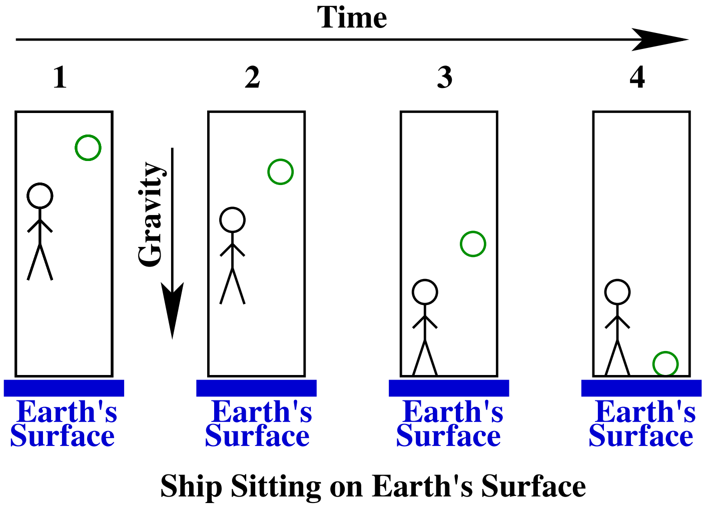

Diagram 5-3:

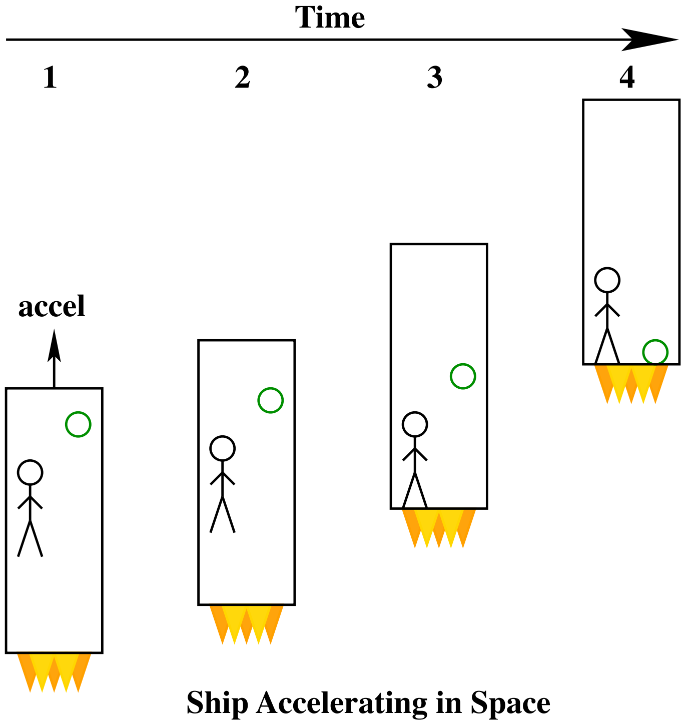

Diagram 5-4:

frame? Well, look back to Diagram 5-4 (where the ship is accelerating in space) and notice the state of the ball and the observer during the first part of that illustration. Even though the ship in that case is not an inertial frame because it is accelerating, the observer and the ball don't begin to accelerate until the bottom of the ship reaches them and begins to push them. Thus, until that point, the ball and the observer are not accelerating. They are shown moving at a constant velocity. Thus, until the bottom of the ship reaches them, the observer and the ball are inertial observers. **Ah**, but as we have pointed out, this situation is supposed to be analogous to the one in Diagram 5-3 (where the ship is sitting stationary on the Earth). If so, then we could argue that the observer and the ball in the first part of Diagram 5-3 (which are in free-fall in the Earth's gravitational field) are what we would now call our inertial observers in the presence of gravity.

So, let's look at one last illustration in which the whole ship is in free-fall as well as the observer and the ball. Diagram 5-5, shows such a situation. Notice that the observer, the ball, and the ship all accelerate at the same rate towards the earth. They each move the same distance during each step shown. Now, look at just the ship and everything in it at each step shown. The observer, the ball, and the sides of the ship are not moving with respect to one another because they are all falling at the same rate. At each step, the ball and the observer are at the same position inside the ship. Therefore, until the ship in Diagram 5-5 reaches the surface of the Earth, the observer wouldn't notice any difference between this situation and the one in Diagram 5-2 (where the ship is floating in space). [!ht]

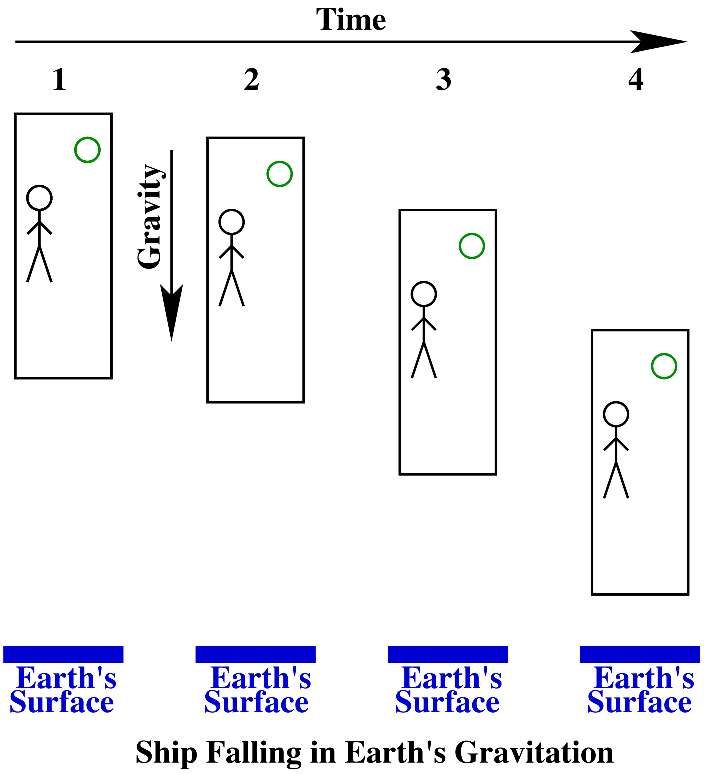

Diagram 5-5:

It certainly seems, then, that a frame which is freely falling in the presence of gravity is actually an inertial frame of reference. As one final test, let's go back to the experiment mentioned earlier in which light rises in the presence of Earth's gravity. As it turns out (though I won't go into the proof) if the light is detected while it is still relatively close to the Earth, and we consider the experiment in a frame of reference which is freely falling near the Earth's surface, then in that frame, the light does not loose energy. Thus, in the freely falling frame of reference, Diagram 5-1 (which depicts an inertial frame of reference) can correctly depict the geometry of the situation.

And so, things are looking deceptively hopeful. In every case we have studied, it seems as if we can continue to use special relativity as-is, even in the presence of gravity, if we simply define "inertial frame" to mean a frame which is in free fall. Then the space-time diagrams we have drawn throughout our discussions would work just fine in the presence of gravity, as long as we understand that they are drawn in free falling frames. However, as I warned earlier, there is a problem here which we haven't solved.

5.3 The Global Break-Down of Special Relativity

Now that we have tried to argue that we can continue using special relativity even when gravity is involved (by appropriately defining a new inertial frame), we are now in a position to explain where the argument breaks down.

Consider Diagram 5-6. There we see a ship which is much wider than the ships we have shown thus far. It is in free fall towards the surface of the Earth, and there are two observers shown, one at either side of the ship. Now, according to our argument, both observers are said to be in inertial frames of reference because they are both in free-fall. However, as they each fall towards the center of the Earth, because they are at great distances from one another, they accelerate in different directions as shown. If one observer looks at the other, he will see that other observer accelerating towards him. But if they are both supposed to be inertial observers, then how can they also each be accelerating in the frame of the other? [!ht]

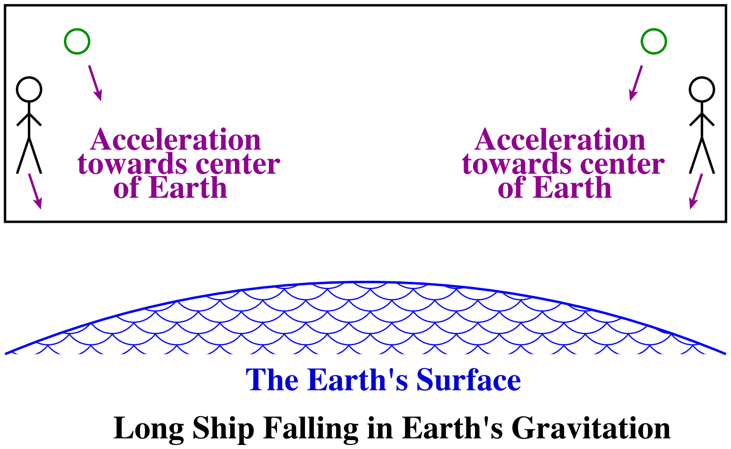
Acceleration towards center of Earth
Acceleration towards center of Earth
The Earth's Surface
Long Ship Falling in Earth's Gravitation

Diagram 5-6:

Also, consider Diagram 5-7 in which there is a ship which is much taller than the ships we have been considering. Here, two observers are again shown, one at the bottom of the ship and one at the top. Because the one near the bottom is much closer to the surface of the Earth, he is accelerating at a greater rate than the other observer. Again, these two observers are both supposed to be inertial observers, yet each is accelerating in the other observer's frame. Further, as the observer on the top continues to accelerate downward, he will eventually be where the observer at the bottom is now. Thus, as time passes, he will fall into a stronger gravitational field, and he will be in a "different" inertial frame than he his now. [!ht]

What does all this say? Well, we have shown that for small distances and over small amounts of time, a free falling frame has all the properties we want in an inertial frame when gravity is present. However, in each of the last two cases above, we have observers who are all free-falling and thus (by our new definition of an inertial frame in the presence of gravity) are all supposed to be in inertial frames. Yet, if we draw a space-time diagram for one of the observers, and extend it so that the other observer can be drawn on the diagram, that other observer will be accelerating on the space-time diagram. Therefore, a space-time diagram which well describes an inertial frame for all of space-time in special relativity can only well describe an inertial frame of reference over a small distance in space and time when a general gravitational field is involved.

This is analogous to the situation in which a flat map can well describe a small, local piece of the curved surface of the Earth (such as a city). However, globally, as you extend the map, it no longer describes the curved surface of the Earth.

We therefore find that when gravity is involved, we can still define an inertial frame of reference **locally** (meaning local in both space and time), but globally, there is no way to define a single, rigid frame of reference which describes an inertial frame of reference everywhere in space-time. Therefore, globally we cannot use special relativity to describe space-time in the presence of a general gravitational field. We must therefore re-think relativity in the presence of gravity.

What we will find is that gravity is actually caused by a curvature of space-time, and like the map trying in vain to describe the curved surface of the Earth, special relativity cannot describe the curved space-time caused by gravity. It is general relativity which describes curved space-time, and for us to fully appreciate it, we will need to discuss some basic ideas used to describe such a geometry.

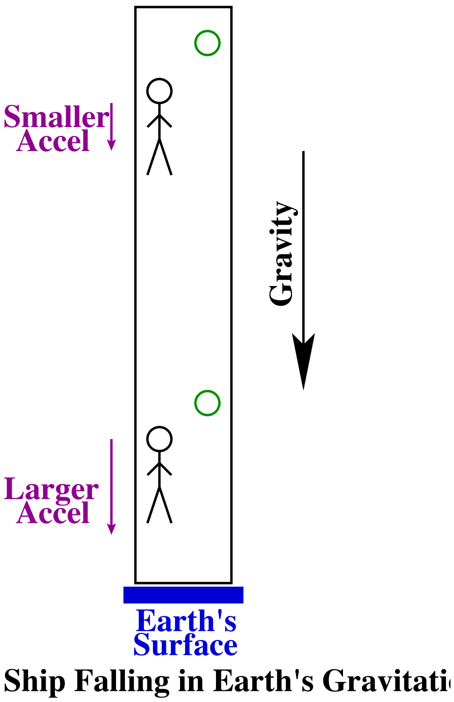

Diagram 5-7:

5.4 Manifolds, Geodesics, Curvature, and Local Flatness

Before we discuss space-time in the presence of gravity, we need to understand some basic geometric concepts which we will use. We will develop these concepts by considering normal, spatial geometry which can be fully grasped using common sense. Applying these concepts to space-time becomes less intuitive (in part because we still aren't that used to thinking of time as just another dimension); therefore, developing them using normal spatial geometry will be beneficial.

First, we introduce the term “manifold”. Basically, for our purposes, you can think of a manifold as a fancy term for a space. The space around us that you are used to thinking of can be called a three dimensional manifold. The surface of a sheet of paper is a two dimensional manifold, as is the surface of a cylinder or the surface of a sphere. Much of our focus on manifolds will involve discussing their geometry. Understanding the geometry of a manifold means understanding the relationships between various points on the manifold and understanding various curves on the manifold as well as knowing how to measure distances on the manifold. Thus, we want to define a few specific notions which will help us understand and explain the geometry of a manifold.

So, next we look at a particular type of path on a manifold called a geodesic. A geodesic is essentially the path which takes the shortest distance between two points on the manifold. On a piece of paper (a flat manifold) the shortest distance between two points is found by following the path of a straight line. However, for a sphere, the shortest distance between two points would be traveled by following a curve known as a great circle. If you imagine cutting a sphere directly in half and then putting it back together, then the cut mark on the surface of the sphere would be a great circle. If you move along the surface of a sphere between two points, then the shortest path you could take would lie on a great circle. Thus, a great circle on a sphere is basically equivalent to a line on a flat manifold–they are both geodesics on their respective manifolds. Similarly, on any other manifold there would be a path to follow between two points such that you would travel the shortest distance. Such a path is a geodesic on that manifold.

Next, we introduce the concept of the curvature of a manifold. There are two different types of curvature: intrinsic and extrinsic. To demonstrate the difference between the two, let's first consider a surface which has only extrinsic curvature. Imagine taking a flat sheet of paper and rolling it as if you were making a cylinder; however, don't let the two ends touch to complete the cylinder. Now, while this two dimensional surface will now look curved in our three dimensional perspective, the geometry of the surface is still the same as the geometry of the flat sheet of paper from which it was made. If you were a two dimensional creature confined to live on this two dimensional surface, there would be no test you could perform to prove you weren't on a flat sheet of paper rather than this cylinder-like surface. Now if you did complete the cylinder, then a two dimensional creature could tell that the global topology of the situation has changed (for example, on a complete cylinder, he could follow a particular path which would bring him around back to where he started). However, this doesn't change the fact that throughout the cylinder, the internal geometry is just like the geometry of a flat sheet of paper from which it was made.

So, for a two-dimensional cylinder, its curvature is only "visible" when viewed from a higher dimensional space (our three-dimensional space). We only say it is curved because a line on the 2-D cylinder can bend away from a straight line in three dimensions. However, The cylinder has no intrinsic curvature to its geometry, so its curvature is extrinsic.

Contrast this with the surface of a sphere. You cannot bend a flat sheet of paper around a sphere without crumpling or cutting the paper. The geometry on the surface of a sphere will then be different from the geometry of a flat sheet of paper. To distinctly show this, let's consider a couple of two dimensional creatures who are confined to the surface of a sphere. Say that they stand facing the same direction at a given, small distance apart from one another on the two dimensional surface, and then they begin walking in the same direction parallel to one another. As they continue to walk beside one another, each will continue in what seems to him to be a straight line. If they do this–if each of them believes that he is following a straight line from one step to the next–then each will follow the path of a geodesic on the sphere. As we said earlier, this means that they will each follow a great circle. But if they each follow a great circle on the surface of a sphere, then they will each eventually notice that their friend walking next to them is moving closer and closer, and eventually they will meet. Now, they started out moving on parallel paths, and they each believed that they were walking in a straight line, but their paths eventually came together. This would not be the case if they performed this experiment on a flat sheet of paper (or on a cylinder). Thus, creatures who are confined to live on the two dimensional surface of a sphere could tell that the geometry of their space was different from the geometry of a flat piece of paper (even though they couldn't "see" the curvature because they are trapped in only two dimensions). That intrinsic difference is due to the intrinsic curvature of the sphere's surface.

This, then, is what we want to note about curvature: There are two types of curvature, extrinsic and intrinsic. Extrinsic curvature is only detectable from dimensions higher than the dimension of the manifold being considered. Intrinsic curvature can be detected and understood even by creatures who are confined to live within the dimensions of the manifold. Thus, just because a manifold may looked "curved" in a higher dimension, that doesn't mean that its intrinsic geometry is different from that of a flat manifold (i.e. its geometry can still be flat–like the cylinder). Thus, the test of whether a manifold is intrinsically curved does not have anything to do with higher dimensions, but with experiments that could be performed by beings confined on that manifold. (For example, if two parallel lines do not remain parallel when extended on the manifold, then the manifold possesses curvature). This is important to us in our discussion of space-time in the presence of gravity. It means that the curvature of the four dimensional manifold of space-time in which we live can be understood without having to worry about or even speculate on the existence of any other dimensions.

As a final note in this introduction to manifolds, I want to mention a bit about local flatness. Note that even though a manifold can be curved, on a small enough portion of that manifold, it will be fairly flat. For example, we can represent a city on our curved Earth by using a flat map. The map will be a very good representation of the city because it is a very small piece of the curved manifold. Earlier I mentioned that over a small enough piece of space-time in the presence of gravity, you can define a frame of reference which is still very similar to an inertial reference frame in special relativity. This gives an indication as to why the geometry of space-time in special relativity is that of a flat manifold, while with general relativity, space-time is said to be curved in the presence of gravity. Still, the space-time of any observer being acted on only by gravity is **locally** flat.

Later we will see how the concepts discussed here will help us in explaining gravity and relativity. Next, however, we want to discuss another property of manifolds which itself will tell us everything we want to know about the geometry of a particular manifold. We will call this property the invariant interval.

5.5 The Invariant Interval

Here we will basically be discussing distances on manifolds, and what we can learn about a manifold based on how we calculate distances on that manifold. We start by discussing the length of a random path on a manifold.

Consider a random path on a flat sheet of paper. We can use an x-y coordinate system to specify any point on the paper and any point on the path. With this coordinate system in place, how can we use it to measure the length of that random path? One way is to break up the path into tiny parts, each of which can be approximated with a straight line segment. Then, if we know how to measure the length of a straight line, we can measure the length of each line segment and add them up to find the approximate length of the path. Now, since the random path doesn't have to be very straight, the line segments we use might not be very good at approximating the path at some point. However, if we break up the path into smaller pieces, then the smaller line segments should do a better job of approximating the curve and giving us the correct length for the path. The smaller we make the line segments, the better our approximation of the path's length will be. The ultimate result of this idea is to figure out what the calculated length would be if we made the line segments infinitesimally small. That would give us the actual length of the curve.

So, the next question is this: How do we calculate the length of a very small (infinitesimal) line segment using our x-y coordinate system? Well, each segment is made up of a component in the x direction (dx) and a component in the y direction (dy) as shown in Diagram 5-8. These components represent infinitesimal distances. The length of the infinitesimal line segment (let's call the length ds) is then given by the following (using the Pythagorean theorem):

$$ds^2 = dx^2 + dy^2 \tag{5:1}$$

(Note that this is the length of a straight line–a geodesic on this manifold–between an initial and a final position which are separated by a distance dx in the x direction and dy in the y direction.) [!ht]

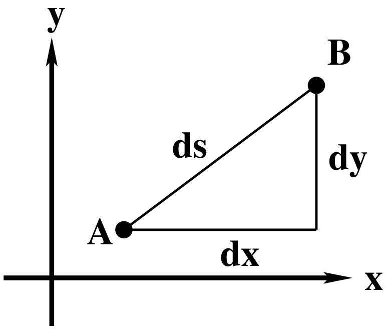

Diagram 5-8:

This distance between two very-nearby points is what I call the invariant interval. Why? Well, first I need to note that there are other types of coordinate systems one could use to locate every point on a flat surface, and that the equation for ds in terms of small changes in each coordinate will depend on the coordinate system you use. However, though the form of the equation will change, the actual distance between two points on the manifold is a physical reality which won't change. The actual interval is independent of the coordinate system you place on the manifold.

Now, Below, I will specifically use ds as defined here (in a flat, x-y coordinate system) to make a comparison with an invariant interval defined using a particular coordinate system on a curved manifold. However, all the arguments I will make can also be made using any other coordinate system on a flat manifold and any other coordinate system on a curved manifold. I simply use two specific ones as solid examples.

So, to demonstrate how the equation for ds will tell us everything we want to know about a manifold, we next need to consider a curved manifold. We will use our old friend the sphere. Let's start by defining a coordinate system on the sphere. Picture a sphere with a great circle drawn on it. Let's call that great circle the equator. Next, consider a point on the equator, and call that point our origin. We want to define two independent coordinates which will allow us to locate any point on the sphere starting from the origin (note: by "independent coordinates" I mean that you can always change your position in one coordinate independent of any change in the other). So, consider some other point on the sphere (call the point "P"), and let's explain how to get to that point using two coordinates. We start by moving either towards the "east" or "west" from our origin in the general direction of "P" (you can define "east" and "west" however you wish). We move along the equator until P is directly north or south of us, and we call the distance we move "L" (L is positive if we move east). Next, we need to move north or south on the sphere to reach P. The distance we move north or south to reach P will be called "H" (H is positive if we move north). That gives us our coordinate system. Every point on the sphere can now be represented by an L-H coordinate pair. The "grid" on the surface of the sphere which represents this coordinate system would be made of latitude and longitude lines such as those on a globe.

Next, we need to figure out what infinitesimal distance (ds) would be associated with moving a small distance in L (dL) and a small distance in H (dH). For the sake of time, I'll just give the answer here. (Note, R is the radius of the sphere we are considering):

$$ds^2 = dH^2 + \cos^2\left(\frac{H}{R}\right) dL^2 \tag{5:2}$$

Remember what this represents. If you start at some point (L,H) on the sphere, and you change your L coordinate by a small amount (dL) and your H coordinate by a small amount (dH) then the shortest distance along the sphere between your first position and your final position would be ds. Note that this distance depends on your H position (because of the "$cos(H/R)$" part of the equation). This is an interesting point because as soon as you start moving from one position to the next, the equation for ds becomes slightly different. We basically think of this difference as negligible as long as dL is very small, but, in fact, the equation is only correct when dL is truly "infinitesimal". Such concepts are generally covered in calculus, and for our purposes, we will just claim that the equation is practically true as long as dL is very small.

So now we come to an important statement to be made in this section: **the form of the invariant interval fully defines the intrinsic geometry of a manifold**. For example, what if we tried to find another coordinate system on the sphere using two independent coordinates (a and b) such that the invariant interval on the sphere would be given by the following:

$$ds^2 = da^2 + db^2? \tag{5:3}$$

Well, because that invariant interval looks just like the formula for ds on a flat sheet of paper ($ds^2 = dx^2 + dy^2$), then it should be impossible for Equation 5:3 to be the invariant interval on the sphere (no matter how we define "a" and "b"). If I drew a grid on a flat sheet of paper and labeled the axes "a" and "b", then Equation 5:3 would appropriately describe the relationship between every single point on that flat manifold given the "a" and "b" coordinate system. Thus, if I define "a" and "b" to be independent coordinates on a sphere, and I claimed that Equation 5:3 described the invariant interval on the sphere given those coordinates, then I'd be saying that Equation 5:3 describes the relationship between every single point on the sphere given the "a" and "b" coordinate system. But that's saying that by appropriately defining "a" and "b", I can make the relationship between all points on the sphere be just like the relationship between every point on a flat sheet of paper. We know that physically, this simply can't be done, because there are intrinsic ways to tell the difference between the geometry of a sphere and the geometry of a flat sheet of paper.

You might be looking back at Equation 5:2 and thinking, "but what if I just define a new coordinate, L' such that $dL'^2 = cos^2(H/R)\, dL^2$? Then I get $ds^2 = dH^2 + dL'^2$, which looks like the invariant interval for a flat sheet of paper." Ah, but look at your definition for dL' and notice that it involves your other coordinate, H. You see that H and L' are **not** independent coordinates. To be valid in our discussion here, the coordinates you use on a manifold must be independent.

So, considering this example of a sphere and a flat sheet of paper, let's make some general points: First, consider some manifold, M1. On M1, we have some (valid) coordinate system, S1. Next we consider two very-nearby points on M1 (call the points P and Q). If we know the distance between P and Q along each of the coordinates (like dx and dy, for example), then we can find some function for ds (the shortest distance on M1 between the very-nearby points) using the coordinates in S1. Now, consider a second manifold, M2. If a (valid) coordinate system, S2, can be defined on that manifold such that ds has the same functional form in S2 as it did using the S1 coordinate system on M1, then the geometry of the two manifolds must be identical.

This indicates that the geometry of a manifold is completely determined if one knows the form of the invariant interval using a particular coordinate system on that manifold. In fact, starting with the form of the invariant interval in some coordinate system on a manifold, we can determine the curvature of the manifold, the path of a geodesic on the manifold, and everything we need to know about the manifold's geometry.

Now, the mathematics used to describe these properties involves geometric constructs known as tensors. In fact, the invariant interval on a manifold is directly related to a tensor known as the metric tensor on the manifold, and we will discuss this a bit later. First, I want to give a very brief introduction to tensors in general.

5.6 A Bit About Tensors

In this section I will introduce just a few basic ideas which will give the reader a feeling for what tensors are. This is simply meant to provide a minimum amount of information to those who do not know about tensors.

Basically, a tensor is a geometrical entity which is identified by its various components. To give a solid example, I note that a vector is a type of tensor. In an x-y coordinate system, a vector has one component which points in the x direction (its x component) and another component which points in the y direction (its y component). If you consider a vector defined in three dimensional space, then it will also have a z component as well. Similarly a tensor in general is defined in a particular space which has some number of dimensions. The number of dimensions of the space is also called the number of dimensions of the tensor. Note that vectors have a component for each individual (one) dimension, and they are called tensors of rank 1. For other tensors, you have to use two of the dimensions in order to specify one component of the tensor. In x-y space, such a tensor would have an xx component, an xy component, a yx component, and a yy component. In three-space, it would also have components for xz, zx, yz, zy, and zz. Since you have to specify two of the dimensions for each component of such a tensor, it is called a tensor of rank 2. Similarly, you can have third rank tensors (which have components for xxx, xxy, ...), fourth rank tensors, and so on.

So that you aren't confused, I want to explicitly note that the dimensionality of a tensor (the number of dimensions of the space in which the tensor is defined) is independent of the rank of the tensor (the amount of those dimensions that have to be used to specify each component of the tensor). In any dimensional space, we can have a tensor of rank 0 (just a number by itself, because it is not associated in any way with any of the dimensions), a tensor of rank 1 (like a vector–it has a component for every one dimension you can specify), a tensor of rank 2 (it has a component for every pair of dimensions you can specify), etc.

Now we look at a very important property of tensors. In fact, it is the property which really defines whether a set of components make up a tensor. This property involves the question of how the tensor's components change when you change the coordinate system you are using for the space in which the tensor is defined. So, let's consider an example in two dimensional space where you go from some coordinate system (call the coordinates x and y) to some other coordinate system (call these coordinates x' and y'). There will be some sort of relationship between the two systems. For example, say we start at some point in this space such that our coordinates are (x, y) and (x', y') (depending on which coordinate system you are using). Now, say we move an "infinitesimal distance" in x (using the first coordinate system). Call that distance dx. When we do so, we may have changed our x' position (using the second coordinate system) by some infinitesimal amount, dx'. Also, we may have changed our y' position by some amount dy'. We can use these concepts of infinitesimal changes to define some relationships between the two systems. We can answer the question "how does x' change when x changes at this point" by noting the ratio, $\frac{dx'}{dx}$. Similarly we can write $\frac{dx}{dx'}$ to denote how much x changes with changes in x' at some point, and $\frac{dy'}{dx}$ denotes how y' changes with changes in x.

Please understand that these are not simply ratios of definite numbers. For example, $\frac{dx'}{dx}$ is not necessarily the inverse of $\frac{dx}{dx'}$ because dx in one expression is NOT the same as dx in the other. The first expression uses dx in the following context: "If I hold y constant and change x by an amount dx, x' and y' might change by amounts dx' and dy'. Take the amount that x' changes (dx') and divide it by the amount I changed x (dx)." The second expression uses dx in the following context: "If I hold y' constant and change x' by an amount dx', x and y might change by amounts dx and dy. Take the amount that x changes (dx) and divide it by the amount I changed x' (dx')." You can see that the dx in the former context does not have to be the same amount as dx in the latter. So, when I write $\frac{dx'}{dx}$ or $\frac{dx}{dx'}$ or $\frac{dy}{dx'}$ etc, you must understand that the form of these ratios (what's on top and what's on bottom) defines how they are produced, and they are not just ratios of definite numbers. (Those who know something of calculus will obviously recognize these terms as simple partial derivatives, but anyway....)

Now, all together there are four of these ratios which denote how the x' and y' coordinates change with changes in x and y:

$\frac{dx'}{dx}$, $\frac{dx'}{dy}$, $\frac{dy'}{dx}$, and $\frac{dy'}{dy}$.

Similarly, there are four more to denote how x and y change with changes in x' and y':

$\frac{dx}{dx'}$, $\frac{dx}{dy'}$, $\frac{dy}{dx'}$, and $\frac{dy}{dy'}$.

In general the values of these ratios can depend on where you are on a manifold, so each ratio is generally a function of x and y (or x' and y', if you like).

Now, we have these ratios which help us relate one coordinate system to another. If we have a tensor defined in this space, then we must be able to use those ratios to find out how the tensor's components themselves change when we go from considering them in one coordinate system to considering them in the other. Let's consider a tensor of rank 1 (a vector) in a two dimensional space. Let the vector, call it V, have an x component (V_x) and a y component (V_y). Then, the rules for finding the x' and y' components of the vector at some point are the following:

$$\begin{aligned} V_{x'} &= \tfrac{dx'}{dx}V_x + \tfrac{dx'}{dy}V_y \\ \text{and} & \\ V_{y'} &= \tfrac{dy'}{dx}V_x + \tfrac{dy'}{dy}V_y. \end{aligned} \tag{5:4}$$

That is the way in which this type of first rank tensor must transform from one coordinate system to another. Note that we can write both equations in Equation 5:4 by using the following:

$$V_{a'} = \sum_{(b=x,y)} \left[\frac{da'}{db}V_b\right] \tag{5:5}$$

In that expression, "a" can be either x or y (so we actually have two equations–those in Equation 5:4). Also, the right side of the equation is a summation where the first term in the summation is found by letting $b = x$, and the second term is found by letting $b = y$. Further, we could make this expression more general by noting that it will be true for a space with higher dimensions when we let "a" be any one of those dimensions and let the sum with b extend over all the dimensions.

The fact that the physical components of a vector do actually transform this way is what makes the vector a tensor. However, we should note that not all types of vectors transform this way.

To show this is so, first we will consider a function which has a value at every point in x-y space. Call the function $f(x, y)$. Such a function is a 0 rank tensor, because at any point in the space, it has some single, numerical value (it does not have components for x and y like a vector does–you can't ask "what's its value in the $x$direction", or "what's its value in the y direction", because it has only a single number at any point). Note that if we change to another coordinate system, the value of f at some physical point in the space will not change. Because it has no x or y component, it is invariant when you change coordinate systems, as are all 0 rank tensors. This is the way all 0 rank tensors must transform when you change coordinate systems–they must be invariant.

Now, back to the point that there are other types of vectors which do not transform as discussed earlier. Let's take the function we were just discussing, $f(x, y)$, at some point and ask "how does it change with small changes in x?" If the function changes by an amount df when we move to another x location a distance dx away, then we can write the expression $\frac{df}{dx}$ to tell how f changes with x. We can do the same in y and have the expression $\frac{df}{dy}$. Then we could define a vector (call it G) which has an x component (G_x) equal to $\frac{df}{dx}$ at every point in x and y, while it has a y component (G_y) equal to $\frac{df}{dy}$ at every point. Now, what if we do this same procedure in the x'-y' coordinate system. First, we need to convert f into a function f'. We do this such that if a point in our space has coordinates (x, y) in one coordinate system while the same physical point has coordiantes (x', y') in the other coordinate system, then we want $f(x, y) = f'(x', y')$. That way f' is the proper representation of f in the primed coordinate system. Now we again find a vector, G, and we will end up with the x' and y' components of the G vector such that $G_{x'} = \frac{df'}{dx'}$ and $G_{y'} = \frac{df'}{dy'}$.

We now want to figure out how to transform G from one frame to another. First, we will look at $G_{x'} = \frac{df'}{dx'}$ which says that $G_{x'}$ comes from knowing how f' changes with respect to x' (i.e. $\frac{df'}{dx'}$). To transform this component of G, we must know how to find $\frac{df'}{dx'}$ using G_x and G_y. This means we will be using information about how f changes with respect to x and y (i.e., using $\frac{df}{dx}$ and $\frac{df}{dy}$). We will also need to use information about how x and y change with respect to x'. Without taking the time to fully explain the calculus involved, perhaps the following equation will not be too surprising:

$$\frac{df'}{dx'} = \frac{df'}{dx}\frac{dx}{dx'} + \frac{df'}{dy}\frac{dy}{dx'} \tag{5:6}$$

Conceptually (though mathematicians would cringe a bit at this explanation) one can imagine canceling out the dx in $\frac{df'}{dx}\frac{dx}{dx'}$ and canceling out the dy in $\frac{df'}{dy}\frac{dy}{dx'}$ to see that in both parts of that equation we are looking at information about $\frac{df'}{dx'}$. In the first case, we are looking at how f' changes with respect to x' by way of how x changes with respect to x', while in the second case we are looking at how f' changes with respect to x' by way of how y changes with respect to x'. Adding these two components together as we do in the above equation gives us a full picture of how f' changes with respect to x' given information about how f' changes with respect to x and y.

We further note that f' and f are actually the same physical function, we just use the prime to indicate which coordinate system we are primarily thinking of. Thus f and f' will both change in the same way with respect to changes in x and y (i.e. $\frac{df'}{dx} = \frac{df}{dx}$ and $\frac{df'}{dy} = \frac{df}{dy}$). We therefore rewrite Equation 5:6 as

$$\begin{aligned} \tfrac{df'}{dx'} &= \tfrac{df}{dx}\tfrac{dx}{dx'} + \tfrac{df}{dy}\tfrac{dy}{dx'} \\ &= G_x\tfrac{dx}{dx'} + G_y\tfrac{dy}{dx'} \end{aligned} \tag{5:7}$$

Note that we have substituted $G_x = \frac{df}{dx}$ and $G_y = \frac{df}{dy}$. The above equation provides the transformation of $G_{x'}$ given the components of G in the (x, y) coordinate system. Similarly, we can also find the transformation of $G_{y'}$. In the end, simply because of the way this vector is defined, it transforms as follows:

$$\begin{aligned} G_{x'} &= \tfrac{dx}{dx'}G_x + \tfrac{dy}{dx'}G_y \\ &\text{and} \\ G_{y'} &= \tfrac{dx}{dy'}G_x + \tfrac{dy}{dy'}G_y \end{aligned} \tag{5:8}$$

As before, we can rewrite these two equations as follows:

$$G_{a'} = \sum_{b=x,y} \frac{db}{da'}G_b \tag{5:9}$$

Note that we are using ratios like $\frac{db}{da'}$ rather than $\frac{da'}{db}$ (which we used earlier). That means that this is a different type of vector (because it transforms in a different way). The vector we discussed earlier (V) is called a contravariant vector, and the fact that it transforms as shown in Equation 5:5 is what defines it as that type of vector. The G vector is called a covariant vector, and it is defined as such because it transforms as shown in Equation 5:9. Usually, we express which type of vector we have by the way we denote its components. For contravariant vectors, we denote their components by putting their indexes (the x or the y) in superscripts:

V^x and V^y

While we denote the components of covariant vectors by putting their indices in subscripts:

G_x and G_y

With this notation, the two different transformations begin to take on an easy to remember form. See if you can figure out how the "upper" indices and the "lower" indices match up on both sides of the two transformation equations when they are written as follows:

$$V^{a'} = \sum_{b=x,y} \frac{da'}{db} V^b \tag{5:10}$$

and

$$G_{a'} = \sum_{b=x,y} \frac{db}{da'} G_b \tag{5:11}$$

Notice that the superscript (or subscript) on one side remains "upper" (or "lower") in the ratio on the other side. Also, note that the summation is always over the index which is repeated on the right side, once in an "upper" position and once in a "lower" position. This basic "formula" helps to produce equations for all transformation in tensor analyses (note this in the next part of this section).

It is interesting to note that in the normal spatial coordinates we are used to using (Cartesian coordinates), $\frac{db}{da'} = \frac{da'}{db}$, and there is no distinction between covariant and contravariant vectors. However, in other systems, the difference is there and must be considered.

Further, we note that with higher rank tensors, they are also defined by the way they transform from one coordinate system to another. For example, consider a second rank tensor, U. It could be that both of its indices are associated with the contravariant type of transformation (note: the following actually denotes four equations because $a'b'$ can be set to $x'x'$, $x'y'$, $y'x'$, or $y'y'$):

$$\begin{aligned} U^{a'b'} &= \tfrac{da'}{dx}\tfrac{db'}{dx}U^{xx} + \tfrac{da'}{dx}\tfrac{db'}{dy}U^{xy} + \tfrac{da'}{dy}\tfrac{db'}{dx}U^{yx} + \tfrac{da'}{dy}\tfrac{db'}{dy}U^{yy} \\ &= \sum_{\substack{\text{c and e vary over} \\ \text{all dimensions}}} \frac{da'}{dc}\frac{db'}{de}U^{ce} \end{aligned} \tag{5:12}$$

Or they could both be associated with the covariant type of transformation:

$$U_{a'b'} = \sum_{c,e} \frac{dc}{da'}\frac{de}{db'} U_{ce} \tag{5:13}$$

Or it could be a mix of the two:

$$U^{a'}_{b'} = \sum_{c,e} \frac{da'}{dc}\frac{de}{db'} U^c_e \tag{5:14}$$

Finally, we will see in the next section that any contravariant tensor also has a covariant form (and vice-versa), and we can transform from one form to the other if we know the geometry of the manifold on which the tensors are defined.

And that about ends our introduction to tensors. To sum up, they are geometric entities which have components denoted by some number of indices. Each index can be any of the dimensions in which the tensor is defined, and the number of indices needed to specify a component of a tensor is called the tensor's rank. We are familiar with 0 and 1 rank tensors (numbers–or "scalars"–and vectors). Finally, the way one transforms a tensor from one coordinate system to another depends on the type of tensor, and it (in fact) defines what it actually is to be a tensor. Each index of a tensor will transform in either a contravariant way or a covariant way.

These are the basic ideas behind tensors, and they allow us to define some very powerful mathematics. If you are familiar with the usefulness of vectors, then you have touched the surface of the usefulness of tensors in general. In the following section, we will look at two particular tensors, and we will see that they can be quite useful.

5.7 The Metric Tensor and the Stress-Energy Tensor

Now that we have had a glimpse at tensors, let's consider a couple that will be important to us. The first is called the metric tensor. I mentioned a couple of sections ago that this tensor is related to the invariant interval for a certain coordinate system on a given manifold. So, let's go back and look at a the two specific invariant intervals which we introduced. First, in normal, x-y, Cartesian coordinates, we have Equation 5:1 duplicated here:

$$ds^2 = dx^2 + dy^2 \tag{5:15}$$

Second, on the surface of a sphere, using the L-H coordinate system which we defined, we have Equation 5:2 duplicated here:

$$ds^2 = dH^2 + \cos^2\left(\frac{H}{R}\right) dL^2 \tag{5:16}$$

Now, let's make this more general by considering an arbitrary, two dimensional manifold and an arbitrary coordinate system on that manifold. Let's call the coordinates "a" and "b". Now, in general, the invariant interval on this manifold is defined in terms of the square of that interval ds^2. The equation for ds^2 involves the infinitesimal distances da and db in second order combinations. By second order combinations, I mean, for example, da^2 or $da\,db$. Thus, in general, the invariant interval will have the following form (note: the g components are generally formulas of "a" and "b"):

$$ds^2 = g_{aa}\,da^2 + g_{ab}\,da\,db + g_{ba}\,db\,da + g_{bb}\,db^2 \tag{5:17}$$

In that equation you see the four components of the metric tensor in this two dimensional, a-b coordinate system. They are the "g's" in the equation. For our x-y coordinate system, we have

$$g_{xx} = 1, \quad g_{xy} = 0, \quad g_{yx} = 0, \quad g_{yy} = 1 \tag{5:18}$$

For our L-H coordinate system, we have

$$g_{HH} = 1, \quad g_{HL} = 0, \quad g_{LH} = 0, \quad g_{LL} = \cos^2\left(\frac{H}{R}\right) \tag{5:19}$$

So, we can construct the invariant interval if we know the metric tensor for a coordinate system on a manifold. Now, remember that we said that the form of the invariant interval for a particular coordinate system tells us everything there is to know about the manifold for which those coordinates are valid. So, now we see that all we need to know is the form of the metric tensor. Once we know g, we know the geometry of the manifold. Using tensor analysis, we can take the metric tensor and find an equation for geodesics on the manifold. We can use it to find out all about the curvature of the manifold. We can even use it to find the dot product (we will discuss this a bit later) of two vectors in a particular coordinate system. Another thing the metric allows us to do is something generally called "raising" or "lowering" indices. Basically, if you consider a tensor with a contravariant index (which transforms in a particular way as discussed earlier), then there is another way to express the tensor as one which has a covariant index (and vice versa). That is to say that the geometric entity represented by the tensor with the contravariant index has another representation which involves a covariant index. For example, consider the tensor A^a, which has a contravariant index, a. There is a corresponding covariant tensor, A_a, which can be found using the metric of the space (and coordinate system) we are dealing with. Here is an example of how you find it (finding A_x when you know A^x) for a coordinate system with some arbitrary coordinates, x and y:

$$A_x = g_{xx}A^x + g_{xy}A^y \tag{5:20}$$

For a general space and coordinate system, you can write this rule as follows (remember, "a" can be any one dimension in the space, so this represents a number of equations):

$$A_a = \sum_{\substack{\text{b varies over}\\ \text{all dimensions}}} g_{ab} A^b \tag{5:21}$$

Similarly, if you know the covariant form of A (A_a) you can find the contravariant form by using the following:

$$A^a = \sum_{\substack{\text{b varies over}\\ \text{all dimensions}}} g^{ab} A_b \tag{5:22}$$

But that equation involves the contravariant form of the metric g^{ab}. In the invariant interval, the metric is expressed in its covariant form g_{ab}. It is therefore important for the reader to remember as we discuss various metrics below, that for all of them we have

$$g^{ab} = \begin{cases} \frac{1}{g_{ab}} & \text{if } a = b \\ 0 & \text{if } a \neq b \end{cases} \tag{5:23}$$

Thus, using the metric tensor, one can "raise" or "lower" any index of a tensor. Remember, what one is really doing is finding a form of that tensor which transforms in a different way.

With this example of how the metric can be used, we will end our discussion of this tensor. To sum up, the metric tensor on a manifold is a very important entity which not only tells us all about the manifold's geometry, but which also provides a very powerful tool which allows us to deal with that geometry mathematically.

The second tensor we want to mention is the stress-energy tensor. I don't want to get too deep into a discussion of the stress-energy tensor, but the reader should know a couple of key points. With the stress-energy tensor, we see our first example of a tensor explicitly defined in four dimensional space-time (though later we will look at the metric tensor defined in 4-d space-time). The stress-energy tensor (T) is also a tensor of rank 2 (like the metric tensor), which gives it 16 components in 4 dimensions. Sometimes we express such a tensor in the form of a matrix as follows:

$$T^{ab} = \begin{bmatrix} T^{tt} & T^{tx} & T^{ty} & T^{tz} \\ T^{xt} & T^{xx} & T^{xy} & T^{xz} \\ T^{yt} & T^{yx} & T^{yy} & T^{yz} \\ T^{zt} & T^{zx} & T^{zy} & T^{zz} \end{bmatrix} \tag{5:24}$$

There you can see the 16 different components. Now, each of these components tell us something about the distribution and "flow" of energy and momentum in a region. More precisely, T contains information about all the stresses and pressures and momenta in a region. For example, The "tt" component of the stress-energy tensor would be the density of the energy in the region (the amount of energy–including mass energy–per unit volume).

As to why the stress-energy tensor is important to us, that will be discussed further in a bit. However, here we can note the following in order to pull us back towards our discussion of relativity and gravity: In Newtonian physics, gravity was caused by the density of mass in an area. However, in SR we find that mass is just a form of energy, and so we might think that the "tt" component of the stress-energy tensor would be the right thing to look at when it comes to gravity. However, if we write a rule using one component of a tensor, then because the value of that component will depend on your coordinate system (or frame of reference in space-time) then the rule will also be frame-dependent. In short gravity would not be an invariant theory, and it would require a preferred frame if we based it only on the "tt" component of T. However, if we use all the components of a tensor to form our theory, then (as it turns out) the theory can be made frame-independent. Einstein thus considered the possibility that the whole stress-energy tensor would need to play a part as the source of gravity. Add to this some insight on curved manifolds and you end up with general relativity, as we will see.

5.8 Applying these Concepts to Gravity

Now that we have discussed manifolds and their properties along with some of the basic concepts of tensors, let's see how all of this applies to relativity and gravitation. First, I will go over the main ideas which lead us from what we have discussed so far to a general relativistic theory. After that, I want to mention a few notes on the physics and the mathematics we will be using given the concepts we have gone over. Next, we will go back and look again at special relativity while applying a bit of our new knowledge. This will show that GR is indeed general, because when applied to space-time without the presence of gravity it will explain a special case–special relativity. Finally, we will look quickly at a specific application of the GR concepts to a space-time in which there is a gravitational field. This application will focus on a particular class of stars and black holes.

5.8.1 The Basic Idea

Let's get started with the basic ideas which combine the concepts we have discussed to produce GR. Here I will simply state the main ideas without an explanation of their application. You will get some feel for their application in our two examples to follow.

So, here are the main claims of GR which involve the concepts we have discussed. First, the space-time in which we live is a four dimensional manifold. On that manifold there is a metric tensor (or just "a metric") which describes the geometry of space-time. The metric can be used to find geodesics on the space-time manifold, and when an object (only being acted on by gravity) goes from one point in space-time to another point in space-time (note: these are not just two points in space, but two points–i.e. events–in space-time), it moves between the points by following a space-time geodesic. Therefore, all the information necessary for us to determine how such objects move through space-time is held within the form of the metric. How, then, do we determine the metric? Well, the metric of space-time in a region is itself determined (in a not-too-trivial way) from the stress-energy tensor (T) which is affecting the region. This then is the new theory of gravity which relativity has produced. The stresses and pressures and momenta in a nearby region produces a stress-energy tensor which, in turn, changes the metric of the nearby space-time (making its geometry "curved"). This forces objects in the region to follow specific paths (geodesics) through the "curved" space-time, and we attribute this motion to gravitational effects.

As a conceptual example, consider a football being thrown from the surface of the earth. Because of the mass of the earth, the space-time the football is traveling through is a curved manifold, and the football follows a "straight line" geodesic in the four dimensional curved space-time. To us, the football's path is curved through three-space, but if we could somehow experience the time dimension as a spacial dimension (i.e. if we were four dimensional beings) and if we followed the path of the football in the four-space, we would seem to be following a straight line on our four dimensional curved manifold. However, in reality, the fourth dimension of time does not act like the other dimensions in our perception of the space-time manifold. Thus we do not see the actual four dimensional path of the football, we only see the path in three dimensions while the fourth component of the path is revealed to us as a dynamic component of the ball's motion through time. That's why we can't see that its path is a "straight line" in curved space-time. The stright-line is revealed to us as curved motion, and we attribute that motion to gravitational effects.

5.8.2 Some Notes on the Physics and the Math

Before we go on to our two examples, I wanted to mention a couple of points about the mathematics which can be used to develop physics in a particular space-time.

First, note that for any space-time there is a four dimensional metric involved. This metric can be used to find the invariant interval between two space-time points. That interval (recall) can generally be expressed as

$$ds^2 = \sum_{\substack{\text{a and b vary}\\ \text{over space and}\\ \text{time dimensions}}} g_{ab}\, da\, db \tag{5:25}$$

Second, consider a vector in our four dimensional space. Such a vector (usually called a four-vector) has four components, three relating to space and one relating to time. Now, in general, the values for these components will depend on the coordinate system/frame of reference in which you are considering the vector. However, we can use the metric to act on two four-vectors to produce an invariant number. In other words, if there are two four-vectors in a space-time, then two different observers using two different frames of reference will each find different x, y, z, and t coordinates which represent those two vectors in their respective frames. However, when they each act on those two vectors in a specific way using their own coordinate systems and using their own representation of the metric, they will each produce the same particular number. The action on the two vectors is called the dot product of the vectors, and many of you may have heard of and used it before (though perhaps you didn't realize you were using the metric–if you have ever had to remember how to produce a dot product in polar coordinates, then you have seen how the metric in that coordinate system affects the way you produce the dot product).

So, consider two four vectors, U and V. Remember that these are simply tensors with either contravariant or covariant components. Now, we can produce the dot product of U with V as follows.

$$U \cdot V = \sum_{a,b} g_{ab}\, U^a\, V^b \tag{5:26}$$

This produces a frame invariant number (a scalar), and if U and V have particular physical properties in space-time, then we can use the dot product to produce frame invariant physical rules in a particular space-time.

For our third note in this section, let's discuss the time between two events. It will be useful for us to find a frame-independent way of expressing that time. To explore this a bit, consider an observer who is not being acted on by any forces other than gravity. Because of gravity, he will simply follow a geodesic through space-time–being at certain points in space at particular times. Now, consider two events which each occur at the position of our observer, but which occur at two different times on our observer's clock. For such events, the time on the observer's clock which ticks off between the two events is called the "proper time" (T, though it is usually denoted using the Greek letter "tau") between those two events. The time this observer reads on his clock does not depend on what any other observer sees or does, and T is therefore a frame-invariant way of specifying a time between two such events. Of course, the time as measured in other frames will be different from T, but every frame will agree that for the one, unique observer who naturally follows space-time curvature to be at the position of both events, T is the proper time which he measures on his clock.

We should note that not all events can be connected by the natural space-time path of an observer because no observer can travel faster than light in that space-time. Any two events which can be connected by an observer's natural space-time path are called "time-like separated", and T can easily be defined for such events.

Now, consider the invariant interval for some observer's space-time path between two particular points. Remember that in general the invariant interval is a function of your position in space-time. Thus, as soon as you start moving down a path, the invariant interval begins to change. We discussed this fact briefly (see page 62) in Section 5.5 and decided that we would deal with it by breaking up the path into small bits and consider the invariant interval at each bit. Therefore, rather than discuss the entire interval between the two events, it is better to consider just one point along our observer's path and look the infinitesimal (ds) at that point. That infinitesimal in four dimensional space-time is generally made up of an infinitesimal change in space and an infinitesimal change in time. However, remember that for the observer and the two events we are considering, both of the events occur right at the observer's position. So, for him there is no spatial distance ($dx' = 0$, $dy' = 0$, and $dz' = 0$) between any two points on the path. Therefore, the invariant interval at any point on his path as calculated using his coordinates must be made up of only changes in his time coordinate (dt'). Thus, the value of the invariant interval at some point on the observer's path is given totally by the infinitesimal change in the proper time ($dT = dt'$, the infinitesimal change in time on our observer's watch). We can therefore write the following (taking the spatial components out of Equation 5:25):

$$ds^2 = g_{t't'}\, dT^2 \tag{5:27}$$

Notice that the component of the metric tensor in the above equation is expressed in the coordinates of the observer we are considering (i.e. we are specifically using t' and not t). This must be the case, because it is only when we measure the infinitesimal invariant interval (ds) using his coordinates that we can disregard any spatial component and write the interval totally in terms of dT. However, since this observer is free falling (only being acted on by gravity), then recall (see page 60) that his local space-time is flat, regardless of the global geometry of the space-time he is in. Thus, for small distances in space and time in his coordinate system (i.e. for infinitesimals like dt') his space-time can be considered to be that of special relativity (flat space-time). We will find out in the next section what g_{tt} is for the flat space-time of SR, and when we plug this into Equation 5:27 we will find that

$$dT^2 = \frac{-ds^2}{c^2}. \tag{5:28}$$

That equation is true for any space-time, because the space-time of the observer is locally flat regardless of the global geometry of the space-time we are considering.

So, how will this help us with the physics? Well, specifically, this gives us a way to define the momentum of an object in any space-time. Consider a free-falling object of mass m. In some coordinate system, the object's position in one coordinate (say "a") can be changing. Note that "a" could be x in an x-y-z coordinate system, r in polar coordinates (which we will discuss later), etc. Now, as the object changes spatial coordinates in this system, it will follow a natural geodesic path through space-time. As the object's position in "a" changes by some infinitesimal amount (da) its own "clock" will tick off some small time (dT–note that this is a proper time because it is measured on the clock of the object itself). In that case, the "a" component of the momentum for that object in this coordinate system will be expressed as

$$p^a = m\,\frac{da}{dT} \tag{5:29}$$

Notice that if we consider the situation where "a" is the time coordinate itself in our system, then we have a sort of "temporal momentum" who's significance will be discussed in the next section. Thus, p^a actually has four dimensions, and is, in fact, a four-vector. Combine this with our discussion of four-vectors above, and we will find some useful physics, as we will see in the following examples.

5.8.3 First Example: Back to SR

The most simple application of the ideas expressed in Section 5.8.2 is one which we have already looked at (though without using the concepts discussed in that section). It is the situation where there is no gravitational field. That is exactly the situation we were considering when we discussed special relativity. In special relativity, there is no gravitational field. All the components of the stress-energy tensor are identically zero.

Now, we will figure out the metric of space-time in such a case by examining what we already know about special relativity. So, let's go back to our space-time diagrams. (By the way, our diagrams only considered one of the spatial dimensions, but we will incorporate the other two in this section.) Consider two observers who start out moving parallel to one another on the diagram. This would mean that they start out with the same velocity in any inertial frame. Well, in special relativity (with no gravitational field) the two observers will continue to remain on parallel paths on the space-time diagram. This is the property of a flat manifold, so in SR, space-time is "flat".

Before we go on, it will be helpful for us to redefine the time variable in our space-time coordinates. Instead of "t", consider the combination "ct" (where c is the speed of light). For convenience, we will simply define a new variable, w, where

$$w = ct \tag{5:30}$$

Then we can use w in place of t in our coordinates. This is actually a fairly natural substitution in a couple of ways: First, note that w has the units of length, just like x, y, and z do. Second, using w on our space-time diagrams makes them a little more general. Why? Well, remember how we defined the units of length and time to be the light-second and the second? We did this so that a light ray would make a line at a 45 degree angle on our diagram. Well, with a w-x coordinate system, this will automatically be the case, regardless of what units you use. To see this, note that the value of t at a certain value of w is just the time it takes for light to travel that length, w (because $t = w/c$). For example, the point $x = 1$ light-second and $t = 1$ second corresponds to the point $x = 1$ light-second and $w = 1$ light-second. So, on both an x-t diagram and on an x-w diagram, a light beam would make a 45 degree angle with the x axis by going through the point (1,1). However, if we wanted to, we could now use a meter as our unit of length. Then, when $w = 1$ meter, t would just be the time it takes for light to travel 1 meter. So, the point $x = 1$ meter, $w = 1$ meter also lies on the light path, and again, that light path would automatically make a 45 degree angle with the x axis by going through the point (1,1). For consistency, we will continue to use units of seconds and light-seconds, but we will now use "w" in units of light-seconds to indicate time in our discussions and diagram (remember, the length "w" just represents the time it takes light to travel that length).

Now, let's look at a change in coordinates on the flat space-time of SR. In space-time, a change in coordinates can represent a change in an observer's frame of reference. So, when we discussed two observers who were moving with respect to one another, we were looking at two different coordinate systems (x-t and x'-t', or now, x-w and x'-w') which both correctly described space-time in SR. This leads us to consider the invariant interval, because we know it must be the same for each of these two coordinate systems. So, let's take a closer look at these coordinate systems on our diagrams and see if we can't define the invariant interval (which, remember, is just another way of writing the metric).

We will specifically want to consider infinitesimal lengths like dx. So, let's look at a small line segment which lies on a particular geodesic–a geodesic we know a little about. That geodesic is the path which light follows. Like anything else being acted on only by gravity, light must follow a geodesic on the space-time manifold. So, for the particular case of a light path, a small segment on that path would have an x component (dx) and a t component (dt); however, we now want to begin thinking of w as the unit which represents time, so we note that a small change in t (dt) represents a change in w of $dw = c\,dt$. Now, since the small distance light travels (dx) divided by the time (dt) it took it to travel that distance is defined as the speed of light, then we have the following:

$$\frac{dx}{dt} = c \quad \text{(where c is the speed of light)} \tag{5:31}$$

which can be rewritten as

$$\frac{dx}{dw} = 1 \tag{5:32}$$

That means that $dx = dw$ (for light). Now, since we always define the invariant interval in terms of the infinitesimal lengths squared, we will actually want to square both sides of that equation and then bring everything to one side so as to get the following:

$$dx^2 - dw^2 = 0 \quad \text{(For light)} \tag{5:33}$$

Now, because the speed of light is the same for all inertial observers, the above equation must be true for all frames of reference. Thus, we might consider the idea that the invariant interval for any small line segment (not just for light) is given in SR by

$$ds^2 = dx^2 - dw^2, \tag{5:34}$$

and this turns out to be the case. The light path, then, is just the case where $ds^2 = 0$.

Now, let's note a few things about this interval. First, it is independent of where you are in space-time. All that matters is the lengths dx and dw, regardless of what actual x and w position you have. This means that the distances (like dx) don't have to be infinitesimal, because the equation remains true regardless of how far you extend dx and dw. Thus, let's consider the case where one side of the line segment is at $x = w = 0$ (the origin). Then dx will be the x distance from the origin to the end of the line segment (which in this case can be as far away as we like), and dw will be the w distance to that point. In other words, for SR, dx and dw can be replaced with x and w when we consider one side of the line segment to be at the origin. Further, consider a point in space-time with coordinates (x, w) in the o observer's coordinates and (x', w') in the o' observer's coordinates. Since the value of the invariant interval is the same for any frame of reference, the following must be true:

$$x^2 - w^2 = x'^2 - w'^2 \tag{5:35}$$

Let's see that this is the case on our space-time diagrams. Diagram 5-9 shows a space-time diagram with two coordinate systems indicated, one for an observer o, and a second for an observer (o') moving with velocity 0.6 c with respect to o. (Note that now we use $w = ct$ for the time axes.) There is also a point marked "*" on the diagram. The x'-w' coordinates for that point are clearly shown to be $x' = 1$ light-second and $w' = 2$ light-seconds (i.e. $t' = 2$ second, remember?). The x-w coordinates are $x = 2.75$ light-seconds and $w = 3.25$ light-seconds. [!ht]

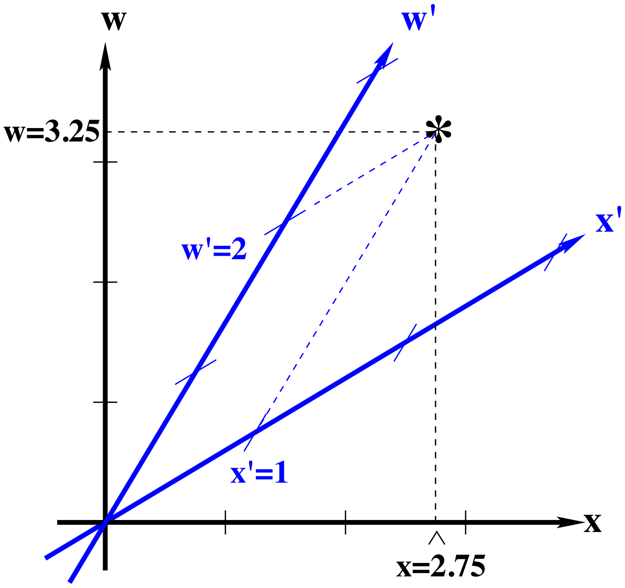

Diagram 5-9:

We therefore find the following:

$$\begin{aligned} ds^2 &= x^2 - w^2 = (2.75)^2 - (3.25)^2 \\ &= -3 \text{ light-seconds}^2 \\ \text{and} & \\ ds'^2 &= x'^2 - w'^2 = (1)^2 - (2)^2 \\ &= -3 \text{ light-seconds}^2 \end{aligned} \tag{5:36}$$

There are a couple notes to make about this outcome. First, of course, we note that $ds^2 = ds'^2$, as it must be. In fact, it is the form of the invariant interval and the fact that it must be invariant from one coordinate system to another that causes the transformation from x-w to $x' - w'$ to look as it does. If the x' and w' axes didn't look the way they do relative to the x and w axes in our diagrams, then the interval would not be invariant. Note that if the "-" sign in the invariant interval were a "+" sign, then the invariant interval would look just like the one for a normal, space-only x-y coordinate system where $ds^2 = dx^2 + dy^2$. Then, the coordinate transformation to x'-w' would be just like a rotation of coordinates (see Diagram 5-10). The "-" sign in the SR interval causes one of the axes to rotate in the opposite direction from the other when we do our space-time coordinate transformation.

Second, note that the interval squared is, in fact, negative. This is not too distressing, because we know that physical lengths on our diagram do not represent the space-time "lengths" which the invariant interval gives us. If they did, then the invariant interval for special relativity would be just like the x-y form of the invariant interval (since the physical lengths on our diagrams are just normal lengths on the flat paper/screen we draw them on). Now, the actual length of an infinitesimal interval on a manifold is usually defined to be the square root of the *absolute value* of ds^2. Thus, we can still make sense of lengths, even when the invariant interval squared is negative. [!ht]

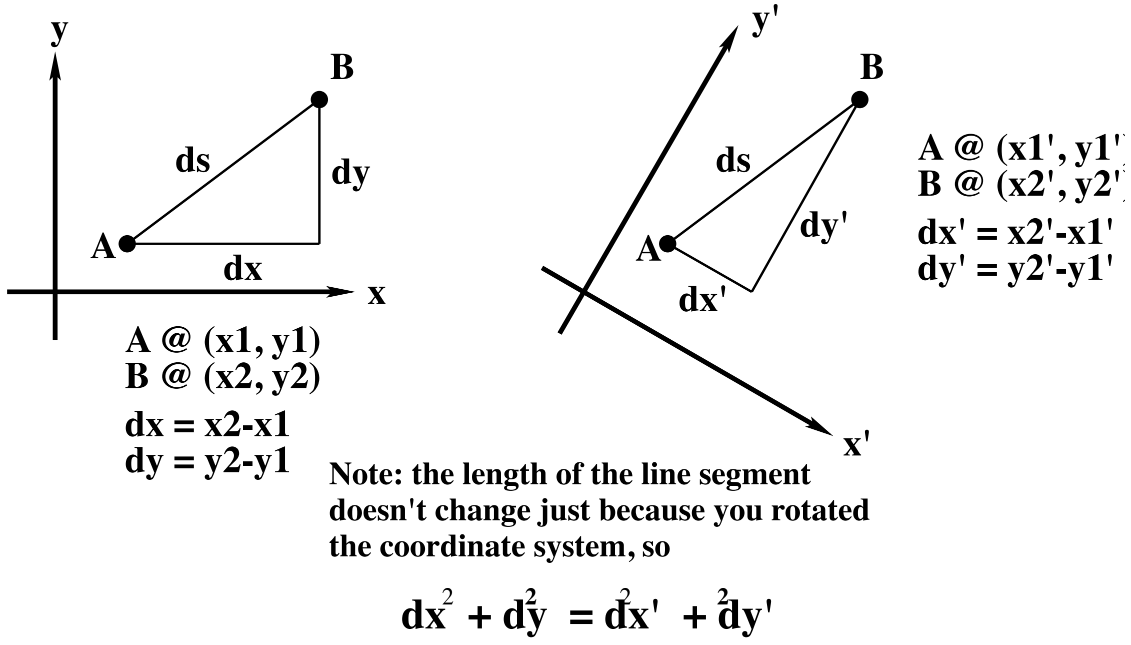
y
B
ds
dy
A
dx
x
A @ (x1, y1)
B @ (x2, y2)
dx = x2-x1
dy = y2-y1
y'
B
ds
dy'
A
dx'
x'
A @ (x1', y1'
B @ (x2', y2'
dx' = x2'-x1'
dy' = y2'-y1'
Note: the length of the line segment doesn't change just because you rotated the coordinate system, so
dx^2 + dy^2 = dx'^2 + dy'^2

Diagram 5-10:

The reader may have noted that thus far in our look back at special relativity we have still only included two of the four dimensions of space-time. The other two (y and z) could actually replace x in any of our discussions, and so they play the same roll in the invariant interval as x does. Therefore, the total four dimensional invariant interval for special relativity is given by

$$ds^2 = dx^2 + dy^2 + dz^2 - dw^2 \tag{5:37}$$

Finally, let's talk about some physics in this space-time using the concepts discussed in the previous section. First, consider the proper time between two time-like separated events. Recall that we defined this time such that:

$$ds^2 = g_{tt}(\text{of SR})dT^2 \tag{5:38}$$

We now know that $g_{ww} = -1$ for SR from the above, so $g_{tt} = -c^2$ for SR. This is how we got Equation 5:28, which is duplicated here:

$$dT^2 = \frac{-ds^2}{c^2}. \tag{5:39}$$

in the previous section. However, since we are now working with w for our time coordinate, we should define $dW = cdT$, and rewrite Equation 5:39 as

$$dW^2 = -ds^2 \tag{5:40}$$

Now, let's consider the observer which followed the t' axes in Diagram 5-9 such that his velocity was $0.6c$. Consider the O observer's frame of reference, and note that if it takes O' a certain time (dw) to travel a certain distance (dx) in the O observer's coordinates, then it must be the case that $dx/dt = 0.6c$. So $dx/dw = 0.6$, or

$$dx = 0.6dw \tag{5:41}$$

This, then, is true all along the w' axes (the line that O' follows through the O observer's coordinate system). So, the invariant interval (considering only two dimensions once again) at any point along the w' axes must be given by the following (using Equation 5:37 with only x and w coordinates and substituting Equation 5:41):

$$\begin{aligned} ds^2 &= dx^2 - dw^2 \\ &= 0.6^2 dw^2 - dw^2 \\ &= -[1 - 0.6^2]\, dw^2 \end{aligned} \tag{5:42}$$

plugging this into Equation 5:40 we find that

$$dW^2 = [1 - 0.6^2]\, dw^2 \tag{5:43}$$

so,

$$dw = \frac{1}{\sqrt{1 - 0.6^2}}\, dW = \gamma\, dW \tag{5:44}$$

Since dW just represents an infinitesimal time as measured on our "moving" observer's clock, and dw an infinitesimal time measured on our clock, Equation 5:44 is just the equation which shows time-dilation effects in SR, and it was quickly derived using our new knowledge.

For another physics consideration, look at the momentum four-vector. We defined this earlier (Equation 5:29) and it is duplicated here:

$$p^a = m\,\frac{da}{dT} \tag{5:45}$$

Again, we want to use $dW = cdT$, and we thus find

$$p^a = mc\frac{da}{dW} \tag{5:46}$$

For us, we consider the situation where "a" is the x dimension. Then, $p^{x'}$ for the "moving" observer himself is zero (because all along the w' axes we have $dx' = 0$ by definition, i.e. he is not moving relative to himself). However, for the O observer (for whom the "moving" observer moves a distance dx in a time dw) we find the following from Equation 5:46 by substituting x for a(Note that from Equation 5:44 we can write $dW = dw/\gamma$, and we are substituting that here. We also use $dw = cdt$ and $v = dx/dt$ in this equation.):

$$\begin{aligned} p^x &= mc\tfrac{dx}{dw/\gamma} \\ &= \gamma\, mc\, \tfrac{dx}{dw} \\ &= \gamma\, m\, \tfrac{dx}{dt} \\ &= \gamma\, mv \end{aligned} \tag{5:47}$$

This is exactly the definition of the momentum we saw in our discussions of special relativity.

However, now we can also look at the time component of the momentum four-vector and figure out what it represents. Again we use Equation 5:46, but here we substitute w for x:

$$p^w = mc\frac{dw}{dw/\gamma} = \gamma\, mc \tag{5:48}$$

But this is just the energy we had defined in SR ($E = \gamma mc^2$) divided by c:

$$p^w = \frac{E}{c} \tag{5:49}$$

And so, we now know all about the components of the momentum four-vector of a particle: three are the spatial components of the momentum of the particle, and the time component represents the energy of the particle divided by c.

As a final bit of physics, consider the dot product (as defined in Equation 5:26) of the momentum four-vector with itself:

$$\begin{aligned} p \cdot p &= g_{ww}p^w p^w + g_{xx}p^x p^x \\ &= -[\tfrac{E}{c}]^2 + p^2 \end{aligned} \tag{5:50}$$

(Note that the total momentum of this observer is p^x, and so we write p^2 in the last line to mean the total momentum squared). Now, recall that the dot product is invariant, so that if any observer measures the energy and momentum of a particle and calculates the above equation in his frame of reference, he must find the same number that any other observer would find in any other frame of reference. This shouldn't come as too much of a surprise if we look back for a moment. Back when we discussed energy and momentum in special relativity, we found in Equation 1:7 that $E^2 = m^2c^4 + p^2c^2$. Thus, we find that the dot product in Equation 5:50 is simply equal to $-m^2c^2$. Since m and c are invariant (remember, m is the rest mass), we could have already known that the formula in Equation 5:50 would be invariant.

We have therefore been able to find all the major physics equations we saw in special relativity by simply apply some tensor analyses using the metric of flat space-time.

So, to sum up, we have found the following: For SR, where there is no gravitational field, space-time has the properties of a flat manifold. The invariant interval of a flat space-time manifold is given by the following:

$$ds^2 = dx^2 + dy^2 + dz^2 - dw^2 \tag{5:51}$$

That interval tells us all about the nature of space-time in SR. The fact that the contribution of the time component (dw) is negative where as the spatial components have positive contributions is what gives the coordinate transformation between different frames of reference its unique form. Thus, it is the negative sign which essentially causes time dilation and length contraction effects, and it is the fact that the speed of light is invariant which causes that sign to be negative.

5.8.4 Second Example: Stars and Black Holes

In this second example, we will briefly look at the description GR gives us for the gravitational field of certain stars. We will also take a look at one of the most widely publicized consequences of GR–black holes.

To make our discussion simpler, the types of stars we will be considering will be spherically symmetric. What does that mean? Well, consider an imaginary sphere with some radius. Place the center of that sphere at the center of the star. If the star is spherically symmetric, then the strength of the gravitational field everywhere on the surface of our imaginary sphere will be exactly the same. For example, a star who's density is spherically symmetric and which is not spinning would work.

Now, it will be helpful for us to discuss the space around the star in terms of spherical coordinates; therefore, I should make sure the reader knows what these coordinates are. Rather than using x, y, and z coordinates for the three dimensional space around the star, we will use r, a, and b coordinates, which I will define here. In Diagram 5-11 I have tried to draw (in three dimensions) an z-y-z coordinate system, and I have marked a point in space, *. There is a line segment drawn from the origin (o) to that point, and the lengths of the x, y, and z components of the line segment are the values for the x, y, and z coordinates of the point, *. These components have been indicated on the diagram using "dotted" lines. Now, note that there is one other dotted line which is not labeled. If you imagine a light shining down on our line segment, then the unlabeled dotted line would be the shadow that light produced on the x-y plane. It is called the projection of the line segment on the x-y plane, but let's just call it "the x-y component" for convenience. [!ht]

Now we can define the r-a-b coordinates for the point, $*$. First, the distance from the origin to the point (the length of the line segment) is the "r" coordinate as indicated on the diagram. Next, the angle between the z axes and the line segment is our "a" coordinate (though it is usually denoted by the Greek letter "theta" (θ)). It too is indicated on the diagram. Finally, there is the angle between x and the x-y component of the line segment. That angle is our "b" coordinate (though it is usually denoted by the Greek letter "phi" (ϕ)), and it is indicated on the diagram as well. Thus, with r-a-b coordinates as defined here, we can specify any point in three dimensional space.

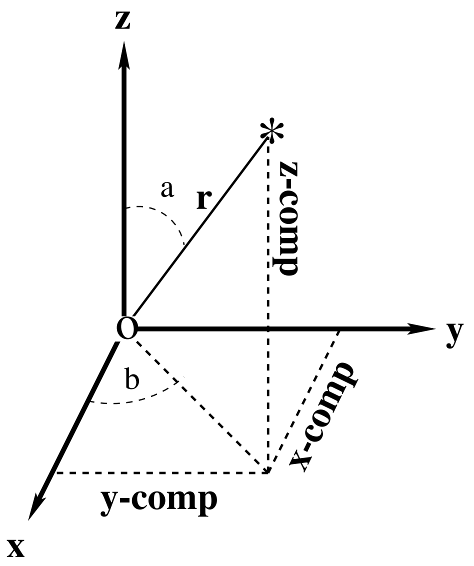

Diagram 5-11:

As a final note about this coordinate system, we should look at the metric of a flat 3-space using these coordinates. For an x-y-z system, the metric is (of course) given by this invariant interval:

$$ds^2 = dx^2 + dy^2 + dz^2. \tag{5:52}$$

However, for our new coordinate system in the same flat 3-space, it is given by the following:

$$ds^2 = dr^2 + r^2 da^2 + r^2 \sin^2(a) db^2. \tag{5:53}$$

For convenience, a new infinitesimal (call it du) is sometimes defined such that:

$$du^2 = da^2 + \sin^2(a)\, db^2. \tag{5:54}$$

Then we can rewrite Equation 5:53 as

$$ds^2 = dr^2 + r^2 du^2. \tag{5:55}$$

We will therefore continue to use du throughout this discussion, but remember it is just a convenient way to write the a and b components of the invariant interval.

Next, let's look at some properties of the star we will be considering. Basically, we will say it has a total mass of m(star) and a radius R. The center of the star will be centered at the origin, o. Finally, we will only be considering the gravitational field outside of the star itself. In general, physicists are interested in the gravitational field inside the star as well, but we will not worry about it that much.

We also want to define a new variable for mass using the Newtonian gravitational constant G. In Newtonian gravitation, the force between two objects of mass m1 and m2 which are a distance r apart is given by

$$F(\text{Newtonian Gravity}) = G\frac{m1\, m2}{r^2} \tag{5:56}$$

(where $G = 6.672 \times 10^{-11} m^3/(s^2 kg)$ and we note that kg is the symbol for kilogram). We will use G to define a new variable, M, such that

$$M = G\,\frac{m(\text{star})}{c^2} \tag{5:57}$$

Notice that M has the units of meters, and so M gives us a way of specifying the mass of an object in units of meters (similar to the way w allows us to specify time in units of meters). It is called the "geometrized" mass. So, using M we can say that an object has a mass of 1 meter, and one can decipher what mass we are talking about in terms of conventional units by using Equation 5:57. As a note, a mass of $M = 1$ meter corresponds to $m(\text{conventional}) = 1.35 \times 10^{27} kg$, the mass of the sun is $M(\text{sun}) = 1477$ meters $(1.989 \times 10^{30} kg)$, and the mass of the earth is $M(\text{earth}) = 0.004435$ meter $(5.973 \times 10^{24} kg)$.

Now, with this information in mind, the next step is to figure out what the metric of the space-time around the star would be because of the stress-energy tensor of the star. Generally, one uses the fact that we are considering spherically symmetric stars in order to make some assumptions about the form of the metric. One then uses this general form to calculate the general form the stress-energy tensor would have. Finally, one uses what we know physically about the star compared to the form of the stress-energy tensor, and one can decipher what equations must have made up the metric in the first place. In the end, one finds a metric for the space-time around this type of star, and for our purposes, we will simply state that end result. Thus, the metric is as follows (expressed in terms of the invariant interval):

$$\begin{aligned} ds^2 &= -\left(1 - \tfrac{2M}{r}\right) dw^2 + \left[\tfrac{1}{(1-\frac{2M}{r})}\right] dr^2 + r^2 du^2 \\ &= g_{ww} dw^2 + g_{rr} dr^2 + g_{uu} du^2 \end{aligned} \tag{5:58}$$

Note that we are using du as defined earlier, and we are using $dw = cdt$ as our time component as discussed in the previous section. Also, we are using M (as defined in Equation 5:57) to denote the mass of the star rather than $m(\text{star})$. This metric is known as the Schwarzschild metric.

The next step, then, is to show that we can get useful physics by considering this metric. We will again (as we did with the Special Relativity discussion earlier) be looking at a particle of mass m, and here we will be interested in its motion in the space-time around the star. Because of the spherical symmetry of the space-time, the motion of such a particle will remain within a plane, and we can orient our coordinate system so that the plane is one where the angle "a" = 90 degrees (and $\sin(a)= 1$). Since the particle doesn't move out of that plane, there is never a change in the angle "a" ($da = 0$). Thus, for this particle, we can consider the metric as follows (putting $\sin(a)= 1$ and $da = 0$ into Equation 5:58):

$$\begin{aligned} ds^2(\text{particle's path}) &= -\left(1 - 2\tfrac{M}{r}\right) dw^2 + \left[\tfrac{1}{(1-2M/r)}\right] dr^2 + r^2 db^2 \\ &= g_{ww} dw^2 + g_{rr} dr^2 + g_{bb} db^2 \end{aligned} \tag{5:59}$$

In the interest of time (because we simply haven't been able to cover everything we need to know about tensor analyses in this text), I will have to simply state a couple of facts which we will use to produce the physics we will look at. Namely, we notice that the form of the metric depends on your particular position in r (because g_{ww}, g_{rr}, and g_{bb} are all functions of r). However, none of the metric's components are functions of w. Because of that, as it turns out, p_w (the covariant form of the time component of the momentum four-vector) is constant throughout the motion of the particle. The metric is also independent of the angle b. This, as it turns out, implies that p_b is a constant. We can therefore define two constants, E and L such that

$$p_w = -Emc \tag{5:60}$$

and

$$p_b = Lmc \tag{5:61}$$

where m is the mass of the particle. These definitions will simplify the equations we will produce below (and they are related to our usual concepts of energy and angular momentum, so the fact that they are constant basically say that energy and angular momentum are conserved as the particle moves).

Now, so far we have only defined the contravariant form of the momentum, p^a. However, when we discussed the metric tensor we learned how to use it to "raise" and "lower" indices. So, we can write the following from Equation 5:22:

$$p^w = g^{ww} p_w + g^{wr} p_r + g^{wb} p_b + g^{wa} p_a \tag{5:62}$$

Note that we are considering the case where the angle "a" is a constant so that $p^a = 0$ in Equation 5:62. Also recall that in Equation 5:23 we noted how to go from contravariant to covariant forms of the metric. For the metrics we are discussing we thus have (note that the metric components come from Equation 5:59).

$$\begin{aligned} g^{ww} &= \tfrac{1}{g_{ww}} &&= \tfrac{-1}{1-\frac{2M}{r}} \\ g^{bb} &= \tfrac{1}{g_{rr}} &&= 1-\tfrac{2M}{r} \\ g^{bb} &= \tfrac{1}{g_{bb}} &&= \tfrac{1}{r^2} \\ &\text{all other covariant metric components} = 0 \end{aligned} \tag{5:63}$$

Thus, only the p_w part remains in Equation 5:62 giving us the following (note that I substitute using Equation 5:60):

$$p^w = \frac{-1}{1-\frac{2M}{r}} p_w = \frac{1}{1-\frac{2M}{r}} Emc \tag{5:64}$$

Similarly we can find the equation for p^b:

$$p^b = g^{bb} p_b = \frac{1}{r^2} p_b = \frac{1}{r^2} Lmc \tag{5:65}$$

Now, recall that in the last section we found that $p \cdot p$ was a constant, $-(mc)^2$. That remains true here, so we find the following:

$$p \cdot p = g_{ww} p^w p^w + g_{rr} p^r p^r + g_{bb} p^b p^b = -(mc)^2 \tag{5:66}$$

We can express each of the parts for that equation by substituting in the metric components from Equation 5:59, using the above equations for p^w and p^b, and writing p^r as $mc\frac{dr}{dW}$ to get the following:

$$\begin{aligned} g_{ww} p^w p^w &= -\left(1-\tfrac{2M}{r}\right)\left[\tfrac{(Emc)^2}{\left(1-\frac{2M}{r}\right)^2}\right] \\ &= \tfrac{-E^2(mc)^2}{1-\frac{2M}{r}} \\ g_{rr} p^r p^r &= \tfrac{1}{1-\frac{2M}{r}}\left[m\tfrac{dr}{dw}\right]^2 \quad \left(\text{note:} \tfrac{dr}{dw} = c\tfrac{dr}{dT}\right) \\ &= \tfrac{\left(\frac{dr}{dT}\right)^2 (mc)^2}{1-\frac{2M}{r}} \\ g_{bb} p^b p^b &= r^2 \tfrac{(Lmc)^2}{r^4} \\ &= \tfrac{L^2 (mc)^2}{r^2} \end{aligned} \tag{5:67}$$

Substitute this into Equation 5:66 and the $(mc)^2$ portions will cancel out on both sides giving this:

$$-1 = \frac{-E^2}{1-\frac{2M}{r}} + \frac{\left(\frac{dr}{dT}\right)^2}{1-\frac{2M}{r}} + \frac{L^2}{r^2} \tag{5:68}$$

From this, we can find the following equation which describes the orbits the particle can take. It is the equation of motion of the particle:

$$\left(\frac{dr}{dT}\right)^2 = E^2 - \left(1-\frac{2M}{r}\right)\left(1+\frac{L^2}{r^2}\right) \tag{5:69}$$

Now, it turns out that if one examine this equation for the case of a circular orbit (where r is a constant and $dr = 0$) and for the case where the mass is small or the orbit is large, we find things to be quite similar to what Newtonian physics predicts. However, it is interesting to note that for orbits for which r can change (elliptical orbits in Newtonian physics) GR predicts something a bit different from Newtonian physics. Basically, in Newtonian physics, the path of the particle in space is a true, closed ellipse. However, with the above equation one finds that the "elliptical" orbit in GR does not close in on itself. Instead, it's as if the ellipse changes position as the particle's orbit goes on. We thus see a difference in the predictions of the two theories, and we will mention this again in the next section.

With this quick look at the physics one can derive using the metric for such a star, we now want to go on and look at a very special case where this metric comes into play. Consider for a moment what would happen if the star's radius were to somehow become smaller than $2M$. Such a thing can theoretically happen for certain stars at the end of their life cycle, (though we won't get into how in our discussion).

So, consider the case where the radius of the star is smaller than $2M$. We can then consider a point above the star for which $r < 2M$. Now look back at the metric of the star. If $r < 2M$ then g_{tt} becomes positive, while g_{rr} becomes negative. That is to say that the time component of the invariant interval will contribute to the interval in the same way that a space-like coordinate did when r was greater than $2M$, and the radial component will contribute in the same way as a time-like coordinate did when r was greater than $2M$. Further, when r was greater than $2M$, we understood that all particles followed a space-time path which took them "forward" in time. Similarly, when g_{rr} becomes negative and d_{tt} becomes positive, (when $r < 2M$) we find that all particles must continue along a space-time path for which r continually decreases. In other words, the point $r = 0$ becomes part of the "future" of every particle/observer for which r is less than $2M$. Thus, such a particle will be doomed to fall in toward the center of the star. One can then imagine that the star itself would be doomed to fall in upon itself completely, becoming nothingness at $r = 0$.

This is known as a black hole (specifically, for the metric we are considering, it is a spherically symmetric black hole), and the radius $r = 2M$ is called the Schwarzschild radius or the event horizon. Any observer with an r coordinate less than $2M$ must fall into the point $r = 0$. Note that at $r = 0$ our metric becomes truly infinite, and as it turns out, that would be a point where physical laws break down. Such a point is called a singularity. We should also note that any signal (even a light signal) which the observer tries to send outside of the event horizon must also fall into the singularity (because all space-time geodesics for $r < 2M$ fall into the singularity). Thus, there is no way to get any information from the singularity to the "outside universe". There is no way for one to "see" the singularity and its destruction of physical laws. In that sense, the singularity's existence isn't a problem for our physical laws outside of the event horizon.

As a last consideration about black holes, one might ask what would happen to an observer who starts where his r coordinate is greater than $2M$ and then falls toward the event horizon. I won't go through the math, but one finds that in our coordinates, the observer will take an infinite amount of time to reach $r = 2M$. However, if we ask about how much time the observer himself reads on his watch as he falls (the proper time) we find that in his coordinates, the time it takes for him to reach the event horizon is finite. To try and understand how this can be, we will start by considering the equation for p^w (the time component of the momentum four-vector) as defined in Equation 5:46:

$$p^w = mc\frac{dw}{dW} \tag{5:70}$$

However, if we look back at Equation 5:64, we can combine it with Equation 5:70 to find the following:

$$\frac{dw}{dW} = \frac{E}{1 - \frac{2M}{r}} \tag{5:71}$$

Rewriting this, one finds that

$$dW = \frac{1 - \frac{2M}{r}}{E}dw. \tag{5:72}$$

So what does that tell us? Well, consider an observer at the coordinate position r. If a small time ticks in our coordinate $w = ct$, then the amount of time which ticks on the observer's clock ($dW = cdT$, where dT is the proper time) depends on the r position of the observer. The smaller his r position (as long as he is above the event horizon) the smaller dW will be for a given dw. This is similar to time dilation in SR, but here it is caused by the gravitational field and not by the relative motion of two observers.

Applying this to our discussion of the observer falling towards the event horizon, we find the following: In our coordinates (w) the clock of the infalling observer (who is constantly falling to smaller and smaller r values) takes longer and longer to tick its next tick. For example, let's say that for the observer's clock, it ticks 10 ticks before it reaches the event horizon. As we mentioned earlier, the coordinate time (w) will have to become infinitely large before the observer will reach the horizon. However, as the observer gets closer and closer to the event horizon, his clock takes longer and longer to tick its next tick. Essentially, in our coordinate system, the observer's clock will never be able to tick the 10th tick. Meanwhile, for the observer, time goes on as usual. For him, therefore, the 10th tick will come, and he will enter the event horizon. However, once in the horizon, he will not be able to send any signals out of the $r = 2M$ event horizon (in our coordinates). Thus, no one with r greater than $2M$ in our coordinates will ever be able to see the infalling observer go into the event horizon. This then explains how we can say that the infalling observer never reaches the horizon according to our coordinate system.

As it was in SR, there are different explanations for how certain outcomes come to be. The explanation depends on what coordinate system you use to explain the occurrences (which means that it depends on your frame of reference). The important point is that the end result of the explanations agree with the each other as far as any physical laws can be applied. In the twin paradox of SR, when the two twins come back together and stand next to one another at the end of the trip, each explanation must agree as to which twin is actually, physically older. For the question of whether an infalling observer reaches the event horizon, regardless of which coordinate system we use, we must agree that the observer is never seen to enter the horizon by any observer outside of the event horizon. The fact that the infalling observer "sees" himself enter the horizon has no physical consequences to the outside world.

Thus, with spherically symmetric stars and black holes, we have found the following: the metric of the surrounding space-time is given by the following (using variables we have defined earlier):

$$\begin{aligned} ds^2 &= -\left(1 - \tfrac{2M}{r}\right) dw^2 + \left[\tfrac{1}{(1-\frac{2M}{r})}\right] dr^2 + r^2 du^2 \\ &= g_{ww} dw^2 + g_{rr} dr^2 + g_{uu} du^2 \end{aligned} \qquad (5{:}73)$$

Symmetries in this metric can be used along with the metric itself to find the equations of motion for a particle which moves within this space-time. Finally, the space-time has interesting consequences for the measurement of space and time for observers at different points in the curved space-time surrounding such stars and black holes.

That ends our look at some examples of the application of GR. The only thing left in our discussion of this theory is to show some experimental evidence for its existence, as we will do in the following section.

5.9 Experimental Support for GR

In this section we will take a look at a few experiments which agree with the predictions of GR.

For the first experiment, we use the effect mentioned in the previous section whereby orbits which were supposed to be elliptical according to Newtonian physics didn't actually close in on themselves according to GR predictions. This effect can be seen as a rotation (or precession) of the "long axis" of the elliptical orbit, whereas under Newtonian theory, this axes doesn't move. Now, for the orbits of most planets, this effect is too small to measure. However, for Mercury (which is closest to the sun and would thus be the most affected) the effect is measurable. In fact, measurements taken during the 1800s showed that Mercury's orbit precessed. Now, much of this could be attributed to effects from the gravity of the other planets, however, after all those effects were taken into account, there was still a small amount of precession which wasn't accounted for. The predictions of GR accounted for the left-over difference. It was Einstein who first pointed this out, and this was the first evidence in favor of GR.

For the second experiment we want to consider, note that light, just like anything else being acted on only by gravity, must follow a geodesic in space-time. One can use the metric introduced in the previous section to figure out how light would travel when passing near an approximately spherically symmetric star. What one finds is that the light would be bent by the presence of the star's gravitational field. Now, one might try to make an argument using special relativity by which light with an energy E would be said to have a "relativistic mass" defined by "$m'' = E/c^2$. One could then figure out how much the light with this "mass" would bend in the presence of a Newtonian-type gravitational field. This, one might hope, could allow the explanation of how light could be bent without considering GR. However, one finds that the amount of bending predicted by this SR-Newtonian method is exactly half as much as the bending predicted by GR. Thus, if we could actually measure the bending of the light, we could figure out which of the two predictions was correct.

Well, experiments to measure such bending can and have been performed using the sun as the source of gravity and using light from particular stars–light which passes near the sun on its way to us–as the light that gets bent (it was Einstein who suggested this test, by the way). Normally, of course, the sun would be too bright to see stars who's light passes near the sun on its way to us. However, during a solar eclipse, the stars can be seen. When one compares the positions of such stars which one sees during a solar eclipse to the positions where the stars should actually be, one finds that the difference can be attributed to the bending of the light as predicted by GR, while the SR-Newtonian prediction was incorrect by a factor of 2.

The third experiment we will look at involves using highly sensitive atomic clocks taken aboard jets. When one compares the reading on such clocks to clocks which remained on the ground, one finds that the difference (though quite small) can only be accounted for completely if one includes calculations for SR effects and acceleration along with the GR effects of having the jet fly at high altitudes where the gravitational field is not as strong as it is on the surface of the earth.

These are a few examples of experimental evidence that exists in favor of GR. In many cases, more data and more precise measurements would be needed to rule out all theories other than GR; however, all the evidence we do have supports the theory.

Part IV

Faster Than Light Travel–Concepts and Their “Problems”

This is Part IV of the “Relativity and FTL Travel” FAQ. It discusses the various problems involved with FTL travel and how they apply to particular FTL concepts. This part of the FAQ is written under the assumption that the reader understands the concepts discussed in Part I of this FAQ which should be distributed with this document.

For more information about this FAQ (including copyright information and a table of contents for all parts of the FAQ), see the Introduction to the FAQ portion.

Chapter 6

Introduction to the FTL Discussion

The following discussion completes the purpose of this FAQ by considering faster than light travel with relativity in mind. After this brief introduction, I will discuss the general problems associated with FTL travel. These problems will apply differently to different FTL concepts, but I need to go over the general idea behind the problems first. After this general discussion of the problems, we will consider their applications to specific FTL concepts. We will also consider possible, conceptual "solutions" to the particular problem that seems to plague all FTL concepts. Finally, because this FAQ is written for the rec.arts.startrek.tech newsgroup, I will go over some notes and arguments for why "warp" drive should be explained in a particular way in order to get around the FTL problems and give us what is seen on the show.

6.1 A Few Notes On The Meaning of FTL Travel

Before we begin the discussion, I wanted to go over the basic idea of what we mean by FTL travel. To do so, we should start by noting that most of space-time through which we would want to travel is fairly flat. For those who have not read Part III of this FAQ, that means that special relativity describes the space-time fairly well without having resorting to general relativity (which applies when a gravitational field is present). Sources of gravity are few and far between, and even if you travel "close" to one, it would have to be a significant source of gravity in order to destroy our flat space-time approximation. Now, some FTL travel concepts we consider will involve using certain areas of space-time which are not flat (and I will go over them when we get there); however, the important thing for us is that all around these non-flat areas, the space-time can be approximated fairly well as being flat.

Thus, for our purposes, we can use the following to describe FTL travel. Consider some observer traveling from point A to point B. At the same time this observer leaves A, a light beam is sent out towards the destination, B. This light travels in the area of fairly flat space-time outside of any effects that might be caused by the method our observer uses to travel from A to B. If the observer ends up at B in time to see the light beam arrive, then the observer is said to have traveled "faster than light".

Notice that with this definition we don't care where the observer is when he or she does the traveling. Also, if some space-time distortion is used to drive the ship, then even if the ship itself doesn't move faster than light *within* that distortion, the ship still travels faster than the light which is going through the normal, flat space-time that is not effected by the ship's FTL drive. Thus, this ship still fits our definition of FTL travel.

So, with this basic definition in mind, let's take a look at the problems involved with FTL Travel.

Chapter 7

The First Problem: The Light Speed Barrier

In this section we discuss the first thing (and in some cases the only thing) that comes to mind for most people who consider the problem of faster than light travel. I call it the light speed barrier. As we will see by considering ideas discussed in Part I, Chapter 1 of this FAQ, light speed seems to be a giant, unreachable wall standing in our way. I note that various concepts for FTL travel may deal with this problem, but here we simply want to talk about the problem in general.

7.1 Effects as One Approaches the Speed of Light

To begin, consider two observers, A and B. Let A be here on Earth and be considered at rest for now. B will be speeding past A at a highly relativistic speed as he (B) heads towards some distant star. If B's speed is 80% that of light with respect to A, then γ for him (as defined in Section 1.4) is 1.6666666... = 1/0.6. So from A's frame of reference, B's clock is running slow and B's lengths in the direction of motion are shorter by a factor of 0.6. If B were traveling at $0.9c$, then this factor becomes about 0.436; and at $0.99c$, it is about 0.14. As the speed gets closer and closer to the speed of light, A will see B's clock slow down infinitesimally slow, and A will see B's lengths in the direction of motion becoming infinitesimally small.

In addition, If B's speed is $0.8c$ with respect to A, then A will see $B's$ energy as a factor of γ larger than his rest-mass energy (Note, I use an equation for energy here defined in Section 1.5, Equation 1:8):

$$E(\text{of B in A's frame}) = \gamma\, m(B)c^2 = 1.666[m(B)c^2] \tag{7:1}$$

where $m(B)$ is the mass of observer B. At $0.9c$ and $0.99c$ this factor is about 2.3 and 7.1 respectively. As the speed gets closer and closer to the speed of light, A will see B's Energy become infinitely large.

Obviously, from A's point of view, B will not be able to reach the speed of light without stopping his own time, shrinking to nothingness in the direction of motion, and taking on an infinite amount of energy.

Now let's look at the situation from B's point of view, so we will now consider him to be at rest. First, notice that the sun, the other planets, the nearby stars, etc. are not moving very relativistically with respect to the Earth; so we will consider all of these to be in the same frame of reference. Remember that to A, B is traveling past the earth and toward some nearby star. However, in B's frame of reference, the earth, the sun, the other star, etc. are the ones traveling at highly relativistic velocities with respect to him. So to him the clocks on Earth are running slow, the energy of all those objects becomes greater, and the distances between the objects in the direction of motion become smaller.

Let's consider the distance between the Earth and the star to which B is traveling. From B's point of view, as the speed gets closer and closer to that of light, this distance becomes infinitesimally small. So from his point of view, he can get to the star in practically no time. (This explains how A seems to think that B's clock is practically stopped during the whole trip when the velocity is almost c. B notices nothing odd about his own clock, but in his frame the distance he travels is quite small.) If (in B's frame) that distance shrinks to zero as his speed with respect to A goes to the speed of light, and he is thus able to get there instantaneously, then from B's point of view, c is the fastest possible speed.

From either point of view, it seems that the speed of light cannot be reached, much less exceeded. This, then, is the "light speed barrier", but most concepts people have in mind for producing FTL travel explicitly deal with this problem (as we will see). However, the next problem isn't generally as easy to get away with, and it probably isn't as well known among the average science fiction fan.

Chapter 8

The Second Problem: FTL, Causality, and Unsolvable Paradoxes

In this section we will explore a problem with FTL travel that doesn't always seem to get consideration. The problem involves ones ability to violate causality in certain frames of reference with the use of FTL travel. While this in itself doesn't necessarily make FTL travel impossible, the ability to go further and produce an unsolvable paradox would make the FTL travel prospect logically self contradictory. So, I will start by discussing the meaning of causality and the problems of an unsolvable paradox. I will then try to show how any form of FTL travel will produce violation of the causality principle. Finally, I will explain how, without special provisions being in place, FTL travel can go further to produce an unsolvable paradox.

8.1 What is Meant Here by Causality and Unsolvable Paradoxes

The principle of causality is fairly straight forward. According to causality, if there is some effect which is produced by some cause, then the cause must precede the effect. So, if for some observer (in some frame of reference) an effect truly happens before its cause occurs, then causality is violated for that observer. Now, recall our discussion in Section 1.1 concerning when occurrences happen in a frame of reference. There I took a moment to explain that when I talk about the order of events in some frame of reference, I mean their actual order, and not necessarily the order in which they are seen. One can imagine a situation whereby I could first receive light from the effect and later receive light from the cause. However, This might be because the effect is simply much closer to me than the cause (so that light takes less time to travel from the effect I observer, and I see it first). After I take into account the time it took the light to travel from each event, then I will find the order in which the events truly occurred, and this will determine whether or not there is a true violation of causality in my frame. This true violation of causality is what I will be talking about, **not** some trick concerning when observers *see* events, but a concept concerning the actual order of the events in some frame of reference.

Now, one can argue that the idea of causality violation doesn't necessarily destroy logic. The idea seems odd–to have an effect come first, and then have the cause occur–but it doesn't have to produce a self-contradictory situation. An unsolvable paradox, however, is a self-contradictory situation. It is a situation which logically forbids itself from being. Thus, when one shows that a particular set of circumstances allows for an unsolvable paradox, then one can argue that those circumstances must logically be impossible.

8.2 How FTL Travel Implies Violation of Causality

I refer you back to Diagram 2-9 (reproduced below as Diagram 8-1) so that I can demonstrate the causality problem involved with FTL travel. There you see two observers passing by one another. [!ht]

The origin marks the place and time where the two observers are right next to one another. The x' and t' axes are said to represent the frame of reference of O' (I'll use *Op*–for O-prime–so that I can easily indicate the possessive form of O as O's and the possessive form of O' as *Op*'s). The x and t axes are then the

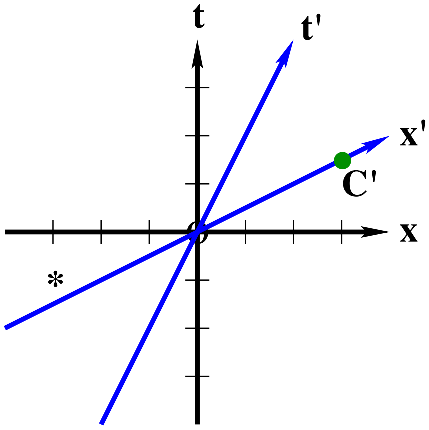

Diagram 8-1: (Copy of Diagram 2-9)

reference frame of the *O* observer. We consider the *O* system to be our rest system, while the *Op* observer passes by *O* at a relativistic speed. As you can see from the two coordinate systems, the two observers measure space and time in different ways. Now, consider again the event marked "*". Cover up the x and t axis and look only at the *Op* system. In this system, the event is above the x' axis. If the *Op* observer at the origin could look left and right and see all the way down his space axis instantaneously, then he would have to wait a while for the event "*" to occur. Now cover up the *Op* system and look only at the *O* system. In this system, the event is below the x axis. So to *O*, the event has already occurred by the time the two observers are passing one another.

Normally, this fact gives us no trouble. If you draw a light cone (as discussed in Section 2.8) through the origin, then the event will be outside of the light cone. As long as no signal can travel faster than the speed of light, then it will be impossible for either observer to know about or influence the event. So even though it is in one observer's past, he cannot know about it, and even though it is in the other observer's future, he cannot have an effect on it. This is how relativity saves its own self from violating causality.

However, consider the prospect of FTL travel with this diagram in mind. As *O* and *Op* pass by one another, the event "*" has not happened yet in *Op*'s frame of reference. Thus, if he can send an FTL signal fast enough, then he should be able to send a signal (from the origin) which could effect "*". However, in *O*'s frame, "*" has already occurred by the time *O* and *Op* pass by one another. This means that the event "*Op* sends out the signal which effects *" occurs after the event which it effects, "*", in *O*'s frame. For *O*, The effect precedes the cause. Thus, the signal which travels FTL in *Op*'s frame violates causality for *O*'s frame. Similarly, since "*" has already occurred in *O*'s frame when *O* and *Op* pass one another, then in his frame an FTL signal could be sent out from "*" which could reach *O* and tell him about the event as the two observer's past. However, for *Op*, the event "*O* learns about * as *O* and *Op* pass one another" comes before * itself. Thus, the signal which is FTL in *O*'s frame violates causality in *Op*'s frame.

In short, for any signal sent FTL in one frame of reference, another frame of reference can be found in which that signal actually traveled backwards in time, thus violating causality in that frame.

Notice that in this example I never mentioned anything about how the signal gets between the origin and *. I didn't even require that the signal be "in our universe" when it was "traveling" (remember our definition of FTL travel in Section 6.1). The only things I required were that (1) the signal's "sending" and "receiving" were events in our universe and (2) the space-time between the origin and "*" is flat (i.e. it is correctly described by special relativity diagrams). Some FTL ideas may invalidate the second assumption, but we will consider them a bit later. We will find, however, that violation of causality still follows from all the FTL travel concepts.

8.3 How We Get Unsolvable Paradoxes

As I mentioned before, violations of causality (as strange as they may be) do not have to truly, logically contradict themselves. However, it isn't too difficult to show (starting with the above arguments) that FTL travel can be used to produce an unsolvable paradox (a situation which contradicts its own existence). As a note, in the past I have called such situations "gross" violations of causality.

I'll illustrate the point with an example (again referring to Diagram 8-1) Remember we said that as *O* and *Op* pass, *Op* can send an FTL message out (from his frame of reference) which effects "*". However, rather than having him send a message out, let's say that *Op* sends out a bullet that travels faster than the speed of light. This bullet can go out and kill someone light-years away in only a few hours (for example) in *Op*'s frame of reference. So, say he fires this bullet just as he passes by *O*. Then the death of the victim can be the event (*). Now, in *O*'s frame of reference, the victim is already dead ("*" has occurred) when *Op* passes by. This means that another observer (stationary in *O*'s frame) who was at the position of the victim when the victim was shot could have sent an FTL signal just after the victim's death, and that signal could reach *O* before *Op* passed by him. So *O* can know that *Op* will shoot his gun as they pass each other.

To intensify the point I will make, we can let the signal which was sent to *O* be a picture of the victim, or even an ongoing video signal of the victim's body. Thus, *O* has evidence of the victim's death before *Op* has fired the weapon (a plain ol' violation of causality). However, at this point *O* can decide to stop *Op* from firing the gun. But if the bullet doesn't go out, and the victim never dies, then why (and how) would a video signal/picture of the victim's dead body ever be sent to *O*? And yet, *O* has that video/picture.

In the end, it is the death of the victim which causes *O* to prevent the victim's death, and that is a self contradicting situation. Thus, if there are no special provisions (which we will discuss later) FTL travel will not only allow violation of causality, but it can also produce unsolvable paradoxes.

At this point, I want to clearly list the various events which must happen to produce an unsolvable paradox in our "FTL bullet" example. Through the rest of our FTL discussion, this will be helpful as a reference listing.

Event Listing and Comments:

1. As observers *O* and *Op* pass by one another (as they are shown in Diagram 8-1) *Op* uses some method to send out an FTL bullet from his reference frame. The event "*O* and *Op* pass one another" will be called the "passing event" from here on.

2. The bullet strikes and kills a victim who's death is the event marked "*" in Diagram 8-1. This event occurs after the passing event in *Op*'s frame of reference, but it occurs *before* the passing event in *O*'s frame.

3. A third observer is at the victim's side as he dies and thus he witnesses the death. This third observer is stationary in *O*'s frame of reference (i.e. his frame is the same as *O*'s), so the victims death ("*") occurs *before* the passing event (when the bullet was fired) in this third observer's frame. Thus, the third observer has witnessed a result which comes from an event in his future–he has information about a future event in his frame of reference.

4. The third observer sends this information about the future to *O* using an FTL signal, and in the third observer's frame of reference, *O* can receive this information before the passing event occurs (and thus before the bullet is fired).

5. *O* receives the message and learns of the victims death before the bullet is fired. He thus knows about the bullet being fired–an event in his own future which will occur at his very location.

6. *O* uses this information to prevent *Op* from firing the bullet, thus causing a self-inconsistent situation–an unsolvable paradox.

It is important to note that the real crux of this problem does not come from the form of the FTL travel used, but from the relationship between the two, ordinary frames of reference for observers (O and *Op*) who never themselves travel FTL. This ordinary relationship (determined by relativity) can be demonstrated through experiments today, and as long as the exact same experiments can be performed in the future to yield the same results, then this argument must still hold. This is the power of this problem, and we will see that the special provisions we will discuss later must concern themselves with the ability of the observers to use the relationship between themselves in order to produce unsolvable paradoxes. Thus, the provisions will not be specifically concerned with the form of FTL travel used or the future theories which might suggest FTL travel, because the problem we have discussed here will be present regardless of either of these considerations.

And so, we have discussed the two problems which arise with FTL travel. Our next job is to consider various, specific FTL concepts in light of these problems. If your not interested in the discussion of the various forms of FTL travel, and you want to take my word for it that they will all suffer from the problem discussed above, then you may want to skip to the ”Special Provisions” section. I'll leave that to the reader.

Chapter 9

FTL Concepts with these Problems in Mind

Next, we want to ask about how one might try to get around these problems. Many of you have heard of ideas which get around the light speed barrier problem. For example, if we can do our traveling in some other, parallel “space”, then we won’t be bothered by the light speed barrier in our own space. However, these ideas have a much harder time getting around the second problem. In fact, to get around the second problem, we will see that special provisions will have to be made.

Therefore, the format of this discussion will involve the following. First, we will look at the various concepts which exist for possibly allowing FTL travel. I will show how each of them allows one to get around the light speed barrier problem, and I will explain how (without special provisions) none of them can bypass the second problem–producing unsolvable paradoxes. Finally, I will introduce some special provisions (beyond the basic assumptions made for the FTL concepts) and show how one can imagine using these provisions in conjunction with some of the FTL concepts to get around the second problem.

9.1 Tachyons (Without Special Provisions)

Tachyons are hypothetical/theoretical particles which would travel FTL. The concept of the tachyon attempts to get around the infinite energy requirements which the light speed barrier problem poses on a particle as it approaches the speed of light. This was accomplished by demanding that the particle have certain characteristics which we will discuss here.

First, consider the energy and momentum. Recall that we can write the energy (E) and the momentum (p) of a particle of mass m as expressed in Equation 1:8 and Equation 1:6 which are duplicated here:

$$E = \gamma\, mc^2 \qquad (9:1)$$

$$p = \gamma\, mv \qquad (9:2)$$

Where γ is defined in Equation 1:5 as $\gamma = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}$. From this we find that $|p\ c|/|E| = |v|/|c|$, which is greater than 1 if v is greater than c. We can thus write

$$E^2 < p^2c^2 \quad \text{(for an FTL particle)} \qquad (9:3)$$

But since we can also express the energy squared as defined in Equation 1:7:

$$E^2 = p^2c^2 + m^2c^4 \qquad (9:4)$$

we find that the only way to get $E^2 < p^2c^2$ is if the mass squared is negative (because then m^2c^4 reduces the sum in Equation 9:4). The mass would then be the square root of a negative number, and such an obviously unreal number is called an imaginary number (imaginary numbers may seem odd, but they have important uses in mathematics). In general we express such imaginary numbers as a product of a real number multiplied by something that symbolizes the imaginary square-root of negative one: $i = \sqrt{-1}$. So, the mass of a tachyon is imaginary. Further, from the equation for γ, we find that it too is imaginary if v is greater than c, but it is also negative because we have the i in the denominator of γ, and $1/i = -i$. (We can show this as follows: start with $1/i = 1/\sqrt{-1}$ and multiply and divide the right-hand side by $\sqrt{-1}$ (which doesn't change the value): $i = \frac{\sqrt{-1}}{\sqrt{-1}\sqrt{-1}}$. The top of that equation is just i, and the bottom is $\sqrt{-1}^2 = -1$. Thus $1/i = i/(-1) = -i$.) That would mean that from Equation 9:1, the energy would still be a real, positive number (because to get E we multiply the i in the imaginary m by the $-i$ in γ to get $-i^2 = -(\sqrt{-1}^2) = -(-1) = +1$). The same would be true for the momentum, $p = \gamma m v$.

I would like to note that I have read elsewhere that the energy would be negative for a tachyon, but this doesn't seem to be the case.

The final interesting property of tachyons I will mention comes from noting that as their velocity increases, the value of their γ will become a smaller, negative, imaginary number (because when $v/c > 1$, $1/\sqrt{1 - v^2/c^2}$ is a negative, imaginary number that decreases as v gets larger). That means that the value of a tachyons energy will decrease as the speed of the tachyon increases–or in other words, as the tachyon loses energy, it gains speed. One result of this is that if a charged tachyon were to exist, then because it would travel faster than light, it would give off a radiation known as Cherenkov radiation. This would take energy away from the tachyon and cause it to go faster and faster, continually giving off more and more energy. Neutral tachyons, however, wouldn't do this.

In any case, we can consider the possibility that tachyons exist and always travel faster than light. They then never have to cross the light speed barrier, and they do not have infinite energy (but their mass is imaginary and their energy decreases as their velocity increases). However, they still cause trouble because of the second problem–if you can use them for FTL communication, they can be used to create unsolvable paradoxes using the same arguments as we used in our "FTL bullet" example.

To explore the question of using tachyons for FTL communication, one can apply quantum mechanics to the energy equation of the tachyon. What one finds is that either (1) the tachyons cannot be localized, or (2) the actual effects of a tachyon cannot themselves move faster than light. In either of these cases, the tachyon cannot be used to produce an FTL signal.

A third idea would also allow the tachyon to exist without the possibility of using the tachyon to send FTL signals. The basic idea is that there would be no way to distinguish between the situation through which you could receive a tachyon and the situation though which you could transmit a tachyon. To show what I mean, consider Diagram 8-1 yet again. From the O frame of reference, a tachyon could be sent "from" * and "to" the origin. However, as long as you cannot distinguish between the transmitter and the receiver, then the Op observer could reinterpret this as a tachyon being sent "from" the origin "to" *. Neither, then, will believe that the tachyon went backwards in time. Obviously, there is no way for a message to be sent (because then you could identify the sender and decide which way the tachyon "really" went), and it wouldn't be quite right to call this FTL travel. However, it would allow tachyons to exist (though uselessly) without causing any problems.

And so, we find that with tachyons, one of the following must be true:

1. Tachyons do not exist,

2. Tachyons exist but cannot be used to send FTL signals, or

3. Tachyons exist and can be used to send FTL signals, but some special provision will keep anyone from using them to produce an unsolvable paradox.

9.2 Using a Special Field/Space/etc. (W/o Special Provisions)

This next concept is often found in FTL travel methods of science fiction. The basic idea is that a ship (for example) can use a special field or travel in another space/dimension in order to "leave" the physics of our universe and thus not be limited by the speed of light.

Again, we see that this concept is basically designed to get around the light speed barrier problem; however, it doesn't deal very well with the problem of producing unsolvable paradoxes.

Though the FTL observer or signal which travels using this concept would leave the realm of our physics, the relationship between two observers (like *O* and *Op*) who stayed behind (within the realm of our physics) would not be effected. This means (if you recall the points made earlier about the "second problem") that the arguments for producing an unsolvable paradox must still hold (unless there are special provisions), because those arguments were based on the relationship between the two observers who themselves never traveled FTL (and thus never left the realm of our physics).

Thus, we very quickly see that with any such methods (as long as no special provisions apply) one can produce an unsolvable paradox.

9.3 "Folding" Space (Without Special Provisions)

Another concept which pops into the minds of science fiction lovers when considering FTL travel is that of "folding" space. Basically, the idea is to bring two points in space closer together in some way so that you can travel between them quickly without having to "actually" travel faster than light. Of course, by our definition of FTL travel in Section 6.1 (where the light you are "racing" against goes through normal space between the starting and ending points) this would still be considered FTL travel.

A frequently used approach for picturing this idea is to think of two dimensions of space represented by a flat sheet of paper. Then consider yourself at some point on the paper (call this point "o"). If you want to travel to some distant point ("D"), you simply fold/bend/crumple/etc the paper and place "o" and "D" close to one another. Then its just a matter of traveling the now short distance between the points.

Again, we see an FTL concept which is built in order to get around the problem of the light speed barrier. However, we will see, once again, that the second problem of FTL travel is not so easily fixed.

We begin to understand this when we consider again the sheet of paper discussed above. Every object in that two dimensional space has a place on the paper. However, because objects may be moving, their position depends on the time at which you are considering them. Basically, if you are sitting at "o", you imagine every point on that sheet of paper as representing space as it is "right now" according to your frame of reference. However, as we have discussed, what is going on "right now" at a distant location **truly** depends on your frame of reference. Two observers at "o" in two different frames of reference will have two different ideas of what events should be represented on the paper as going on "right now". This difference in simultaneity between different frames of reference is what allowed for the "unsolvable paradox" problem to exist in the first place. Thus, even though you "fold" the paper so that you don't "actually" travel faster than light, you don't change the fact that you are connecting two events at distant points (your departure and your arrival) which in another frame of reference occur in the opposite order. (In the other frame of reference, you aren't just bending space, you're bending space-time such that you travel backwards in time.) It is that fact which allowed the unsolvable paradoxes to be produced.

In the end, unless special provisions are present, one can use this form of FTL travel in our FTL bullet example (I refer you back to the listing of events (see page 93) in Section 8.3). *Op* will fold space in his frame of reference to connect the passing event with the event "*", while the third observer will fold space from his frame of reference to connect the event "he sees the victim die" with an event "*O* learns of the victims death before the FTL bullet is sent". Thus, you can used this method to produce an unsolvable paradox as we discussed earlier.

9.4 Space-Time Manipulation (Without Special Provisions)

The final concept we will discuss before looking at special provisions is what I call space-time manipulation. The idea is to change the relationship between space and time in a particular region so that the limitation of light speed no longer applies. This is basically confined to the realm of general relativity (though the more simplified concept of "changing the speed of light" can also be handled by the arguments in this section). We won't worry too much about the particulars of how GR can be used to produce the necessary space-time, because the arguments that will be made will apply regardless of how you manipulate space-time in the region of interest.

There are two general types of space-time manipulation to consider. The first I will call "localized", because the space-time that is effected is that surrounding your ship (or whatever it is that is traveling FTL). A basic example of this is the idea for FTL travel is presented in a paper by Miguel Alcubierre of the University of Wales (the paper is available via the world wide web[1]). In the paper, Alcubierre describes a way of using "exotic matter" (matter with certain properties which may or may not exist) to change the space time around a ship via general relativity. This altered space-time around the ship not only keeps the ship's clock ticking just as it would have if the ship remained "stationary" (in its original frame of reference), but it also "drives" the ship to an arbitrarily fast speed (with respect to the original frame of reference of the ship before it activated the FTL drive).

The second type is thus "non-localized", and it involves the manipulation of space-time which at least effects the departure and arrival points in space-time (and perhaps effects all the space-time between). A basic example of this is the idea of a wormhole. A wormhole is another general relativity concept. Again, exotic matter is used, but here space-time is effected so that two distant locations in space are causally connected. You can enter one "mouth" of the wormhole and exit from the other very distant "mouth" so as to travel FTL (by our definition in Section 6.1).

Both of these concepts get around the light speed barrier problem, but again we will argue the case for the problems with unsolvable paradoxes. To do this, we will first carefully describe the situation in which a couple of FTL trips will occur. Let's call the starting point of the first trip "A". B will then be the destination point of that trip. Also, consider a point (C) which is some distance to the "right" of B ("right" being defined by an observer traveling from A to B), and finally consider a corresponding point (D) which is to the right of A. Diagram 9-1 uses two dimensions of space (no time is shown in this diagram) to depict the situation (at least from some particular frame of reference). [!ht]

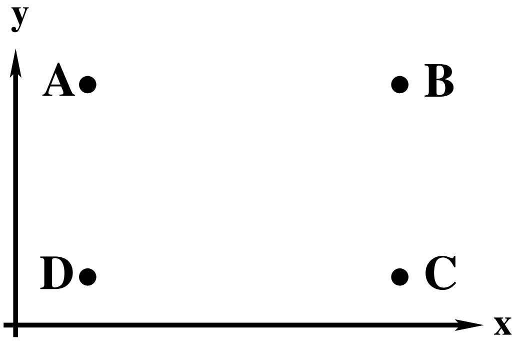

Diagram 9-1: (x and y are spatial dimensions)

Now, let's go back to the FTL bullet example through which we first explained the unsolvable paradox problem. In this case, the FTL bullet travels from A to B through space-time manipulation. (The event "the bullet leaves A" is event (1) in our list (see page 93) from Section 8.3). This means that all the space-time along the bullet's path between A and B might be affected by the space-time manipulation. Thus, we can no longer assume (after the bullet's trip) that a space-time diagram such as those we have drawn (which only apply to special relativity, not GR) will still apply. However, the space between D and C does not have to be effected by the FTL drive. Because of that we can make our argument by considering the following events:

a. *Op* sends an FTL bullet from A to B (using space-time manipulation) as the "passing event" occurs

b. The bullet strikes and kills a victim at B (event "*" in Diagram 8-1).

c. The third observer witnesses the death. However, now (because the FTL travel of the bullet may have changed the space-time between A and B) we can no longer assume that our space-time diagram of the situation is correct. It may be that with the changed space-time, this third observer's frame of reference no longer has the victim's death occurring before the passing event. However, we can continue as follows:

[1] http://arXiv.org/abs/gr-qc/0009013

d. The third observer sends a signal over to C using ordinary (slower-than-light) means.

e. An observer at C sends an FTL signal to D. Since the space-time between C and D need not be effected by the bullet's FTL travel, our space-time diagrams can be applied.

f. An observer at D receives the signal before event "a" (and thus before the bullet effected any space-time).

g. The observer at D can now send a signal over to *O*, and *O* can receive it before event "a" occurs.

The above events show that even though the space-time may be changed between A and B during the bullet's trip, the *O* observer can still know about and use the fact that the victim was killed in order to prevent the victims death. We use the same arguments we did in the section concerning the "second problem" (Section 9.1), except that the two FTL portions (the bullet and the signal from the third observer) are sent from two different locations so that neither is affected by the other's effects on space-time. Thus, as long as there are no special provisions, this form of FTL travel will still allow for unsolvable paradoxes.

9.5 Special Provisions

Thus far, we have seen that the second problem is not easily gotten around using any FTL concept. However, we have also insisted during our arguments that none of these FTL concepts include "special provisions". The specific provisions we were referring to will be discussed here. Basically, these are ideas which allow one to bypass the second problem in some way, and the ideas are generally not specific to any one form of FTL travel. They don't require that you bend space-time in some way or that you travel in some other universe or that you be made of some specific form of matter when you do your FTL traveling. What they do require is for the universe itself to have some particular property(ies) which, in conjunction with whatever form of FTL travel you use, will prevent unsolvable paradoxes.

There are four basic types of provisions, but we can express the general idea behind them all before we look at each one specifically. Recall that in producing the unsolvable paradox in our "FTL bullet" example, there was a series of events listed (see page 93), each of which had to occur to produced the paradox. The provisions simply require that at least one of these events be prevented from occurring. With the first and second provisions we will discuss, no restrictions necessarily have to be placed on the actual FTL travel, and any of the events (even those not directly dealing with the FTL travel) can be the "disallowed" event. The other two provisions place restrictions on the actual FTL travel in certain cases in order to prevent the unsolvable paradox.

9.5.1 Parallel Universes

In the first provision, one of the events in our list is not so much prevented as it is "transferred" to or from another (parallel) universe or reality. For example, say *O* has just received the information about the victim who dies at the "*" event, and *O* is waiting to stop *Op* from firing the FTL bullet. However, before he stops *Op*, he could find himself transferred to a parallel universe. In this universe he is able to stop *Op* from firing the bullet. The unsolvable paradox is resolved because the information about the death at "*" was not from the universe in which *O* stopped *Op*. Instead, *O* brought the information from a very similar parallel universe when he came over.

As another example, the bullet which killed the victim could have appeared from a parallel universe rather than being sent from *Op* in "our" universe. In this case, it is the "other universe bullet" which kills the victim. This bullet could seem to come from *Op* in our universe, though it actually came from an *Op* in the parallel universe. So, *O* is lead to believe that the bullet came from his own *Op*, and *O* stops *Op* from firing the FTL bullet. However, he doesn't prevent the death of the victim because the bullet which did the killing came from the "other universe *Op*". Again, the paradox is resolved.

Now, in that second case, the FTL bullet wasn't just performing FTL travel, but was involved with inter-dimensional travel. However, the second FTL signal in which the information is sent from the third observer to *O* (event number 4 in our list (see page 93)) was allowed. Thus, though this provision can effect the FTL trips, it doesn't have to forbid either of them.

In the end, as long as one of the events is forced to transfer to or from a parallel universe, there will be no unsolvable paradox (although why or how the inter-universe transfer would occur is left unanswered). Also, we should note that this provision could be applied with any of the FTL concepts we have discussed in order to allow them to exist without being self-inconsistent.

9.5.2 Consistency Protection

The second provision is what I am calling "consistency protection". The idea is that the universe contains some sort of built-in mechanism whereby some event in our list of events (see page 93) would not be allowed to occur.

An example of such a mechanism can be found when we look at the situation through quantum mechanics. (A theory of Steven Hawking called the "chronology protection conjecture" (CPC) attempts to do just that–the jury is still out on this theory, by the way, and will probably be out for a long time.) In quantum mechanics (QM), we do not think in certain terms of whether or not an event will occur in the future given everything we can possibly know about the present. Instead we consider the probability of an event (or string of events) occurring. One form of consistency protection would insist that QM prevents the unsolvable paradoxes because the probability of all the events occurring so as to produce an unsolvable paradox is identically zero.

Under this explanation using QM, our bullet example would be resolved through arguments similar to this: It may be that the *Op* observer is unable to produce the FTL bullet (perhaps his FTL gun fails), thus averting the paradox. If he is able to get the FTL bullet on its way, then perhaps the bullet will end up missing its mark. If it does hit the victim, then perhaps the victim's friend will be unable to send an FTL signal back to the *O* observer (perhaps his FTL message sender fails). If the signal to *O* gets sent, it still might not be received by *O*. If *O* receives it, he may be unable to stop *Op* from firing the bullet. In any case, this particular QM explanation would insist that one of these events must not occur, because the quantum mechanics involved forces the probability of all of the events occurring to be zero.

To sum up, this provision requires that some mechanism exists in the universe that would prevent at least one of the events from occurring so that the unsolvable paradox does not come about. This mechanism does not have to specifically target any of the FTL trips/messages which one might want to make/send, but it could disallow any of the events which must be present for the unsolvable paradox to occur. We should also note that this provision (just like the last) can be apply regardless of the FTL concept used.

9.5.3 "Producing" Restricted Space-Time Areas

This provision is sort of an extension on the previous one, but its mechanism specifically targets the FTL travel so as to restrict one of the FTL trips or messages one must use to produce an unsolvable paradox. Remember that in the list of events for our FTL bullet example, there were two different FTL portions (the FTL bullet and the FTL message from the third observer to O). This provision would cause the sending or receiving of one of these "messages" to strictly prohibit the sending or receiving of the other. I will try to illustrate the basic way in which such restrictions could work to always prevent unsolvable paradoxes. I will then give an example where this provision is implemented with a particular FTL concept.

For the illustration, we need to consider each of two possibilities within our FTL bullet example. In the first possibility, the *Op* observer is allowed to send his FTL bullet which strikes the victim, but that FTL trip must then restrict the third observer's ability to send the FTL message to *O*. In the second example, the third observer happens to decide to send some FTL signals to *O* at some point before the event "*" (which is the event in our example that usually marked the victim's death). Now, we let the third observer continue to send those FTL signals until some point after "*". Then, if the victim dies at "*" because of the FTL bullet, then since the third observer is sending FTL signals to *O* at that point, he would be able to tell *O* about the victim's death, and the paradox would still be possible. Thus, in this second case, the FTL bullet must not be allowed to strike the victim (the FTL travel of the bullet is restricted because the third observer sends FTL signals to O).

So, how would these restrictions work in these two possible cases? Well, as it turns out, if all unsolvable paradoxes are going to be averted while only placing restrictions on particular FTL trips, then there must be a very specific provision in place. To explain this, we will look at both possible situations, and consider diagrams which explain each one. (Note that these diagrams are drawn a little differently from Diagram 8-1 so as to better show the point I am trying to make here.) [!ht]

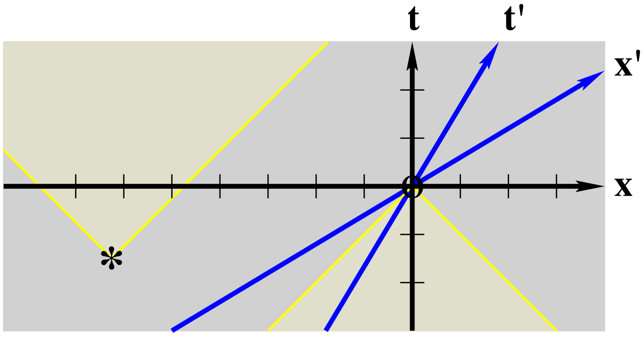

Diagram 9-2: (Case 1–The FTL bullet is allowed to strike at the event "*")

In this diagram we mean to illustrate case one in which the FTL bullet leaves the "passing event" (i.e. the origin, "o") and is "received" by the victim who immediately dies at event "*". Now, I have also drawn parts of two light cones (marked in yellow). One part is the "upper half light cone of the event '*'," and the other is the "lower half light cone of the passing event, 'o"'. The upper half light cone of "*" contains all events which an observer at "*" (like the third observer in our bullet example) can influence without having to travel FTL. All observers agree that all events in this area occur some time after "*" (as discussed in Section 2.8). Also, the lower half light cone of "o" contains all the events which could effect "o" (which, remember, is the event at which the FTL bullet is sent) through non-FTL means. Thus, as long as no FTL signal/traveler can leave as an event in the upper half light cone of "*" and be received as an event in the lower half light cone of "o", then *all* unsolvable paradoxes will be averted. There would be no way for the third observer to witness the death of the victim and afterwards get a signal to O before the bullet is fired.

Now, that seems to be straight forward. We just need to make this provision: When an FTL signal is transmitted as event T, and it is received as event R, then it must be impossible for any information to be sent as an event in R's upper ("future") light cone and end up being received as an event in T's lower ("past") light cone. If the universe restricted FTL travel in this way, it would be impossible to produce unsolvable paradoxes.

However, we can see that the matter can get a little complicated when we consider things from O's frame of reference (which is also the frame of the third observer). In this frame, after the third observer witnesses the victim's death at "*", the event "the bullet leaves" hasn't occurred yet. He might then argue that no FTL signal has yet been sent which would keep him from sending a FTL message to O. The problem with his argument is that he has already witnessed the result of the FTL bullet being sent (even if it hasn't occurred in his frame yet). Thus, any FTL signal he tries to send to O (in the lower half light cone of the origin/passing event/bullet-being-fired event) must be prevented from being received by O.

Ah, but what if he (the third observer) just happened to decide to start sending FTL signals to O (just to chat) before the bullet strikes the victim? That leads to our second case. Here, then, is a diagram we will use to describe this second case. [!ht]

Now, there are a few extra events here. The point "s" marks the point where the third observer starts sending FTL signals to O while "T" marks the point where he finishes sending those FTL signals. The point "R" marks the point where O receives the last message which was sent at "T". Now, here we have drawn the upper and lower half light cones of interest, and according to our discussion above, it would be impossible for Op to send his bullet at the origin, "o" (which is in the upper half light cone of R) and have it "received" by the victim at "*" (which is in the lower half light cone of T). So, according to that argument, the bullet doesn't strike while the third observer is sending FTL signals to O, and so the third observer never tells O about the victim's death.

However, this doesn't **have** to be what happens, and we might just end up back at the first case. You see, either (1) the signals sent by the third observer are all successful, and the FTL bullet is restricted from striking the victim at "*" (that's the second case); or (2) the FTL bullet does strike the victim at "*" and any FTL signals that the third observer sends after "*" are restricted from reaching the O observer before

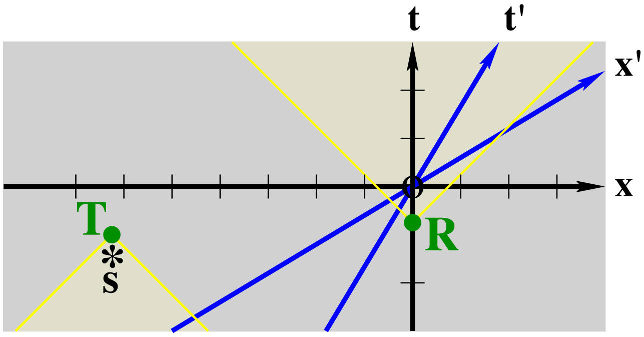

Diagram 9-3: (Case 2–The FTL bullet may not be allowed to strike at the event "*")

the bullet is fired (this is the first case, even though the third observer *was* sending signals to *O* just *before* the bullet hit). The obvious question, then, is "which one of these two cases actually occurs?" The answer happens to be, "it really doesn't matter." You see, as long as one or the other does occur, the situation remains self consistent and no self inconsistent paradoxes are produced. Roll some dice and pick one, if you like, or let some unknown force decide which happens. It really doesn't matter for our argument. Is that a bit odd? Yes. Is it self-inconsistent so as to produce unsolvable paradoxes? No.

Finally, as an example to show this provision in action with a particular FTL concept, let's consider a case where space-time manipulation is used via a wormhole. Recall that in our discussion of this FTL concept in Section 9.4, we showed that one can still produce unsolvable paradoxes. Notice, that there still must be two FTL parts (we discussed one FTL "trip"–the bullet–from A to B and another–an FTL message–from C to D). Now, to prevent the paradox, the existence of the wormhole that allows the bullet to travel from A to B could forbid the existence of the wormhole that allows the FTL message to go from C to D. This is a situation where case 1 applies, and here the way the provision is satisfied comes from the conceptual ability of one wormhole's existence to forbid the existence of another wormhole.

And so, we have a provision which simply restricts (in a very particular way) certain FTL trips because of other FTL trips. We have found that there doesn't have to be a discernible answer to the question of whether trip A disallows trip B or trip B disallows trip A, but as long as it is one case or the other, this provision will keep all situations self consistent and thus avoid unsolvable paradoxes.

9.5.4 A Special Frame of Reference for the purpose of FTL Travel

The fourth and final provision is (again) something of an extension to the previous one. This provision also forbids certain FTL signals, but it does so in a very specific and interesting way (there will be no question as to which trips are allowed and which are not). To explain this provision, I will start by describing a situation through which the provision could be applied. I will then explain how the provision works, given that particular situation.

Now, as I describe the situation, I will use the idea of a "special field" to implement the "special frame of reference". However, it isn't necessary to have such a special field to imagine having a special frame of reference. I am simply using this to produce a clear illustration.

So, join me now on a journey of the imagination. Picture, if you will, a particular area of space (a rather large area–say, a few cubic light-years if you like) which is permeated with some sort of field. Let this field have some very particular frame of reference. Now, in our imaginary future, say we discover this field, and a way is found to manipulate the very makeup (fabric, if you will) of this field. When this "warping" is done, it is found that the field has a very special property. An observer inside the warped area can travel at any speed he wishes with respect to the field, and his frame of reference will always be the same as that of the field. This means that the *x* and *t* axes in a space-time diagram for the observer will be the same as the ones for the special field, regardless of the observer's motion. In our discussion of relativity, we saw that in normal space, a traveler's frame of reference depends on his speed with respect to the things he is observing.

However, for a traveler in this warped space, this is no longer the case.

For example, consider two observers, A and B, who both start out stationary in the frame of reference of the field. Under normal circumstances, if A (who starts out next to B) began to travel with respect to B, then later turned around and returned to B, A would have aged less because of time dilation (this is fully explained in Section 4.1 of Part II if you are interested). However, if A uses the special property of this field we have introduced, his frame of reference will be the same as B's even while he is moving. Thus, there will be no time dilation effects, and A's clock will read the same as B's.

Now, for the provision we are discussing to work using this special field, we must require that all FTL travel be done while using this field's special property. How will that prevent unsolvable paradoxes? Well, to demonstrate how, let's go back to our FTL bullet example and consider one of two cases. In case 1, we will let *Op*'s frame of reference be the same as the frame of reference of our special field. With this in mind, let's go through the events listed (see page 93) in Section 8.3 once again; only this time, we will require any FTL travel to use the special property of the field we have discussed.

So, here is the new list of events given that the special frame of reference of the field is the same as *Op*'s frame. Remember, our new provision requires that any FTL trip will have to use the property of our special field, thus the object/person/message traveling FTL will be forced to take on the frame of reference of our special field (*Op*'s frame in this example). (It may be good for you to review the original list (see page 93) before reading this one):

1. Again (just as in our original argument), as observers *O* and *Op* pass by one another, *Op* uses some method to send out an FTL bullet. This time, as the FTL method is activated, our new provision requires the bullet's frame of reference to become the frame of reference of the special field. However, since *Op*'s frame is the same as that of the special field in the case we are considering, the bullet will still be sent out from *Op*'s frame of reference, just as it was in our original argument.

2. Again, the event marked "*" occurs after the "passing event" in *Op*'s frame, so again the bullet can travel FTL to strike and kill a victim at "*", and again that event occurs *before* the "passing event" in *O*s frame.

3. Again, a third observer (who is in *O*'s frame of reference) witnesses the victim's death, and again the death will have occured *before* the bullet was sent in his frame of reference. Thus again this third observer will have information about an event which will happen in his future.

But that is where the "agains" stop. You see, in the original argument event (4) was possible in which the third observer sends this information about the future to *O* via an FTL signal. In the frame of reference of *O* (and the third observer), that FTL signal could be sent after the victim's death and arrive at *O* before the passing event (when the bullet was fired). But now, as the FTL signal is sent, it must take on the frame of reference of the special field. That frame of reference is the frame of *Op*, and in that frame the victim dies *after* the bullet is fired. So, in the new reference frame of the message (forced on it by the provision we are making) the bullet has already been sent, and thus the FTL message cannot be received by *O* before the bullet is sent.

From the frame of reference of the third observer, he simply cannot get the FTL signal to go fast enough (in his frame) to get to *O* before the bullet is sent. From *Op*'s frame of reference (that of the special field) any FTL signal (even an instantaneous one) can theoretically be sent using our provision. However, from *O*'s frame (and that of the third observer) some FTL signals simply can't be sent (specifically, signals that would send information back in time in *Op*'s frame of reference–look again at Diagram 8-1 to make this clear). This prevents the unsolvable paradox.

We can also consider case 2 in which the special frame of reference of the field is the same as *O*'s frame of reference. In this case, any FTL traveler/signal/etc must take on *O*'s frame of reference as it begins its FTL trip. Thus, as *Op* passes *O* and tries to send the FTL bullet from his frame of reference, the bullet will have to take on *O*'s frame as it begins is FTL trip. But in *O*'s frame of reference, the event "*" has already occurred by the time *O* and *Op* pass one another. Therefore, from the FTL bullet's new frame of reference (forced on it by the provision we are making), it cannot kill the victim at the event "*" since that event has already occurred in this frame. Thus, the paradox is obviously averted in this second case as well because of our provision.

So, in the end, if all FTL travelers/etc are required to take on a specific frame of reference when they begin their FTL trip, then there will be no way an unsolvable paradox can be produced. This is because it takes two different FTL trips from two different frames of reference to produce the paradox. Under this provision, if you are sending tachyons, the tachyons must only travel FTL in the special frame of reference. If you are folding space, the folding must be done in the special frame of reference. If you are using the special field itself to allow FTL travel, then you must take on the field's frame of reference. Etc. If these are the cases, then there will be no way to produce an unsolvable paradox using any of the FTL concepts.

As a final note about this provision, we should realize that it does seem to directly contradict the idea of relativity because one particular frame of reference is given a special place in the universe. However, we are talking about FTL travel, and many FTL concepts "get around" relativity just to allow the FTL travel in the first place. Further, the special frame doesn't necessarily have to apply to any physics we know about today. All the physics we have today could still be completely relativistic. In our example, it is a special field that actually has a special place in the physics of FTL travel, and that field just happens to have some particular frame of reference. Thus, the special frame does not have to be "embedded" in the makeup of the universe, but it can be connected to something else which just happens to make that frame "special" for the specific purpose of FTL travel.

And so, we have seen the four provisions which would allow for the possibility of FTL travel without producing unsolvable paradoxes. For the case of the real world, there is no knowing which (if any) of the provisions are truly the case. For the purposes of science fiction, one may favor one of the provisions over the others, depending on the story one wishes to tell.

Chapter 10

Some Comments on FTL Travel in Star Trek

Since this document was originally created for the rec.arts.startrek.tech newsgroup, it seems appropriate to take all we have discussed and apply it to what we see in Star Trek. Of course, it would be foolish to assume (unfortunately) that the writers for the show take the time to learn as much about these concepts as we now know, and I am certainly not implying that a conscious effort was made to incorporate what we know to be true in a consistent way on the show (after all, this is Star Trek ☺). However, interestingly enough, if we apply the concepts correctly, we can explain most of what Star Trek has shown us. That is what I will try to do here.

10.1 Which Provision is Best for Explaining Warp Travel

First, we might want to consider the four provisions and try to decide which one would best fit Trek so that everyday warp travel couldn't be used to produce unsolvable paradoxes.

So, let's consider both the first and second provisions. In these cases, neither of the two FTL trips in our FTL bullet example will necessarily be forbidden. So, if we consider that example yet again, we can make the following argument: Let *Op* be the Enterprise. Then, rather than sending a bullet, the Enterprise could itself travel from the origin to "*". It could then (through ordinary acceleration) change its frame of reference to match *O*'s. Then it could travel from "*" (or just after "*"–we have to give them a little time to do their acceleration) back to the *O* observer, and it could get to *O* **before** it ever left for its first FTL trip (i.e. we put the Enterprise in place of the FTL signal sent by the third observer). Thus, since neither the first or second provision has to forbid any of these actions, the Enterprise could use everyday warp travel via this method to easily travel back in time without having to do something as dangerous as zipping around the sun (as they have had to do on the show).

In addition, if the first provision governed normal warp travel, then making different trips from different frames of reference would introduce the possibility that you would find yourself being transferred to another parallel universe to prevent unsolvable paradoxes. Also, if the second provision governed normal warp travel, it would require Star Trek ships to be careful as to which frames of reference they were in when they decided to enter warp. After all, they may not want to accidentally meet themselves from a previous trip (in which case the universe may destroy them to protect self consistency). So, there seems to be some daunting arguments against using either the first or second provision to keep ordinary warp travel from producing unsolvable paradoxes in Trek.

Okay, what about the third provision? With that provision it would be impossible to use ordinary warp travel as a "time machine". However, this provision does cause certain noticeable restrictions on some FTL trips (remember, it allows certain FTL trips to prevent other FTL trips). There could be cases where the Enterprise would be prevented from completing its warp trip on time because of an FTL signal sent by someone else. We certainly don't see that on the show (not surprisingly). So, considering this provision, I can't easily point out any arguments to support using it to keep warp travel from being self inconsistent.

This leaves us with the fourth provision, and I think you will see that it the provision of choice for the purposes of Trek. Of course, this fourth provision must involve some special frame of reference; therefore, we might first ask about where this special frame might come from. Thus, I will make a proposal for answering such a question in the next section, and then I will present what I believe are strong arguments for using the fourth provision to keep normal warp travel from being self inconsistent in Trek.

10.2 Subspace as a Special Frame of Reference

When we discussed the fourth, "special frame of reference" provision, I introduced the idea of a field which had a particular frame of reference. For Star Trek, we can imagine subspace to be this field, and we can let it pervade all of known space. Then, subspace (or at least some property of subspace) would define a particular frame of reference at every point in space. When you entered warp, you would take on the frame of reference of subspace and keep it, regardless of your velocity with respect to subspace. This would ensure that normal, everyday warp travel would not produce unsolvable paradoxes (as we discussed in Section 9.5.4).

So, what does this provision give us that the third provision didn't? Well, by assuming that subspace defines a special frame of reference, we can explain some interesting points on the technical side of Trek. For example, in the "Star Trek the Next Generation Technical Manual" (and in other sources) we see that the different warp numbers correspond (in some way) to different FTL speeds. But when they say that Warp 3 is 39 times the speed of light, we must ask what frame of reference this speed is measured in. With subspace as a special frame of reference, it would be understood to mean "39 times the speed of light in the frame of reference of subspace."

The same idea can be applied to references made to impulse-drive-only speeds. In the Technical Manual, they mention efficiency ratings for "velocities limited to 0.5c." They also mention the need for added power for "velocities above 0.75c." But these velocities are all relative, and so we must ask why these normal, slower than light velocity of the Enterprise should matter when considering efficiencies, etc. After all, the Enterprise is always traveling above 0.5 c in **some** frame of reference and above 0.75c in some other frame of reference. However, since impulse is supposed to use a subspace field to "lower the mass of the ship" (so that it is easier to propel), we could argue that the speed of the ship with respect to subspace (assuming subspace defines a special frame of reference) would effect efficiencies, etc.

Further, there is a much more documented example which refers to warp 10. As many of you know, warp 10 is supposed to be infinite speed in the Next Generation shows. That means that the event "you leave your departure point" would be simultaneous with the event "you arrive at your destination". But, as we have discussed, the question of whether two events are simultaneous or not truly depends on the frame of reference you are in. So, we ask, in what frame of reference is warp 10 actually infinite speed. Again, we can use the frame of reference of subspace to resolve this issue. Warp 10 would be understood to be infinite speed in the frame of reference of subspace.

Finally, using this provision, there would be a standard, understood definition for measuring times, lengths, etc. Times would be measured just as it would tick on a clock in the frame of reference of subspace, and distances would be measured just as they would be by a ruler at rest in the subspace frame of reference. Basically, the feeling we have for the way things work in every day, non-relativistic life would be applicable to Trek by using the subspace frame of reference as a standard, understood reference frame.

And so, I believe that the fourth provision gives us the best explanation for how normal, everyday warp travel in Trek could be self consistent.

10.3 The "Picture" this Gives Us of Warp Travel

Given the previous discussion, we see that the fourth provision seems to fit Star Trek like a glove. Thus, it may be best for us to view warp travel in Star Trek like this: Subspace is a field which defines a particular frame of reference at all points in known space. When you enter warp, you are using subspace such that you keep its frame of reference regardless of your speed. Not only does this mean that normal warp travel cannot be used to produce unsolvable paradoxes, but since in warp your frame of reference would no longer depend on your speed as it does in relativity, relativistic effects in general do not apply to travelers using warp. Since relativistic effects don't apply, you also have a general explanation as to why you can exceed the speed of light in the first place.

(As a note, this is similar to Alcubierre's idea for "warp" travel (mentioned earlier), but in his idea the traveler did not take on a "special" frame. Instead, he took on the frame he had before entering warp, but that allows two trips from two different frames of reference to produce an unsolvable paradox. If we add subspace as a special frame of reference to Alcubierre's idea, we could get a self consistent situation which would be very similar to what we see in Trek.)

For more information on how this might conceptually work in the science fiction world of Trek (at least one way I imagine it) you may want to read my other regular post, "Subspace Physics"[1]. Here, however, we can at least use this "picture" of warp to consider how the outside universe might appear to someone traveling at warp speed. Remember, at any point the warp traveler's frame of reference it is as if he is sitting still in subspace's reference frame. We could illustrate the way such an observer would picture a particular event by using the following idea: Picture a string of cameras, each a distance (d) away from the one before it. Let these cameras all be stationary in the frame of reference of subspace, and let them all be pointed at the event of interest. Further, let each camera have a clock on it, and let all the clocks be synchronized in the subspace frame. Then, we can set each camera to go off with the time between one camera flash and the next being d/v (where v is the FTL velocity of the observer we want to illustrate). Then, each picture is taken in the subspace frame of reference, but the string of pictures (one from each camera) would form a movie in which each frame was taken from a different place in space from the previous frame. Thus, we can use this to produce a film of how an event would look to a warp traveler.

Of course, in Trek they have subspace sensors which do all their seeing for them (faster than light, of course). However, the above does illustrate one's ability to use this view of warp travel to answer various technical questions.

10.4 Some Notes on Non-Warp FTL Travel and Time Travel in Trek

Now, there are cases in Trek where FTL travel exists without necessarily using subspace (and thus the subspace frame of reference would not apply and would not prevent unsolvable paradoxes). For example, if the wormhole in Deep Space Nine is assumed to be the same as a wormhole we theorize about today, then it wouldn't need to deal with subspace to allow FTL travel. (Now, what they call a wormhole doesn't necessarily have to be what we call a wormhole, but for this illustration, let's assume it is). So, if the wormholes in Trek aren't bounded by the subspace frame of reference, we could imagine a situation whereby they could be used to cause unsolvable paradoxes. This is true for any form of FTL travel in Trek which might not use subspace. However, I propose that in cases where subspace isn't used (so that its special frame of reference could not prevent unsolvable paradoxes) then the first or second provision, "parallel universes" or "consistency protection", would apply. In that way, we can allow for non-warp/non-subspace-using FTL travel in Trek while still preventing unsolvable paradoxes.

Further, consider time travel in Trek. Actual time travel couldn't be accomplished by using subspace alone (the subspace frame along with the fourth provision would prevent it). However, I propose again that such travels in time should not be able to produce unsolvable paradoxes because the "parallel universes" or "consistency protection" provisions would apply (since subspace alone couldn't be in use to produce the time travel).

For example, consider the Star Trek: The Next Generation episode, "Time's Arrow" (in which Data's severed head is found on 24th century Earth, and Data eventually travels back in time to (unintentionally) leave his head behind to be found). Now, after the head was found, one of the crew (let's say Riker, just to use an example) could decide to try to produce an unsolvable paradox. Riker may decide to do everything in his power so as to keep Data from going back in time. He may even try to destroy Data and his head to accomplish this task. Of course, Riker isn't the type of person to do this, but what if he was? Well, in that case, he would be trying to produce an unsolvable paradox, and the first or second provision would prevent it. For the first provision, the head found in the 24th century might have actually come from a parallel

[1] http://www.physicsguy.com/star-trek/subspace-physics/

universe. For the second provision, we could imagine various ways in which Riker might fail in his task of trying to keep data from going back in time. Further, we could consider the case in which he would succeed in producing an unsolvable paradox and we could insist that such situations would destroy themselves or prevent themselves from ever happening.

Such a situation is seen in a particular Voyager episode. In this episode, members of the crew are caught in a "subspace fissure", and they travel back in time. By the end of the episode, their trip back in time has produced a self-inconsistent situation. That series of events then becomes impossible and ceases to exist by the closing credits. This could be seen as a result of having the "consistency protection provision" apply to a case where the subspace frame of reference is bypassed via "subspace fissures".

So, even though we can be relatively sure that this was not the intention of the writers, the situations shown do seem to comply with the concepts we have developed.

10.5 To sum up...

To sum up, we have found that by introducing a special frame of reference which would be "attached" to subspace, and by further insisting that any type of FTL/time travel done without using subspace be governed by the "parallel universe" or "consistency protection" provisions, we will not only have a self consistent universe for our Star Trek stories, but we can also (coincidentally) explain many of the "but how come...?" questions which some Star Trek episodes produce.

Chapter 11

Conclusion

In Part I of this FAQ, I presented some of major concepts of special relativity, and here in Part IV, we have discussed the considerable havoc they play with the possibility of faster than light travel. I have argued that the possibility of producing unsolvable paradox is a very powerful deterrent to all FTL concepts. Further, we have introduced four basic provisions, at least one of which must be in place so that FTL trips/signals (sent using any of the FTL concepts) cannot be used to produce unsolvable paradoxes. Finally, we looked at the science fiction of Star Trek while considering all that we had discussed. We concluded that warp travel could be governed by the fourth provision (via subspace defining a special frame of reference) while all other FTL travel (or time travel) could be governed by the first or second provisions. This, I believe, best explains what we see on Star Trek.

If you have not read Part II or Part III of this FAQ, and you are interested in learning more about relativity (special and general), then you may want to give them a look.

As the end result of producing this FAQ, I hope that I have at least informed you to some extent (or perhaps just helped to clarified your own knowledge) concerning relativity and the problems it poses for FTL travel.

Jason Hinson