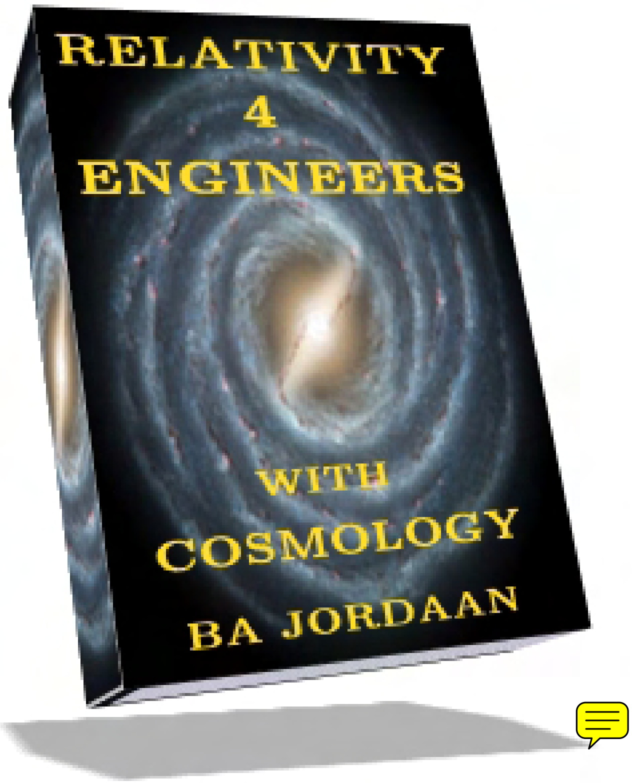
# RELATIVITY 4 ENGINEERS



# EINSTEIN'S RELATIVITY

# FOR

# ENGINEERS

## The engineer's resource

## for

## Relativity and Cosmology

**Burt A. Jordaan**

# Contents

# List of Figures

# Preface

why this book was written
and
what it is about

As an engineer working in the aerospace industry, the author started to collect and read popular books about the subjects of relativity and cosmology, partly because they are intriguing, and partly because they are loosely connected to his profession.

Easy to read and enjoyable as such popular books are, they do however tend to leave a lot of unanswered questions in the enquiring mind. One of the problems with popular books is that the publishers want a mass market and therefore they virtually ban mathematics from such books.

The author next turned to technical books on the subjects. Here he was confronted with two problems. Firstly, prior technical knowledge, mostly only known inside the trade, seems to be required. Secondly, (especially in general relativity) the mathematical treatment is pretty hard to comprehend.

After a lot of grinding in his spare time, the author succeeded in extracting something more ‘engineering-like’ from the formidable mathematics. This gave rise to this book. It is an attempt to bridge the gap between the popular book and the highly technical book. It is full of mathematics, but at a level that engineers can easily “connect to”.

It is true that mathematics does not make for easy armchair reading, especially if the derivations of formulas are part of the text. However, it is almost impossible to express the subtleties of cosmology and relativity purely in words. Even with the aid of pictures, pure text cannot generally give the understanding that a few formulas can.

This book starts with a general introduction to relativity, giving a flavor of both the special and the general theory. Here just a very light sprinkling of tensors, as used in differential geometry, is utilized. From then on, the math is a lot more accessible.

The book next proceeds through the special case of zero gravity (special relativity), including all the usual issues. The treatment concentrates on spacetime intervals, avoiding the usual 'controversies', like "I think your clock runs slower than mine, but then you may think that my clock runs slower than your's." Thorough discussions of the Lorentz transformation and also accelerated frames of reference are included.

The next major part concentrates on gravitation theory (general relativity). It deals with all the more common issues of gravitation, including gravitational redshift, acceleration with and without movement, black holes, orbits, gravitational lensing, tidal gravity and gravitational waves.

All this is done for the simplest relativistic case only, namely Schwarzschild coordinates. This holds for a gravitational source that is permanently at rest at the centre of the coordinate system and is non-rotating. The treatment given in this text avoids the use of tensors, but is mathematically rigorous, for as far as it goes.*

*Tensors are needed in many of the derivations or proofs, but they are skipped where possible.

The treatment on gravitational acceleration includes some rigorously derived 'quasi-Newtonian' formulae, which should appeal to engineers to a larger degree than what most relativity books offer. The 'quasi-Newtonian' acceleration is used to plot orbits in more or less the way engineers simulate dynamical systems.

The chapters on tidal gravity and gravitational waves are kept as practical (and engineering-like) as is feasible. Finally, the 'modern' way of doing relativity, i.e., the post-Newtonian formalism, is discussed briefly. This is probably the 'interface' that engineers working in the space industry are most likely to encounter.

The last major part of this text deals with the essential elements of cosmology. It covers the standard big bang theory, starting from a simple model that leads up to the proper mathematics and the conclusions that can be drawn from it. Included are the expansion laws for various models, redshift interpretations and look-back time.

Then the essentials of inflation theory is covered rather loosely (engineering-like). Finally the contemporary views on vacuum-energy driven universal expansion are discussed. Relativity does play a role in theoretical cosmology, but it is slight in the treatment offered here.

There is quite a bit of mathematics involved, but it generally requires no more than a high school background. Further, the meaning of all formulae used is discussed, so it is not necessary to read them intensively in the first place. They are however right there in the text as an aid in understanding the subject matter.

The late Richard Courant once said: "In applied mathematics, the most important problem is the transition of the problem of reality to the formulation of a mathematical model, followed by the mathematical analysis of the model, and the next step, translation of the results again into the language of reality". This book is mostly about the last step, with just a sprinkling of the first two.*

*Courant was an applied mathematician, perhaps most famous for his ‘finite element method’, although he did not name it as such.

Where understanding of the subject is greatly enhanced by the derivation of a formula, such derivations are contained in boxes in the same chapter, or they appear in an appendice at the end of the book. The text can be followed without reading them, but enquiring minds might be compelled towards them.

The author deviates from the current practice of inserting boxes in the body of a chapter, simply because he finds it very distracting when reading a technical book. The boxes are grouped together at the end of each chapter.

The illustrations used in this book are generally computer generated, using the formulae in the book, which are rigorously correct, barring blunders, misprints or where indicated as approximations.

This text does not propose any new theory for cosmology or relativity (as has been coming regularly from sources outside of those fields, including from engineers). It rather attempts to make the existing theories more accessible to engineers and the like.

The author has found that some of the ‘proposed new theories’ contain pretty convincing arguments, only to find it being more convincingly refuted by experts in the trade—provided one can understand what the experts say!

This text projects a different view—an engineering view if you like—of the subjects, drawing from the works of specialists, to whom the author is greatly indebted. References to their works are made throughout the text. It is not expected that the specialists will read this book, but if any of them do, the engineering view may appear a trifle shallow for their liking.

This may be so, but the idea is simply to ‘bootstrap’ engineers so that they may perhaps understand better what relativists and cosmologists speak and write about. In the end, engineers are needed to design some of the sophisticated instruments and machinery needed in the search for knowledge about space, the universe and physics in general.

Special thanks are due to some of my engineering colleagues, who suffered numerous draft articles during the research and formulation of this project. Whatever the degree of clarity of this work might be, it would have been far less clear without their feedback.

Lastly, I dedicate this work to my family, who suffered severe timesharing with the book. Like Dr. John Peacock said in the preface to his excellent book **Cosmological Physics:** “... the conflicting demands for any spare time inevitably make it difficult for families and [the writing of] books to coexist peacefully.” [Peacock]

Burt A. Jordaan
Centurion, South Africa

# Chapter 1

# Introducing Engineers to Relativity

space, time and
gravitation,
briefly

One of the best introductions to his theory of relativity is found in Einstein's essays [Einstein], in which he stated:

"The theory of relativity is that physical theory which is based on a consistent physical interpretation of the concepts of motion, space and time. The name 'theory of relativity' is connected with the fact that motion from the point of view of possible experience always appears as the *relative* motion of one object with respect to another.

"Motion is never observable as 'motion with respect to space' or, as it has been expressed, as 'absolute motion.' The 'principle of relativity' in it's widest sense is contained in the statement: The totality of physical phenomena is of such a character that it gives no basis for the introduction of the concept of 'absolute motion'; or shorter but less precise: There is no absolute motion." Einstein continued with:

"The development of the theory of relativity proceeded in two steps, 'special theory of relativity' and 'general theory of relativity.' The latter presumes the validity of the former as a limiting case and is its consistent continuation."

Einstein then briefly described both the special and the general theories of relativity. It is worth to emphasize the fact which Einstein stated above—it is not two theories; general relativity completely includes special relativity as a limiting case where the gravitational field is negligible. Such a limiting case can happen in empty space, far away from massive objects and in the absence of acceleration. It can also happen closer to massive objects, inside a small laboratory that is in free-fall, where the gravitational field is practically not detectable inside the laboratory.

## 1.1 The special theory of relativity, briefly

Einstein was lead to his special theory of relativity by his believe that there is no way to detect absolute motion. This dictated that the measured speed of light must be the same in all inertial frames of reference.* Einstein called this the *"Light-principle"*, or L-principle for short.

Einstein realized that of if this principle does not hold, there must be a 'rest-frame' for light. This means that we could in principle set up an inertial frame in which light would not propagate in the forward direction at all (if the frame moves at the speed of light relative to the aether). Einstein reportedly contemplated if he would still be able to see his own face in a mirror if they were both at rest in such a frame.

We can extend this to say that radars as we know them would not work in such a moving inertial frame. Even at less extreme speeds, standard radars would report wrong distances, with errors that depend on direction of movement. More about radar measurements later.

Einstein realized that it is paradoxical to assume the same light ray can actually move with the same speed $c$ (in an absolute Newtonian sense) relative to all inertial frames. This would require that light adapts it's "absolute speed" to the frame that measures it. He decided that either time intervals or distance measurements (or both) must change if measured by observers in different inertial frames that are in relative motion.

Exactly how these intervals must change, Einstein found in the transformation equations that Lorentz has developed before. These equations transforms time and distance measurements from one inertial frame to another precisely as required by the L-principle.

We will deal with the Lorentz transformation in a later chapter. For the purpose of introduction, it requires that for any two events, $A$ and $B$ that occurs in space and time, there is a quantity called the *spacetime interval* that remains unchanged, irrespective of in which inertial frame the components of the spacetime interval are measured.

The spacetime interval can be 'spacelike', 'lightlike' or 'timelike', as defined below:

$$\Delta s^2 = \begin{cases} (\Delta space)^2 - (\Delta time)^2 & \text{if } \Delta space > \Delta time \quad \text{(spacelike)}, \\ 0 & \text{if } \Delta space = \Delta time \quad \text{(lightlike)}, \\ (\Delta time)^2 - (\Delta space)^2 & \text{if } \Delta space < \Delta time \quad \text{(timelike)}, \end{cases}$$

in geometric units, where $c = 1$ so that $\Delta time$ and $\Delta space$ are expressed in the same units. Most engineers would probably prefer this to rather be expressed in the normal SI units of metres and seconds. It can be done

*Inertial frames are uniformly moving coordinate systems, far away from gravitational or any other form of influence, where inertia is isotropic, meaning a given force will cause the same acceleration on identical masses in whatever direction the force is applied.

by simply replacing all references to time by $ct$, thus converting seconds to metres. The spacetime interval will look then like this:

$$\Delta s^2 = \begin{cases} \Delta\mathbf{x}^2 - c^2\Delta t^2 & \text{if } \Delta\mathbf{x} > c\Delta t \quad \text{(spacelike)}, \\ 0 & \text{if } \Delta\mathbf{x} = c\Delta t \quad \text{(lightlike)}, \\ c^2\Delta t^2 - \Delta\mathbf{x}^2 & \text{if } \Delta\mathbf{x} < c\Delta t \quad \text{(timelike)}. \end{cases}$$

The author attempts to use SI units throughout, but here and there it is so much clearer if the constant $c$ is not cluttering the equations that geometric units are being used. It is usually very clear which units are under consideration.

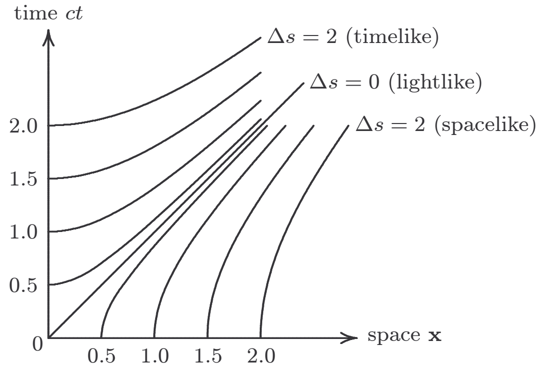


**Figure 1.1:** Spacetime intervals plotted for $\Delta s$ ranging from 0 to 2 in 0.5 steps, both lightlike and timelike. For $\mathbf{x}$ large, they all approach the $\Delta s = 0$ (lightlike) line asymptotically. Note the precise symmetry of spacelike and timelike intervals around the lightlike interval.

Figure 1.1 illustrates the three types of interval on a standard spacetime diagram. A timelike interval is the only type where an observer, traveling slower than light, can be present at both events. This is so because there is enough time to cover the distance between the two events at a speed slower than that of light.

It then follows that a spacelike interval is the type where nothing, not even light, can be present at both events. In relativistic jargon, the two events are not causally connected, or stated more simply, the two events could not have influenced each other.

A lightlike interval is the borderline between the above two intervals and is only applicable to some types of waves and to massless particles—light, radio waves, gravitational waves, etc—things that move at the speed of light.

Spacelike intervals are normally denoted by $\Delta s$ and timelike intervals with $\Delta\tau$, so that $\Delta s = -c\Delta\tau$.

The fact that the (spacetime) interval remains unchanged, irrespective of which inertial system measures it, may appear to be completely unremarkable. In Newton mechanics, where time intervals and space intervals remain unchanging when you change your inertial frame of reference, the interval will obviously remain unchanged.

Newton however, demands that the *measured speed of light* is different in different inertial reference frames. In order to conform to Einstein's *L-principle* (the invariance of the measured speed of light), either $\Delta$t*ime* or $\Delta$s*pace* or *both* must change if you switch between inertial reference frames that is moving relative to each other.

In order to satisfy the L-principle and leave the interval unchanging, both must change in a very specific way.

**The speed connection** If we take the timelike interval between two events $A$ and $B$ as

$$c\Delta\tau = \sqrt{c^2\Delta t^2 - \Delta s^2}$$

and factorize $c^2\Delta t^2$ out from the righthand side, we get

$$c\Delta\tau = \sqrt{1-\dot{\mathbf{x}}^2}\Delta t, \qquad (1.1)$$

where $\dot{\mathbf{x}} = \frac{\Delta s}{c\Delta t}$, which can be interpreted as the uniform speed that an observer must maintain (relative to the reference frame) to be present at both events.

The arrow $AB$ in figure 1.2 represents an observer that leaves event $A$ at time $t_A$ and arrives at event $B$ at time $t_B$, as measured in the coordinate system $\mathbf{x}, ct$. In the coordinate system of the moving observer $(\mathbf{x}', ct')$, the arrival time is at time $t'_B$, so that $\Delta t' = t'_B - t'_A = \Delta\tau$.

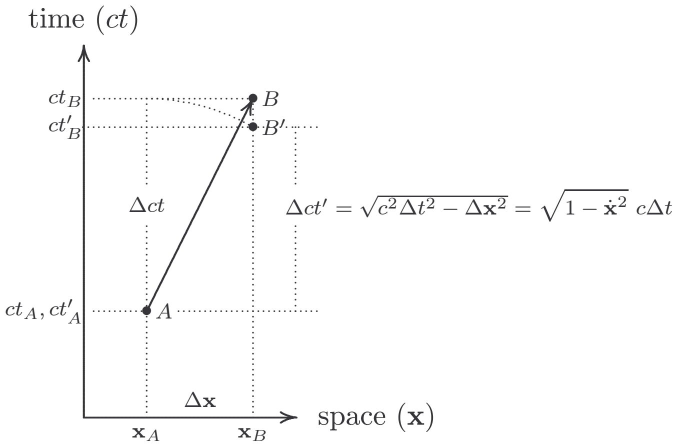


**Figure 1.2:** The arrow $AB$ represents a uniformly moving observer that is present at both events $A$ and $B$. In order to keep the interval $\Delta s$ invariant, the moving observer must measure a time interval of $\Delta t' = \sqrt{1-\dot{\mathbf{x}}^2}\Delta t$.

By definition, the 'moving' observer is stationary in an inertial frame that moves at a speed $\dot{\mathbf{x}}$ relative to the original reference frame. Since the observer is present at both events, the two events are separated in the observer's space by $\Delta\mathbf{x}' = 0$. The interval $\Delta\tau$ must be identical for both inertial frames, so

$$c\Delta\tau = \sqrt{c^2\Delta t^2 - \Delta\mathbf{x}^2} = \sqrt{c^2\Delta t'^2 - 0} = c\Delta t',$$

meaning the time difference between the two events must be measured by the moving observer as

$$\Delta t' = \Delta\tau = \sqrt{1 - \dot{\mathbf{x}}^2}\Delta t. \tag{1.2}$$

This statement counters the argument that is sometimes expressed, namely that special relativity is ambiguous in that either of the two observers can be considered as moving relative to the other one, so either could be considered as having a clock that is 'slow' when compared to the other's.

Here the situation is not symmetrical—one observer is present at both events and the other one is not. It is true that any one of the observers can be chosen as 'stationary' and the other one as moving relative to this 'stationary' frame of reference.

However, if the two (inertial) observers are moving relative to each other, *only one of them can be present at both events.** The time interval measured by that observer is called the *propertime interval*, $\Delta\tau$. Propertime is an extremely important concept in relativity theory.

The above interpretation has nothing to do with the fact that distant observers will detect events with a time delay caused by the finite speed of light. By 'measure the time difference' we mean that the event times have been corrected for the time that light takes to travel from the event to the observer.

This does of course mean that the distance between the observer and the events must be known, which brings us to the way inertial observers will measure the distance between the two events.

Let Pam be the observer that moves between events $A$ and $B$ at a speed $\dot{\mathbf{x}}$ relative to Jim. Let event $A$ happens as the two of them pass each other. There is no problem to understand how Pam measures the distance between the events. She is after all present at both and the distance in her inertial frame is zero.

How does Jim, who is *not* present at both events, measure the distance between events $A$ and $B$? Let event $B$ be a flash of light, generated by Pam after she has moved some distance away from Jim.

Equip Jim with a good radar with which he can constantly monitor Pam's distance as she moves away from him. Jim can read the distance ($\Delta\mathbf{x}$) of event $B$ at the moment he observes the light flash, directly from his radar.

The fact that by the time Jim observes the flash from event $B$, Pam will be some distance past the position of the event, does not influence Jim's

*Provided that the two events do not both happen at the place and moment where the two observers pass each other—then they will both measure $\Delta t = \Delta\mathbf{x} = 0$, which is not an interesting experiment.

confidence in his measurement. The return signal of his radar and the flash of event $B$ started out at precisely the same place and time and came to him at the same speed—the speed of light.

It is fairly obvious that Jim will measure a longer distance between the two events than Pam. Pam is the 'moving observer', who here measured the distance as zero!

However, Pam was stationary in her own inertial frame of reference, so for her Jim was the 'moving observer'. The only—and very significant—difference between Pam and Jim is that Pam was present at both events and Jim was not.

Inertial observers that are moving relative to Pam cannot also be present at both events.* They will all measure time and distance intervals between the two events that are longer then those measured by Pam.

The reason for laboring the observation of the time and space intervals between events is this: events in empty space give us something 'tangible' to base comparisons between inertial frames on.

It does not say whose clock is running faster or slower than anybody else's. It does say unequivocally who will *measure* the shorter time and distance between two events—it the observer who is present at both events.

One of the classic 'tests' of special relativity's predictions is the case of the muon particles. They are created high in the earth's atmosphere by cosmic rays hitting oxygen atoms. The muons have such a short 'half-life'* that even if they travel at the speed of light,* virtually none of them could possibly make it to the surface of the Earth.

Yet they are routinely detected in laboratories on Earth in reasonable abundance. The secret lies in the fact that the muons are present at two events, $C$ (their creation) and $D$ (their detection on the ground), while we as observers on the ground are clearly not present at both events.

On Newtonian grounds, we predict that the time that the muons would take to reach the ground is much too long for any appreciable number to survive. Because of their high speed, the muons experience a time difference between events $C$ and $D$ that is much shorter than what Newton would have predicted, allowing a lot of them to make it to the ground.

So far, this introduction was an 'engineering-like' attempt to acquaint the reader with the all important spacetime interval. So before we go any

*Remember, they are inertial observers, so they cannot turn around in any way.

*Half-life is a statistical parameter, meaning the time in which half of the particles (on average) would have decayed.

*Which they don't, but they come quite close to the speed of light.

further, it will be appropriate to give a brief overview of how the relativists view and express the spacetime interval.

## 1.2 The formal spacetime metric, briefly

Spacetime, in the absence of gravity, is expressed by the Minkowski metric with a *line element*

$$\begin{aligned} ds^2 &\equiv \eta_{\mu\nu}\, d\mathbf{x}^\mu d\mathbf{x}^\nu \qquad (\mu, \nu = 0, 1, 2, 3) \\ &= -c^2 dt^2 + dx^2 + dy^2 + dz^2, \end{aligned} \tag{1.3}$$

where $\eta_{\mu\nu}$ is the Minkowski metric tensor. The indices $\mu$ and $\nu$ indicate which component of 4-space is under consideration (i.e., $t, x, y$ or $z$). In this notation, $\mathbf{x}^\mu$ does not mean $\mathbf{x}$ raised to the power $\mu$, but rather that $\mu$ is an index that indicates how $\mathbf{x}$ must be summed.

This 'Einstein summation convention' sums over repeated indices, e.g., the $\mu$ and $\nu$ in $\eta_{\mu\nu}$ and $d\mathbf{x}^\mu d\mathbf{x}^\nu$, without the implicit summation sign being used. For example, if both $\mu$ and $\nu$ range from 0 to 1 only, the summation will result in

$$\eta_{\mu\nu} d\mathbf{x}^\mu d\mathbf{x}^\nu = \eta_{\mu\nu}(d\mathbf{x}^0 d\mathbf{x}^0 + d\mathbf{x}^0 d\mathbf{x}^1 + d\mathbf{x}^1 d\mathbf{x}^0 + d\mathbf{x}^1 d\mathbf{x}^1). \tag{1.4}$$

When $\mu$ and $\nu$ range from 0 to 3, the summation will naturally have all combinations up to $d\mathbf{x}^3 d\mathbf{x}^3$, i.e., 16 terms in all. Each $d\mathbf{x}^\mu d\mathbf{x}^\nu$ term is multiplied by the corresponding element of the metric tensor $\eta_{\mu\nu}$, which is best represented by a 4x4 matrix.

For Minkowski spacetime the $\eta_{\mu\nu}$ matrix is relatively simple, presented here as a "bordered matrix" for clarity:

$$(\eta_{\mu\nu}) = \begin{array}{c c} & \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \end{array} & \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array},$$

where only the diagonal elements are non-zero and they are unitary, indicating 'flat' spacetime. It means that only terms with $\mu = \nu$ remains after multiplication and the coefficients are all unity, so that

$$\begin{aligned} \eta_{\mu\nu} d\mathbf{x}^\mu d\mathbf{x}^\nu &= -(d\mathbf{x}^0)^2 + (d\mathbf{x}^1)^2 + (d\mathbf{x}^2)^2 + (d\mathbf{x}^3)^2 \\ &= -c^2 dt^2 + dx^2 + dy^2 + dz^2. \end{aligned} \tag{1.5}$$

This may look like a very round-about way to achieve a simple result. And what is more, to rigorously prove that the metric tensor $\eta_{\mu\nu}$ has the form shown above, requires quite complex tensor analysis. We will skip that and accept the $\eta_{\mu\nu}$ matrix at face value. The complexity is the price paid for mathematical generality.

The $d\mathbf{x}^\mu d\mathbf{x}^\nu$ terms can represent many things, not just 4-space coordinates. Further, $ds^2 = \eta_{\mu\nu}\ d\mathbf{x}^\mu d\mathbf{x}^\nu$ is not restricted to 'flat' spacetime, as we will see later. Also, the formalism can handle virtually any coordinate system. We can replace the Cartesian coordinate system $(x, y, z)$ with an equivalent spherical coordinate system $(r, \theta, \phi)$, as shown in figure 1.3, where

$$\begin{aligned} x &= r\cos\theta\sin\phi \\ y &= r\sin\theta\sin\phi \\ z &= r\cos\phi. \end{aligned}$$

When $dx, dy$ and $dz$ are computed from the above and substituted into the Cartesian line element, the line element for spherical coordinates becomes

$$ds^2 = -c^2dt^2 + dr^2 + r^2d\phi^2 + r^2\sin^2\phi\ d\theta^2, \tag{1.6}$$

valid for 'flat' spacetime.

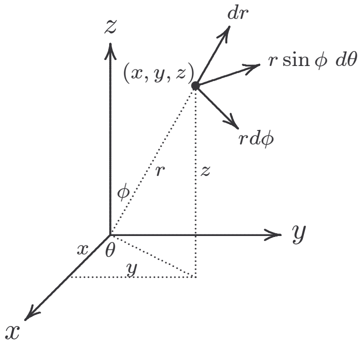


**Figure 1.3:** Small changes in $r$, $\theta$ and $\phi$ create an 'orthogonal coordinate system' $dr$, $rd\phi$ and $r\sin\phi\ d\theta$ at point $x, y, z$. Note that while $dr$ signifies radial displacement, the other two directions signify transverse (or tangential) displacements relative to the origin.

This gives the elements of the $\eta_{\mu\nu}$ matrix for spherical coordinates as

$$(\eta_{\mu\nu}) = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & r^2\sin^2\phi \end{pmatrix}.$$

The above spherical form of the metric is not important in flat spacetime, but is very convenient in the curved spacetime environment of general relativity, as will be discussed later in this chapter.

The spacetime line element $ds$ corresponds to a spacelike interval $\Delta s$. We have seen before that a timelike interval, normally indicated by $\Delta\tau$, can be obtained from $c^2\Delta\tau^2 = -\Delta s^2$. So the timelike line element can be written as

$$c^2d\tau^2 = -ds^2 = c^2dt^2 - dx^2 - dy^2 - dz^2.$$

Some authors prefer to work solely with the timelike interval as the metric of spacetime, e.g., [Faber]. This is presumably because most of the intervals that we can observe and measure are timelike.

**The Lorentz transformation.** A Dutch physicist H.A. Lorentz is credited with a set of transformation equations that transformed space and time between inertial frames in relative motion. Historically, Lorentz was not the first person making the suggestion,* but he was the first to publish the set of equations, now known as the *Lorentz transformation*

Lorentz did not discover special relativity though. He simply found a mathematical way to transform measurements made on objects moving through the aether (or absolute space) that made them conform to the null result of the aether-drift experiment of Michelson and Morley.

In essence, his equations transformed the space interval ($\Delta\mathbf{x}'$) and time interval ($\Delta t'$) as measured by a frame moving relative to the aether, to the 'absolute' space interval $\Delta\mathbf{x}$ and 'absolute' time interval $\Delta t$.

Lorentz had no physical theory for why his transformations seem to agree with experiments. He did postulate that length contraction might be a physical reality, but he could not explain why time had to transformed too, other than that it explained observations. The Lorentz transformation equations in SI units are:

$$\Delta\mathbf{x} = \frac{\Delta\mathbf{x}' + \dot{\mathbf{x}}c\Delta t'}{\sqrt{1-\dot{\mathbf{x}}^2}} \tag{1.7}$$

$$c\Delta t = \frac{c\Delta t' + \dot{\mathbf{x}}\Delta\mathbf{x}'}{\sqrt{1-\dot{\mathbf{x}}^2}} \tag{1.8}$$

where $\dot{\mathbf{x}} = \frac{d\mathbf{x}}{cdt}$, the speed relative to the aether. Einstein's contribution was that these equations can be used to transform time and space directly between any two inertial frames in relative motion and not just between an inertial frame and the absolute (aether) frame. In short, they conform to the principles of relativity.

The meaning of Einstein's interpretation of the Lorentz transformation is simply this: measure a space interval and a time interval in any inertial frame. Through the invariance of the spacetime interval, the transformation tells you what the value of the space and time intervals will be in any other inertial frame.

The 'absolute frame of reference', the aether, was not required at all. As we will see later, things are sometimes much simpler when we do not have to contend with the aether—especially in one-way Doppler measurements.

## 1.3 The general theory of relativity, briefly

Einstein's *general theory of relativity* is essentially a theory about the gravitational fields generated by massive objects. It is also about the dynamics of objects moving in such gravitational fields.

*G.Fitzgerald, an Irish physicist, first postulated a contraction in the direction of movement. Relativistic length contraction is commonly known as *Lorentz/Fitzgerald contraction*.

The objects can be massless particles like photons that always move at the speed of light relative to every inertial reference frame; or they can be massive objects that always move at speeds less than that of light relative to every inertial reference frame.

The gravitational field can be thought of as a 'deformation' of the fabric of spacetime caused by massive objects. 'Test objects' move 'as straight as possible' through this deformed spacetime.

All cases are locked up in Einstein's field equations, *the Einstein equation* for short. With $c = 1$, $G = 1$,* it is given by

*This means that geometric units are being used for simplicity and clarity.

$$R_{\mu\nu} - \frac{1}{2} R \, g_{\mu\nu} = T_{\mu\nu} \qquad \text{(where } \mu, \nu = 0, 1, 2, 3\text{)}, \tag{1.9}$$

where $R_{\mu\nu}$ is the *Ricci tensor*, $R$ the *Ricci scalar*, $g_{\mu\nu}$ the generalized form of the metric tensor $\eta_{\mu\nu}$ and $T_{\mu\nu}$ the *energy-momentum tensor*. The Ricci tensor is a contraction of the *Riemann curvature tensor* and the Ricci scalar is the trace of the Ricci tensor.

So the equation tells us how the energy and momentum in space (the right hand side) cause the curvature of spacetime (the left hand side). The curvature of spacetime influences the movement of massive bodies through spacetime, thus changing the momenta. So there is 'cross-talk' between the left- and right hand sides, making the full equation very, very difficult to solve.

For any given situation there are up to 10 different $T_{\mu\nu}$ values to be established in terms of energy, distance and time—not 16, because the tensor is symmetrical, meaning $T_{01} = T_{10}$ etc. Various solutions to these equations presumably represent every possible equation of motion that exists in the macroscopic universe.

We will not dig into all that complexity, but rather attempt to provide some intuitive feel for certain specific solutions. The first, and perhaps best known exact solution to Einstein's field equation was derived by Karl Schwarzschild in 1916, only months after Einstein published his general theory of relativity.

This solution provides the gravitational field outside of an isolated, spherically symmetrical, non-rotating mass, permanently at rest at the origin of a 3-d spherical coordinate system $r, \theta, \phi$ (as in figure 1.3).

In such a case $R_{\mu\nu} = T_{\mu\nu} = 0$, but all the components making them up are not zero (the components simply sum to zero). The solution is then simpler, though not trivial at all. The spacetime metric of the gravitational field is obtained through the *metric tensor*, just like for flat spacetime, as

$$ds^2 \equiv g_{\mu\nu} \, d\mathbf{x}^\mu d\mathbf{x}^\nu,$$

or

$$c^2 d\tau^2 \equiv -g_{\mu\nu} \, d\mathbf{x}^\mu d\mathbf{x}^\nu,$$

with

$$(g_{\mu\nu}) = \begin{array}{c} \\ cdt \\ dr \\ d\phi \\ d\theta \end{array} \begin{array}{c} \begin{array}{cccc} cdt & dr & d\phi & d\theta \end{array} \\ \begin{pmatrix} g_{00} & 0 & 0 & 0 \\ 0 & g_{11} & 0 & 0 \\ 0 & 0 & g_{22} & 0 \\ 0 & 0 & 0 & g_{33} \end{pmatrix} \end{array},$$

so that the metric becomes

$$c^2 d\tau^2 = -(g_{00}\ c^2 dt^2 + g_{11}\ dr^2 + g_{22}\ d\phi^2 + g_{33}\ d\theta^2). \tag{1.10}$$

By (tediously)* solving for the $T_{\mu\nu}$ of the energy-momentum tensor and casting them into $g_{\mu\nu}$ form, the following values are obtained for the non-zero elements of the metric tensor:

*Found in most general relativity texts. The details fall outside of the scope of this book.

$$\begin{aligned} g_{00} &= -(1 - \tfrac{2GM}{rc^2}) && (1.11) \\ g_{11} &= (1 - \tfrac{2GM}{rc^2})^{-1} = -1/g_{00} && (1.12) \\ g_{22} &= r^2 && (1.13) \\ g_{33} &= r^2 \sin^2\phi. && (1.14) \\ & && (1.15) \end{aligned}$$

This gives the 'Schwarzschild metric' as

$$c^2 d\tau^2 = (1 - \tfrac{2GM}{rc^2})c^2 dt^2 - \frac{dr^2}{1 - \frac{2GM}{rc^2}} - r^2 d\phi^2 - r^2 \sin^2\phi\ d\theta^2. \tag{1.16}$$

It is easy to see that when $rc^2 \gg 2GM$, i.e., far from the central mass, the metric reduces to the 'flat' Minkowski spacetime of special relativity.

Because of the 'awkwardness'* of using, especially, $\sqrt{-g_{00}}$ in many places, $g_{00}$ will be replaced by $g_{tt} = -g_{00} = 1 - 2GM/(rc^2)$, meaning the 'time-time' coefficient of the Schwarzschild metric.

*Some texts, e.g. [Pathria] use the convention of labeling indices from 1 to 4 and making time the 4$^{th}$ coefficient $g_{44}$.

To make it clear that the usage is nonstandard, the other coefficient that is regularly used, $g_{11}$, will be relabeled $g_{rr}$, loosely meaning the 'radial-radial' coefficient. Now, if we express $d\tau$ in terms of $dt$, like we did for special relativity, we get

$$c^2 d\tau^2 = [g_{tt} - g_{rr}\frac{dr^2}{c^2 dt^2} - \frac{r^2(d\phi^2 + \sin^2\phi\ d\theta^2)}{c^2 dt^2}]\ c^2 dt^2.$$

Since $dr^2$ is a radial spatial displacement squared and $r^2(d\phi^2 + \sin^2\phi\ d\theta^2)$ is a transverse spatial displacement squared, we can write

$$d\tau^2 = [g_{tt} - g_{rr}\frac{v_r^2}{c^2} - \frac{v_t^2}{c^2}]\ dt^2. \tag{1.17}$$

where $v_r$ and $v_t$ are the radial and transverse coordinate velocity components respectively.

This is a most illuminating expression of the spacetime metric. It tells us that, compared to coordinate time flow $dt^2$, propertime flow $d\tau^2$ is reduced by three terms inside the bracket.

The first is a static term: $g_{tt}$, which is always less than unity. The other two are velocity related terms: $g_{rr}v_r^2/c^2$ and $v_t^2/c^2$. We will first examine the static term in more detail.

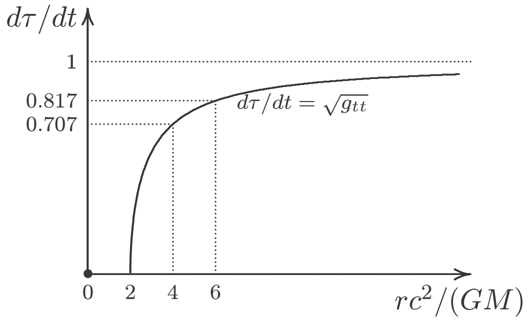


**Figure 1.4:** The gravitational time dilation factor (or propertime flow $d\tau/dt$), against coordinate radial distance $r$ from mass $M$. The region $2\frac{GM}{rc^2} < r <= 6\frac{GM}{rc^2}$ has special significance (see text for details). Far from mass $M$, when $r \to \infty$, $d\tau/dt \to 1$.

As shown in figure 1.4, the 'rate of propertime flow' $d\tau/dt$ is zero at $r = 2GM/c^2$ and then increases rapidly until $r = 4GM/c^2$, where the slope of the curve is unity (45 degrees). After that, the curve starts to approach unity asymptotically.

We will later see that the *measured* static gravitational acceleration (i.e., the initial acceleration of an object kept stationary and then released so that it free-falls), is proportional to the slope of the curve $d\tau/dt$ against $r$, at least to a first approximation.

This suggests that the measured gravitational acceleration at $r = 2GM/c^2$ will approach infinity. This radial distance is called the Schwarzschild radius $r_S$ and the spherical surface associated with $r_S$ is called the event horizon of a static black hole. Because of the infinite acceleration, nothing, not even light, can escape from within the event horizon.

The velocity related terms are a bit more subtle. We have met the velocity time dilation factor of flat spacetime: $d\tau^2/dt^2 = 1 - v^2/c^2$, where $v^2 = v_r^2 + v_t^2$, the square of the vector sum of the radial and transverse components of velocity $v$.

In curved spacetime, the vector sum differs, because radial velocity is affected by the curvature of space. It is illuminating to write the Schwarzschild solution in the following form:

$$\frac{d\tau^2}{dt^2} = g_{tt}[1 - g_{rr}^2\frac{v_r^2}{c^2} - g_{rr}\frac{v_t^2}{c^2}]. \tag{1.18}$$

This can be viewed as the product of a gravity related time dilation factor

($g_{tt}$) and a velocity related time dilation factor $(1 - v^2/c^2)$. One can guess the vector summation equation from $d\tau/dt$ above as:

$$v_{lo}^2 = g_{rr}^2 \frac{v_r^2}{c^2} + g_{rr} \frac{v_t^2}{c^2}, \tag{1.19}$$

where $v_{lo}$ is the velocity as measured by the *local observer* and $v_r, v_t$ are the coordinate (i.e. distant observer) velocities. A local observer must be inertial (free falling) and momentarily stationary in the reference frame at the time and place of the measurement. The timelike metric can then be simply written as

$$d\tau^2 = g_{tt}(1 - v_{lo}^2)\ dt^2, \tag{1.20}$$

relating proper time and coordinate time by the product of the gravitational time dilation (redshift) and a simple velocity time dilation.

Note that despite the local velocity being measured by a locally stationary observer, the equation gives the rate of the locally moving clock ($d\tau$) as a function of the rate of the coordinate clock ($dt$).

This shows very clearly that special relativity is a special case of general relativity. Special relativity rules when the gravitational field is weak or absent ($g_{tt} = 1$)—then it is only velocity time dilation that occurs.

From the above we have the very useful transformation formulae between local ($v_{lo}$) and coordinate ($v_{co}$) velocities in Schwarzschild spacetime (recall that $g_{rr} \geq 1$ and $g_{tt} = 1/g_{rr}$)

$$v_{r(lo)}^2 = g_{rr}^2\ v_{r(co)}^2, \tag{1.21}$$

$$v_{t(lo)}^2 = g_{rr}\ v_{t(co)}^2, \tag{1.22}$$

$$v_{r(co)}^2 = g_{tt}^2\ v_{r(lo)}^2, \tag{1.23}$$

$$v_{t(co)}^2 = g_{tt}\ v_{t(lo)}^2. \tag{1.24}$$

To make sense out of the velocity transformations, remember that someone with a slower clock (the local observer) will measure a shorter time and thus a higher speed than someone with a faster clock (the distant observer). This explains the transverse velocity transformation, but why the additional factor $g_{rr}$ for radial velocities?

This is caused by the gradient of curved space. Near the source of gravity, distances appear to be 'compressed' in the radial direction, as viewed by the distant observer. Therefore, radial movement appears to the distant observer to slow down by a further factor $g_{tt}$, as shown in figure 1.5, where geometrized units have been used for clarity, i.e. $\bar{M} = GM/c^2$, or $G = 1,\ c = 1$.

The 'compression' of radial distances by a gravitational field can be viewed as caused partially by gravitational time dilation and partially by the gradient of curved space (actually, partially here means that both have an equal share in the outcome). The gradient of curved space at a distance $r$ from a mass $M$ is given by

$$\frac{dz}{dr} = \sqrt{\frac{2\bar{M}}{g_{tt}\ r}}, \tag{1.25}$$

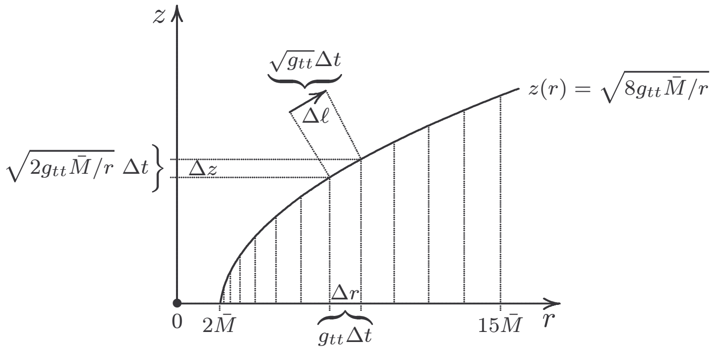

$z$
$\underbrace{\sqrt{g_{tt}}\Delta t}$
$\Delta\ell$
$z(r) = \sqrt{8g_{tt}\bar{M}/r}$
$\sqrt{2g_{tt}\bar{M}/r}\ \Delta t\}$
$\Delta z$
$\Delta r$
$0$
$2\bar{M}$
$g_{tt}\Delta t$
$15\bar{M}$
$r$

**Figure 1.5:** Proper radial distance increments ($\Delta\ell$) against coordinate radial distance increments ($\Delta r$), showing both gravitational time dilation (redshift) and space curvature. The segments $\Delta\ell$ represent the proper distance that light travel in time interval $\Delta t$, which become shorter closer to the origin due to gravitational time dilation. Then the gradient of the local space curve $z(r)$ causes the projection onto the coordinate radial axis to be 'compressed' further.

giving $z$ as a function of $r$ (after integration, with the expanded $g_{tt}$)*

*An easy to follow derivation of $z(r)$ is given in [MTW], section 23.8.

$$z(r) = \sqrt{\frac{8g_{tt}\bar{M}}{r}} + \text{constant}, \qquad (1.26)$$

as used in figure 1.5. This figure illustrates so much of the gravitational field around a static, spherically symmetric mass, that it warrants a closer look. Firstly, if space had no curvature, i.e., $z(r) =$ constant, there would still have been an apparent contraction in the radial direction, as observed by the distant observer, i.e.,

$$\Delta\ell = \sqrt{g_{tt}}\Delta t,$$

the distance that light propagates in coordinate time interval $\Delta t$.

If clocks slow down near the central mass, then so does the propagation of light, at least as viewed by the distant observer. A local observer cannot detect this 'slowing down' of light, because if your clock and your measuring rod* changes precisely in step, you cannot detect the 'slowing down'.

*All distance measurements are directly or indirectly based on the speed of light.

A distant observer can, in principle, detect the 'slowing down' of light by measuring the round trip travel time of light, beamed from a large distance to the mass and being reflected back from the surface of the mass.*

This is the effect of gravitational redshift alone. Because of the gradient of curved space, the effective movement of light in the radial direction is still 'slower' than that. From the gradient equation and since $dr^2 = d\ell^2 - dz^2$ (see figure 1.5), it is easy to show that

$$dr = \sqrt{g_{tt}}\ d\ell = g_{tt}\ dt, \qquad (1.27)$$

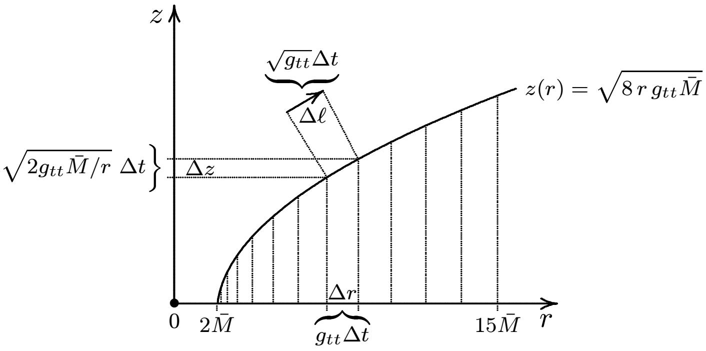


**Figure 1.5:** Proper radial distance increments ($\Delta\ell$) against coordinate radial distance increments ($\Delta r$), showing both gravitational time dilation (redshift) and space curvature. The segments $\Delta\ell$ represent the proper distance that light travel in time interval $\Delta t$, which become shorter closer to the origin due to gravitational time dilation. Then the gradient of the local space curve $z(r)$ causes the projection onto the coordinate radial axis to be 'compressed' further.

giving $z$ as a function of $r$ (after integration, with the expanded $g_{tt}$)*

*An easy to follow derivation of $z(r)$ is given in [MTW], section 23.8.

$$z(r) = \sqrt{8\,r\,g_{tt}\,\bar{M}}\ +\ \text{constant}, \tag{1.26}$$

as used in figure 1.5. This figure illustrates so much of the gravitational field around a static, spherically symmetric mass, that it warrants a closer look. Firstly, if space had no curvature, i.e., $z(r) =$ constant, there would still have been an apparent contraction in the radial direction, as observed by the distant observer, i.e.,

$$\Delta\ell = \sqrt{g_{tt}}\Delta t,$$

the distance that light propagates in coordinate time interval $\Delta t$.

If clocks slow down near the central mass, then so does the propagation of light, at least as viewed by the distant observer. A local observer cannot detect this 'slowing down' of light, because if your clock and your measuring rod* changes precisely in step, you cannot detect the 'slowing down'.

*All distance measurements are directly or indirectly based on the speed of light.

A distant observer can, in principle, detect the 'slowing down' of light by measuring the round trip travel time of light, beamed from a large distance to the mass and being reflected back from the surface of the mass.*

This is the effect of gravitational redshift alone. Because of the gradient of curved space, the effective movement of light in the radial direction is still 'slower' than that. From the gradient equation and since $dr^2 = d\ell^2 - dz^2$ (see figure 1.5), it is easy to show that

$$dr = \sqrt{g_{tt}}\ d\ell = g_{tt}\ dt, \tag{1.27}$$

*A black hole will not reflect light, but a neutron star will work nicely.

the projection of $d\ell$ onto the coordinate radial distance axis. Therefore, the radial 'contraction' factor of curved space is identical to the contraction factor of gravitational redshift, both being $\sqrt{g_{tt}}$.

This means that as far as distant observers are concerned, light moving precisely radially relative to a central mass 'slows down' to $g_{tt}c$.

Light moving (momentarily)* in a purely transverse direction relative to a mass, is slowed down just by the gravitational redshift factor, to $\sqrt{g_{tt}}\ c$.

*There is a case where light can be in orbit around a black hole, but more about that later.

Since the full effect can only be measured indirectly, it is not called a 'slowing down of light', but rather a 'delay of light'.*

*It is called the *Shapiro delay*, after the man who first measured it accurately, as is discussed further in chapter 5.

## 1.4 Summary of this introduction to relativity

We have seen that in the gravity-free space of special relativity, there is a quantity called the spacetime interval that is invariant, i.e., it has the same value, no matter which inertial observer measures the components of the interval.

This lead us to the conclusion that an observer that is present at two spacetime events will always measure a time interval that is shorter than what any observer that is not present at both events will measure. This time interval (measured by the observer present at both events) is called the propertime between two events.

The more formal view of the spacelike and the timelike interval was then derived, using just a sprinkling of tensor algebra. We then moved on to a fairly loose discussion of Einstein's field equations and an even more loose derivation of the metric for the gravitational field was presented.

This lead us to the (disturbing?) realization that the speed of light varies in Schwarzschild coordinates with the direction of movement—at least as measured by a distant observer. This forced us to accept that distant observers will measure different velocities than what local observers will measure. There are however relatively simple transformation equations for velocities as measured locally and remotely.

Underlying to this is the fact that time and distance are measured differently by distant clocks and local clocks. In particular, distant clocks runs faster than local clocks and the difference becomes more noticeable if the local clock is moving relative to the coordinate system.

Now, to move forward, we will discuss a few topics in special relativity that are of fundamental importance—clock synchronization, energy, momentum and Doppler shift. They are the sort of things that many engineers use daily and may perhaps sometimes wonder how relativity influences their work.

# Chapter 2

# Special Relativity for Engineers

clocks
energy/momentum
Doppler shift

In this chapter, we will remain (more or less) inside the realm of special relativity and look at the measurement problem posed by the seemingly peculiar behavior of light, how clocks are synchronized and how energy, momentum and Doppler shift are represented in special relativity.

## 2.1 The peculiar nature of light

We have seen that there is apparently no inertial frame in which light (or any other electromagnetic propagation) is at rest. For every uniformly moving material object there must exist an inertial frame in which the object is at rest. In that frame of reference, space and time appears completely normal, whatever "normal" might mean.

Light does not have a "normal" reference frame and it is not wise to try and cast it into such a frame. As far as we know, the reference frame for light is singular and totally undefined. On a spacetime diagram, this becomes clearer if we draw the time axis and the space axis for a moving material object.

As we will analyse a little later, the time and space axes for an inertial observer that is moving relative to a reference frame is skewed, as shown in figure 2.1. When the speed of such an observer approaches the speed of light, the $t$ and $\mathbf{x}$ axes tend to become one and the same. In this limit only spacetime events where $ct = \mathbf{x}$, i.e. light-like intervals, are possible.

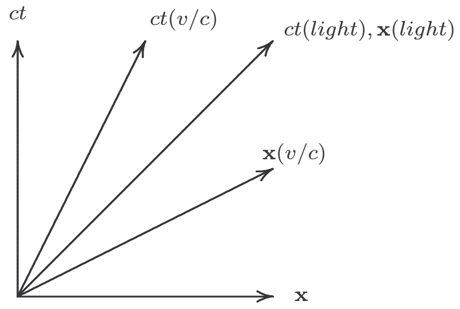


**Figure 2.1:** A Minkowski spacetime diagram showing the effective time and space axis of a material object moving at velocity $v$ in a reference system $ct, \mathbf{x}$. As the velocity approaches that of light ($v \to c$), the moving system's time and space axis almost coincide and all points approach a singular, undefined state relative to the reference frame.

We are bound to use this strange phenomenom to observe and measure most of what we are interested in. Distances, for example, are always measured directly or indirectly by using some form of electromagnetic propagation.

One can perhaps say that if we use light to measure distance, it is completely natural that we will always get the same value for the speed of light!

It is not quite that simple though. There are other factors that also have to come into play. This brings us to the question: how is the speed of light measured?

**Measuring the speed of light.** The most common procedure for obtaining the speed of light is to time the two-way travel of light over a known space interval. The reason for specifying a two-way experiment is that the same clock can then be used for the timing of the travel time of light.

As soon as the one-way speed of light needs to be measured, one needs two clocks—one at the starting point and one at the end point of the space interval. These two clocks need to be synchronized somehow, otherwise one have no idea of how long the light took to cover the space interval.

As we will soon see, there are more than one way in which two clocks can be synchronized—and the method chosen will influence the answer obtained for the one-way speed of light.

**The Galileian (or Newtonian) synchronization of clocks.** In Galileian spacetime, it is mostly a trivial exercise to synchronize remote clocks, because one can put two clocks together and synchronize them and then move one clock to a new location without upsetting the mechanism in any way. If the clocks are identical, they will stay synchronized.

If this method is not practical for some reason, one can send a messenger from the one clock to the other and, provided that we know the precise distance and the speed of the messenger, we can synchronize the clocks perfectly. The "messenger" could in principle be any material object that we shoot from the one clock's position to the other's at a known speed.

The "object" can also be light (photons). However, we will then have to know the precise velocity vector of the inertial frame relative to the aether. Why? Because in Newtonian dynamics, the velocity vector of the transmitter relative to the aether influences the speed of light.

Even if we did know the velocity of the inertial frame relative to the aether, it would be somewhat circular to synchronize clocks by means of light signals and than use those clocks to measure the speed of light.

**The Einstein synchronization of clocks.** The trivial synchronization method mentioned for the Galileian spacetime (moving one clock) cannot be used in Minkowski spacetime. Any relative movement of the clocks may desynchronize the clocks, at least to some degree.

We can however use a "messenger" with a constant known speed, because that is equally valid in Minkowski spacetime. Einstein used his L-principle to postulate that we can use light to synchronize any two clocks that are stationary in an inertial system.

However, light is just like a particle traveling at constant speed in all directions in an inertial frame. In this case, the measurement would suffer from the same circularity as mentioned above (using light to synchronize clocks for the purpose of measuring the speed of light).

**Thought experiment to measure the one-way speed of light.** Somewhere in free space, construct a large symmetrical (orthogonal) cross with arm lengths $2L$ from end to end. Let the construction move inertially and without rotation relative to the distant stars. Equip the cross with four identical atomic clocks (p,q,r,s), one at each end of an arm, as shown in figure 2.2.

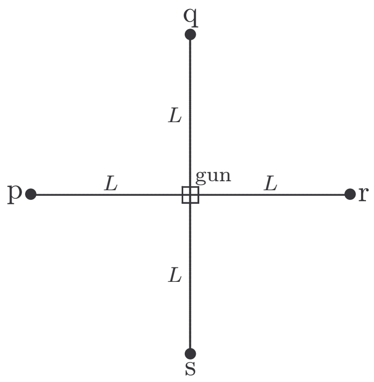


**Figure 2.2:** Four clocks (p,q,r,s) are situated at the ends of a perfect cross formed by bars of length $2L$. The gun at the centre can shoot single particles simultaneously and with equal speeds along all four bars.

Place a special gun in the exact centre of the cross in such a way that it can shoot material particles simultaneously at a known speed towards each of the four clocks. Now fire the gun at an arbitrary time and set each clock to a predetermined time (say zero) when the particles are detected at the corresponding clock.

If we ignore quantum uncertainty (Heisenberg's uncertainty principle), the clocks are then synchronized, because the particles will arrive simultaneously at the four clocks, at least in the inertial frame in which the cross is at rest.

Now the four synchronized clocks can be used to measure the one-way speed of light between any two of them. Let the inertial frame move at a constant speed $v$ relative to the aether in the direction a-c.

According to *Galileian mechanics*, the one-way speed of light will depend on the direction of the light path for the four orthogonal directions as follows

$$\begin{aligned} c_{pr} &= c - v \\ c_{rp} &= c + v \\ c_{qs} &= \sqrt{c^2 - v^2} \\ c_{sq} &= \sqrt{c^2 - v^2}. \end{aligned}$$

The first two are self explanatory, while the latter two come from simple Pythagorean geometry, as the reader can easily verify.

In Einstein's special relativity, the measured speed of light will be the same for all four directions and is simply equal to $c$. Since there is no aether in this theory, the speed of the inertial frame relative to the aether (if it exists) is zero at all times ($v = 0$).

So how do we know that Einstein had it right? Well, the thought experiment above is for all scientific purposes equivalent to the Michelson-Morley aether-drift experiment of the early 1900s.

Although Michelson and Morley did not measure the speed of light, they have shown unequivocally that there is no significant difference in the speed of light in directions normal to and directions parallel to the movement of the Earth through the hypothetical aether.

At that stage, the only known movement of the Earth was the orbital speed of the Earth around the Sun, some 30 km/s. The small fluctuations that Michelson and Morley measured were much smaller than what was expected for such a speed. Many other experiments with improved accuracy followed, but no one was ever able to detect the expected aether-drift of Earth.

Many sceptics questioned (and are still questioning up to this day) whether a two-way experiment necessarily proves Einstein right. Some sceptics insist that a specific experiment to verify the one-way speed of light be performed in a moving inertial frame.

So why has it not been done? Well, scientists are so confident that Einstein had it right that they design particle accelerators and systems like the GPS by using the details of Einstein's relativity theory. Some scientists say that every time someone uses a GPS receiver on Earth, it is a test of the one-way speed of light in a moving inertial frame.

Perhaps not quite, but if Einstein had it wrong, you can bet your bottom dollar that linear particle accelerators and the GPS system would not have worked as advertised.

So the answer as to why a one-way test has not been done—it would be money wasted to re-establish an already well-known experimental fact. The money is more wisely spent on things that utilize that known fact!

**And what if the speed of a frame changes?** Firstly, while there is any acceleration, the frame is not inertial. The moment the acceleration stops, the frame will again move at a constant speed and will be inertial. How would Galileo (or Newton) have viewed the one-way speed of light in this new inertial frame?

Say we know that the frame is moving at a speed $v$ in the direction a-c. If the change in that speed was $\Delta v$, the frame would now move at $v + \Delta v$ relative to the aether. The Galileian speed of light for the four orthogonal directions would then be

$$\begin{aligned} c_{pr} &= c - (v + \Delta v) \\ c_{rp} &= c + (v + \Delta v) \\ c_{qs} &= \sqrt{c^2 - (v + \Delta v)^2} \\ c_{sq} &= \sqrt{c^2 - (v + \Delta v)^2}. \end{aligned}$$

In this (Newtonian) view, the four clocks would still be synchronized after the acceleration, in the sense that they would at all times measure "universal time" and therefore the measured speed of light should come out as in the last set of equations.

There are immediately some objections that one can raise about this view and this set of equations. Firstly, even not allowing $v + \Delta v$ to become greater than $c$, the values of $c_{pr}$, $c_{qs}$ and $c_{sq}$ can become zero. This means that light cannot always propagate in those directions.

Secondly, the restriction $v + \Delta v < c$ also implies that $v + \Delta v + v_p < c$, where $v_p$ is the speed of a particle relative to the gun. Therefore one cannot always check the synchronization of those clocks by means of particles shot from the centre, as used before.

Thirdly, even if we allow $v + \Delta v + v_p > c$ so that a particle can reach the clocks, they would no longer seem to be synchronized in the sense that they read the same time when the four particles arrive.

That is unless $v + \Delta v + v_p \rightarrow \infty$, which seems absurd. In short, the concept of simultaneity becomes problematic. Lastly, experimental results do not support the Galileian view.

In Einstein's view the measured speed of light would still be simply $c$ in all directions relative to the frame. How could this be? Well, we would agree that the synchronization of the clocks could still be done by means of the particle gun, as before.

If the gun shot photons instead of massive particles, one can call on the wave-particle duality and argue as follows: the particle view would make the photon synchronization method equivalent to (say) an electron synchronization method. All that is required is that the speed of the specific particles are the same and remain constant during the synchronization procedure.

If one could use photons to synchronize the clocks, then the speed of light is guaranteed to come out as $c$ during any subsequent measurement thereof! Various forms of experimental results seem to support Einstein's view so far.

This brings us to the very interesting question: would the four clocks that were synchronized before the speed change still be synchronized after the speed change $\Delta v$? The answer is no.

It has been proven (and experimentally confirmed) that acceleration *per se* does not affect the time keeping of well built atomic clocks in any measurable way. However, inertial frames in relative motion will each have their own definition of simultaneity. Einstein called this the *relativity of simultaneity*.

It can be illustrated on a Minkowski spacetime diagram as in figure 2.3. The $t'$ axis is given by the equation $\mathbf{x} = (\Delta v/c)(ct) = \Delta vt$ and the $\mathbf{x}'$ axis by $ct = \Delta v\mathbf{x}/c$, found by setting $t' = 0$ and $\mathbf{x}' = 0$ in the appropriate Lorentz transformation equation.

For a relative speed of $\Delta v$, this gives a synchronization offset over a length $L$ of $\Delta vL/c$ metres, or $\Delta vL/c^2$ seconds. In the $t', \mathbf{x}'$ inertial frame, any line parallel to the $\mathbf{x}'$ axis (like the dotted line $p' \rightarrow r'$) is a line of constant time in that frame, also called a line of simultaneity [Faber].

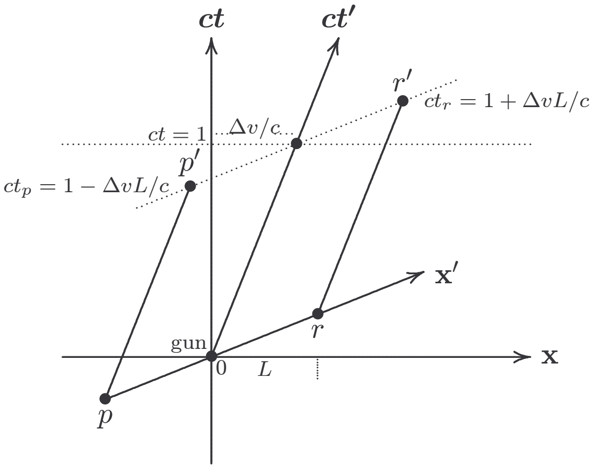


**Figure 2.3:** A Minkowski spacetime diagram where the primed frame is moving at a speed $\Delta v = 0.4c$ relative to the reference frame. The length of the arms are $L = 0.5$ units.

A good question at this stage: where does the synchronization offset come from? The easiest way to understand it is to consider the following.

While the frame is moving inertially, fire two particles within a short time $T$

in the direction of clock $r$. Fire the particles as close to the speed of light as available energy allows, so that the initial speed of the particles relative to the frame will approach $c$. *

*The particles must not be photons, because in this context, we want to synchronize the clocks so that we can measure the speed of photons.

Immediately after the second particle is fired, accelerate the frame at a moderate rate in the direction of particle movement until just before the first particle arrives at clock $r$.

Say the frame's speed changed by $\Delta v$, so that the speed of the particles relative to the frame will now be $c - \Delta v$. The change in time separation between the two particles passing clock $r$ will be (in seconds per original particle period $T$)

$$\Delta T = \frac{\Delta v}{c - \Delta v} T \cong \frac{\Delta v}{c} T \quad \text{(for } \Delta v \ll c\text{)}$$

Convert this to the period change per particle travel time ($t \cong L/c$, which is how long the acceleration lasted), obtaining

$$\Delta t \cong \Delta T \frac{t}{T} \cong \frac{\Delta v}{c^2} L$$

seconds per particle travel time, or

$$c\Delta t \cong \frac{\Delta v}{c} L \tag{2.1}$$

metres per particle travel time. This corresponds to the synchronization offset as used in the Minkowski spacetime diagram. This result is not exact in Newton mechanics, although very close due to the low acceleration and low relative speed assumptions. As indicated by the agreement with the Minkowski spacetime diagram (based on the Lorentz transformation), the result is exact in special relativity.

There is one more interesting point to be cleared up. In which direction must the two clocks be adjusted after the acceleration? The rear clock ($p$) must clearly be advanced by the synchronization offset and the front clock ($r$) set back by the same amount.

In the Minkowski spacetime diagrams (figures 2.3 and 2.4), clock $p$ is shown to be behind clock $r$, because the diagrams portray the view before the synchronization adjustments were made to the moving clocks—it is the view of the reference frame. It is only in the view of the moving frame that the new definition of simultaneity is valid.

Figure 2.4 shows the paths of the two particles as lines $0 \rightarrow p'$ and $0 \rightarrow r'$.

The particles are moving isotropically relative to the moving frame, at $v'_p$ in both directions. In the reference frame, they do not appear to move isotropically. This is expected, because the particles have the forward motion of the gun added to the firing velocity.

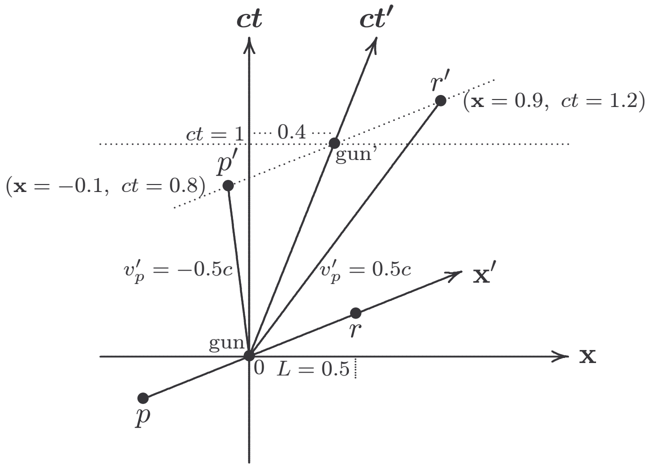


**Figure 2.4:** The paths of two particles moving at $v'_p \pm 0.5c$ relative to the gun respectively. The primed frame moves at $\Delta v = 0.4c$ and $L = 0.5$ units, as before.

From figure 2.4, we can see that the forward moving particle's speed in the reference frame is $v_{p+} = 0.9c/1.2 = 0.75c$ and the backward moving particle's speed is $v_{p-} = -0.1c/0.8 = -0.125c$ in the reference frame.

We have indirectly constructed the rule for *relativistic addition of velocities*:

$$v = \frac{v_1 + v_2}{1 + v_1 v_2/c^2}. \tag{2.2}$$

If we plug in $v_1 = 0.4c$ and $v_2 = 0.5c$, we have

$$v_+ = \frac{0.4 + 0.5}{1 + 0.4 \times 0.5}c = 0.9c/1.2 = 0.75c$$

and

$$v_- = \frac{0.4 - 0.5}{1 + 0.4 \times -0.5}c = -0.1c/0.8 = -0.125c,$$

in agreement with the Minkowski spacetime diagram analysis.

This does not agree with Galileian addition of velocities, where the forward particle would have a speed of 0.4c+0.5c = 0.9c, and the backward particle would have a speed of 0.4c-0.5c = -0.1c.

Now with the synchronization of the clocks established, we can investigate how the two frames will measure the speed of light. Let one frame fire an omni directional pulse of photons when the origins of the two inertial frames coincide.

It is immaterial which frame does the firing, unlike for the case of material particles, where it does matter. Both frames will observe the photons to move isotropically in both directions and hence both will measure a photon velocity $c$, as indicated in figure 2.5.

Obviously, to measure the one-way speed of light, each reference frame must have its own observers posted at appropriate positions. This is so that they can read their synchronized clocks at the respective locations.

In the $\mathbf{x}', t'$ frame, the observers can just ride with clocks $p$ and $r$ respectively. In the reference frame $x, t$, two observers, with clocks synchronized in the reference frame must be positioned at a distance of plus and minus $L$ light-seconds from the origin.

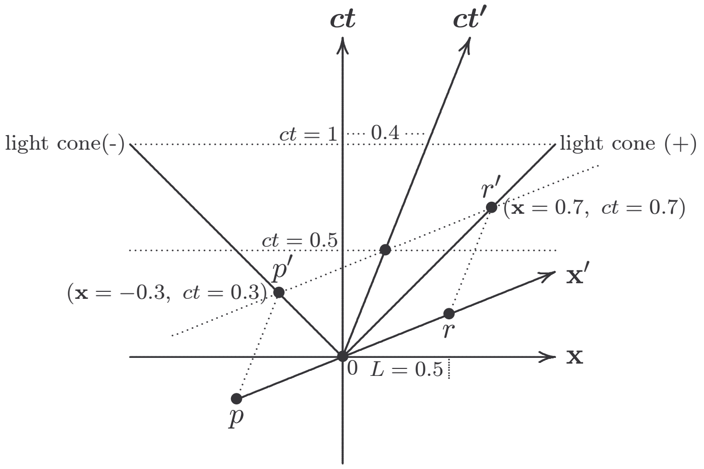


**Figure 2.5:** The light cone and the points of interception ($p'$ and $r'$) of a light pulse originating at the origin, by the two clocks $p$ and $r$. Note how the same light pulse spreads isotropically in both directions for both reference frames.

One can check the isotropy of light by means of the relativistic addition of velocities rule:

$$v_{light+} = \frac{0.4 + 1}{1 + 0.4 \times 1} c = c$$

and

$$v_{light-} = \frac{0.4 - 1}{1 + 0.4 \times -1} c = -c,$$

as is expected from the Minkowski spacetime diagram.

Note that nowhere in this analysis time dilation and/or Lorentz contraction were mentioned or brought into consideration. These concepts were neatly "brushed under the carpet" by the Lorentz transformation used in determining the synchronization offset.*

> *Lorentz transformations are essentially done in the equations for the $t'$ and the $\mathbf{x}'$ axes.

If the grid lines of the moving reference frame are drawn, time dilation and length contraction become visible, as shown in figure 2.6. The horizontally and vertically measured distances between the oblique gridlines are obviously shorter than those in the reference frame. The relative lengths correspond to the time dilation and length contraction of the Lorentz transformation.

But how do we know the separations of the grid lines of the moving frame? The $t'$ axis is "calibrated" by using the invariant timelike interval to determine the point $t' = t$ on the $t'$ axis. Likewise, the $\mathbf{x}'$ axis is "calibrated" by using the invariant spacelike interval to determine the point $\mathbf{x}' = \mathbf{x}$ on the $\mathbf{x}'$ axis.

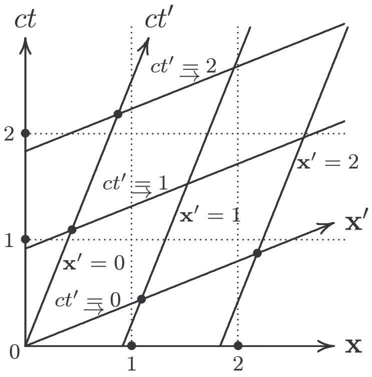


**Figure 2.6:** Although the oblique grid line distances of the moving coordinate system are longer than in the reference frame, the horizontal and vertical distances are shorter than the corresponding reference frame distances. This illustrates time dilation and Lorentz contraction on a Minkowski spacetime diagram.

It can be visualized by plotting the curves of the intervals on a spacetime diagram, as shown in figure 2.7. The curves are the hyperbolas

$(ct)^2 - \mathbf{x}^2 = \Delta s^2$ where $\Delta s = 2, 1, 1, 2,$ timelike and spacelike, as required.

Since $(ct')^2 - (\mathbf{x}')^2$ also equals $\Delta s^2$, an invariant quantity, the scale of the

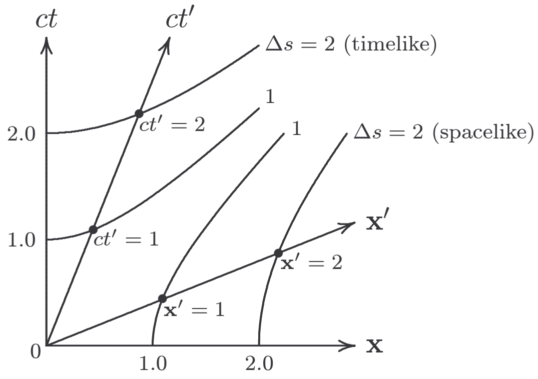


**Figure 2.7:** By making either $t'$ or $\mathbf{x}'$ zero in the equations for the spacetime interval, the intersections of the hyperbolas with the moving time and space axes can be determined. This gives the "calibration" of the moving coordinate system as observed in the reference coordinate system.

$t'$ axis is found by making $\mathbf{x}' = 0$ in the timelike equation (with $ct' > \mathbf{x}'$). This give the point where the curve crosses the $t'$ axis. Likewise, the scale of the $\mathbf{x}'$ axis is found by making $t' = 0$ in the spacelike equation (with $\Delta\mathbf{x}' > ct'$).

Time dilation and length contraction are not "real" or absolute in special relativity, i.e., in the absence of gravity. In a way, time dilation and Lorentz contraction are just spacetime projections of one inertial frame onto another.

Absolute time and space of Galileo and Newton were replaced by the spacetime interval, which is absolute and the same in all inertial frames. This is the essence of Einstein's special relativity. And as far as we can tell so far, Einstein was right.

## 2.2 Mass, energy and momentum

In special relativity, mass, energy and momentum are also transformed between two inertial frames in relative motion. The meaning of Einstein's famous equation $E = mc^2$ is that the *rest energy* equals the *rest mass* of an object, where the $c^2$ is just a conversion constant between the conventional (SI) units for mass and for energy.

If the body moves relative to the reference frame, then it's energy increases and it can be thought to have a *moving mass* that is larger than it's rest mass. We will use the symbol $\epsilon$ for moving energy (or moving mass), to distinguish it from the Newtonian energy $E$.

In conventional Newton dynamics, the momentum of a moving object is $p = mv$, where $v$ is the velocity of the mass $m$ relative to the reference frame. Newton dynamics also defines the kinetic energy of a moving mass as $E = \frac{1}{2}mv^2$.

In Einstein's special relativity, the momentum of an object moving with velocity $v$ relative to the reference frame is equivalent to the Newton momentum divided by the time dilation factor $\sqrt{1 - v^2/c^2}$:

$$p = \frac{mv}{\sqrt{1 - v^2/c^2}}. \tag{2.3}$$

The total energy $\epsilon$ is the vector sum of the rest energy and the relativistic kinetic energy, or

$$\epsilon = \sqrt{(mc^2)^2 + p^2c^2}, \tag{2.4}$$

shown graphically in figure 2.8. Note how the value of Newton momentum ($mv$) graphically relates to the relativistic momentum and energy. By solving the triangle, we get

$$\epsilon = \frac{mc^2}{\sqrt{1 - v^2/c^2}}, \tag{2.5}$$

showing that the total energy of a moving mass is the rest energy divided by the time dilation factor due to velocity.

The "strange" quantity $mvc$ in figure 2.8 comes from multiplying rest energy $mc^2$ by the value $v/c$, keeping the units of the horizontal and vertical axes the same. The horizontal axis can be viewed as the relativistic kinetic energy. The value $mvc$ scales linearly with the speed, from zero to $mc^2$ and thus rotates the total energy vector from the vertical to the horizontal.

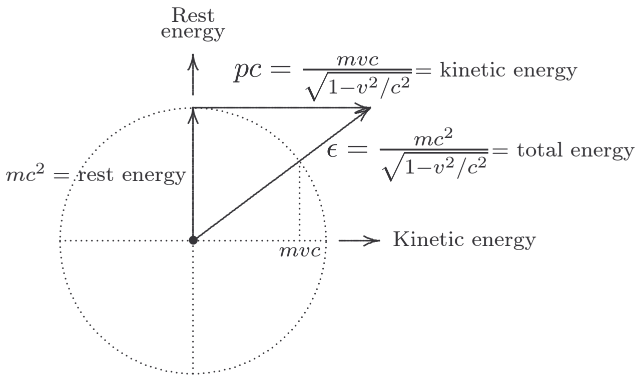

Rest energy
$pc = \frac{mvc}{\sqrt{1-v^2/c^2}}$ = kinetic energy
$\epsilon = \frac{mc^2}{\sqrt{1-v^2/c^2}}$ = total energy
$mc^2$ = rest energy
$mvc$
Kinetic energy

**Figure 2.8:** Rest energy ($mc^2$), kinetic energy ($pc$) and total energy ($\epsilon$) of a body moving with velocity $v$. When $v \to c$, $mvc \to mc^2$ which is the radius of the circle. The momentum vector and the total energy vector then tend towards being parallel to each other and will meet at a very large momentum. Both momentum and total energy tend to infinity as velocity approaches the speed of light.

If $v \ll c$ the energy equation can be approximated by

$$\epsilon = mc^2 + \frac{1}{2}mv^2, \tag{2.6}$$

stressing the fact that for low velocities, Einstein's energy equation is equivalent to adding the rest energy ($mc^2$) to the Newton kinetic energy ($\frac{1}{2}mv^2$).*

This simple relationship breaks down when velocity becomes a significant portion of the speed of light and $\epsilon$ diverges to infinity as $v$ approaches the speed of light. The only correct interpretation is that total energy equals the vector sum of the rest mass and the relativistic kinetic energy.

When we attempt to accelerate an object to a velocity close to the speed of light, the amount of energy required is enormous. It is as if the high speed object converts most of the energy applied to it into mass instead of into extra velocity.

However, suppose we could set up an inertial frame with the same velocity as the object, so that the object is effectively at rest in the frame. If someone in this frame would measure the mass of the object, perhaps by accelerating it slightly, the mass of the object would turn out to be just it's rest mass.

Rest mass is an absolute value in the sense that all frames of reference would get the same measured mass for the same object, provided it is at rest in that frame.* If the mass is moving, the measured mass depends on how fast the object is moving relative to the frame that measures it.

* If some value $a \ll 1$, then $1/\sqrt{1-a^2} \approx 1/(1-a/2) \approx 1 + a/2$.

> *Strictly, for rest mass to be constant, the object must remain at constant temperature and pressure as well.

This is contrary to Newtonian dynamics. There the measured mass of an object remains constant, irrespective of how fast it moves relative to the frame that measures it.

In Newton's view, if an observer riding with the object measures some acceleration ($a$) caused by some force ($F$) acting on the object, then an observer in some other inertial frame,in relative motion to the first one, can do the same measurements and obtain the same measured acceleration.

This is so because, in Newton mechanics, all inertial frames will measure the same change in velocity and thus the same change in momentum in the same time interval. Velocity and momentum simply adds up in a linear fashion for all (Newton) inertial frames:

$$\dot{\mathbf{x}}_2 = \dot{\mathbf{x}}_1 + \Delta\dot{\mathbf{x}}$$

and

$$p_2 = p_1 + \Delta p.$$

In relativity, velocity and momentum do not add up linearly. It can be shown that velocity adds up as

$$v_2 = \frac{v_1 + \Delta v}{1 + v_1 \Delta v / c^2}, \tag{2.7}$$

the 'relativistic addition of velocities rule' that we met before. Momentum adds up as

$$p_2 = \frac{p_1 + \Delta p}{\sqrt{1 - v_1^2/c^2}\ \sqrt{1 - \Delta v^2/c^2}}. \tag{2.8}$$

where $v_1$, $v_2$, $p_1$ and $p_2$ are as measured in the inertial *reference frame*, while $\Delta v$ and $\Delta p$ are measured in an inertial frame moving at velocity $v_1$ relative to the reference frame.

It is equivalent to a mother ship that cruises inertially (the reference frame) and sends off a probe that reaches a stable velocity $v_1$ relative to the ship, so that the probe is the moving inertial frame. Then the probe shoots out a projectile with velocity $\Delta v$ relative to the itself (the probe). The equations tell us how the mother ship will measure the velocity and momentum of the projectile.

*The relativistic momentum summation is equivalent to the Newton summation divided by the product of two time dilation factors, one due to the base velocity and one due to the $\Delta$velocity.*

Once the new momentum is known, the new velocity can be extracted from the relationship $p_2 = mv_2/\sqrt{1 - v_2^2/c^2}$, yielding the velocity addition rule given above.

It is clear from the equations that in the extremes, momentum can tend to infinity, while velocity can never exceed 1. As a check, put $v_1 = c$ and $\Delta v = c$ into both the velocity addition and the momentum addition rules.

## 2.3 One way Doppler shift

To conclude this chapter on special relativity essentials, we will briefly examine the relativistic Doppler shift, as measured between inertial frames in relative motion.

First we will look at the one way Doppler shift. We start with the Newton Doppler shift for the general case where both the transmitter (Tx) and the receiver (Rx) are moving. By 'moving' we mean moving relative to Newton's "absolute frame of rest"—the frame where the speed of light equals $c$.

In such a case, one is forced to perform a linear transformation from the moving transmitter frame to the absolute frame and again from the absolute frame to the moving receiver frame.

In this section, the normalized speed parameter $\dot{\mathbf{x}} = v/c$ will be used for clarity. Let the transmitter Tx move at a speed $\dot{\mathbf{x}}_{\mathsf{Tx}}$ and the receiver Rx at a speed $\dot{\mathbf{x}}_{\mathsf{Rx}}$, both relative to the absolute frame.

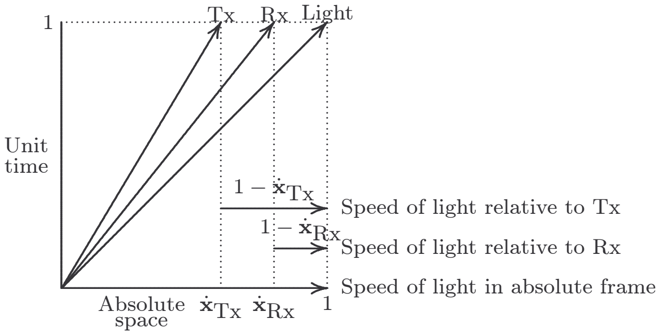

1
Tx
Rx
Light
Unit time
$1 - \dot{\mathbf{x}}_{\mathrm{Tx}}$
Speed of light relative to Tx
$1 - \dot{\mathbf{x}}_{\mathrm{Rx}}$
Speed of light relative to Rx
Speed of light in absolute frame
Absolute space
$\dot{\mathbf{x}}_{\mathrm{Tx}}$
$\dot{\mathbf{x}}_{\mathrm{Rx}}$
1

**Figure 2.9:** In Newton dynamics the speed of light changes relative to the moving transmitter (Tx) and the moving receiver (Rx). This causes the wavelength received by Rx to be a factor $\frac{1-\dot{\mathbf{x}}_{\mathsf{Tx}}}{1-\dot{\mathbf{x}}_{\mathsf{Rx}}}$ times the wavelength transmitted by Tx. The diagram was drawn for $\dot{\mathbf{x}}_{\mathsf{Tx}} = 0.6$ and $\dot{\mathbf{x}}_{\mathsf{Rx}} = 0.8$, giving $\lambda_{\mathsf{Rx}}/\lambda_{\mathsf{Tx}} = 2$ or $\Delta\lambda/\lambda_{\mathsf{Tx}} = 1$.

When the transmitter wavelength ($\lambda_{\mathsf{Tx}}$) is transferred to the absolute frame, the wavelength becomes

$$\lambda_a = \lambda_{\mathsf{Tx}}(1 - \dot{\mathbf{x}}_{\mathsf{Tx}}), \tag{2.9}$$

where $1 - \dot{\mathbf{x}}_{\mathsf{Tx}}$ is the Newton speed of light relative to the Tx frame (see figure 2.9).

As a check, consider what would happen if the transmitter could move at near the speed of light relative to the absolute frame—the wavelength transmitted to absolute space would approach zero, i.e., approaching infinite frequency.

When the wavelength in the absolute frame is transferred to the Rx frame, the wavelength becomes

$$\lambda_{\mathsf{Rx}} = \lambda_a/(1 - \dot{\mathbf{x}}_{\mathsf{Rx}}), \tag{2.10}$$

where $1-\dot{\mathbf{x}}_{\mathsf{Rx}}$ is the speed of light relative the Rx frame. Here, if the receiver could move at the speed of light relative to absolute space, the received wavelength would be infinite, i.e., no signal would be received.

From these values, the ratio of received to transmitted signal wavelength can be found as

$$\frac{\lambda_{\mathsf{Rx}}}{\lambda_{\mathsf{Tx}}}=\frac{1-\dot{\mathbf{x}}_{\mathsf{Tx}}}{1-\dot{\mathbf{x}}_{\mathsf{Rx}}}. \tag{2.11}$$

The *Doppler shift* is defined as the change in wavelength as a fraction of the transmitted wavelength,

$$\Delta\lambda/\lambda_{\mathsf{Tx}}=(\lambda_{\mathsf{Rx}}-\lambda_{\mathsf{Tx}})/\lambda_{\mathsf{Tx}}, \tag{2.12}$$

which works out to be

$$\frac{\Delta\lambda}{\lambda_{\mathsf{Tx}}}=\frac{1-\dot{\mathbf{x}}_{\mathsf{Tx}}}{1-\dot{\mathbf{x}}_{\mathsf{Rx}}}-1. \tag{2.13}$$

The reason for having labored this point is that one-way Newton Doppler shift does not depend on the relative speed between transmitter and receiver, *but rather on their speeds relative to the absolute frame of reference*, measured in the direction of the line of sight (i.e. radial speeds).

It is only at very low speed that the usual approximation $\Delta\lambda/\lambda_{\mathsf{Tx}}=\dot{\mathbf{x}}$ holds, where $\dot{\mathbf{x}}$ is the relative radial speed. For high speeds, how would one extract the relative speed between receiver and transmitter $(\dot{\mathbf{x}}_{\mathsf{Rx}}-\dot{\mathbf{x}}_{\mathsf{Tx}})$ from the measured Doppler shift?

It is only possible if you know at least one of the speeds relative to the absolute frame as well, because there are many combinations of $\dot{\mathbf{x}}_{\mathsf{Tx}}$ and $\dot{\mathbf{x}}_{\mathsf{Rx}}$ that will give the same Newton Doppler shift. For example, speeds of 0.6c and 0.8c will give the Doppler shift

$$\frac{\Delta\lambda}{\lambda_{\mathsf{Tx}}}=\frac{1-0.6}{1-0.8}-1=1,$$

which is the same as what you would get for speeds of 0.96c and 0.98c, i.e.,

$$\frac{\Delta\lambda}{\lambda_{\mathsf{Tx}}}=\frac{1-0.96}{1-0.98}-1=1,$$

while the relative speeds differ by an order of magnitude (0.2c versus 0.02c). This is a consequence of the involvement of the absolute frame of reference where light has a constant speed.

In special relativity, there is no absolute reference frame, at least as far as observables are concerned, so one simply choose either transmitter or receiver as the reference frame.

Let us choose the transmitter as the 'stationary' reference frame and let the receiver move away from it at a radial speed of $\dot{\mathbf{x}}$. The received wavelength would be like the Newton receiver moving through absolute space

$$\lambda'_{\mathsf{Rx}}=\lambda_a/(1-\dot{\mathbf{x}}_{\mathsf{Rx}}).$$

Since the receiver's distance measurements (and thus also wavelength measurements) are Lorentz contracted by $\sqrt{1-\dot{\mathbf{x}}^2}$, the received wavelength will be

$$\lambda_{\mathsf{Rx}} = \frac{\sqrt{1-\dot{\mathbf{x}}^2}}{1-\dot{\mathbf{x}}}\,\lambda_{\mathsf{Tx}} = \sqrt{\frac{1+\dot{\mathbf{x}}}{1-\dot{\mathbf{x}}}}\,\lambda_{\mathsf{Tx}}.$$

If we choose the receiver as the 'stationary' reference frame, then the transmitter is moving at $-\dot{\mathbf{x}}$ relative to the receiver. Again it is like the Newton case of transmitter transferring signal to the absolute rest frame (with negative velocity), i.e.,

$$\lambda'_{\mathsf{Rx}} = \lambda_a[1-(-\dot{\mathbf{x}}_{\mathsf{Rx}})].$$

Since the moving transmitter's distance measurements (and thus also the wavelength transmitted) are Lorentz contracted, the received wavelength will be

$$\lambda_{\mathsf{Rx}} = \frac{1+\dot{\mathbf{x}}}{\sqrt{1-\dot{\mathbf{x}}^2}}\,\lambda_{\mathsf{Tx}} = \sqrt{\frac{1+\dot{\mathbf{x}}}{1-\dot{\mathbf{x}}}}\,\lambda_{\mathsf{Tx}},$$

which boils down to the same value as for the stationary transmitter. The usual Doppler shift expression is then

$$\frac{\Delta\lambda}{\lambda_{\mathsf{Tx}}} = \sqrt{\frac{1+\dot{\mathbf{x}}}{1-\dot{\mathbf{x}}}} - 1. \qquad (2.14)$$

It can be easily shown* that when $\dot{\mathbf{x}} \ll 1$, this expression reduces to the Newtonian approximation $\frac{\Delta\lambda}{\lambda_{\mathsf{Tx}}} \approx \dot{\mathbf{x}}$.

This then, is the case for one way Doppler shifts. What happens if we consider the two way Doppler shifts, as applicable to Doppler radar?

## 2.4 Two way Doppler shift

This little section is specially for radar engineers, who might have got a little anxious, reading about the 'errors' they are making with their usual Doppler radar equations.

Here is some consolation: when a two way situation is considered in a Newton absolute frame, the effect of the absolute frame of reference cancels out!

One can laboriously go through a process of multiple translations to the absolute frame and back (four translations in all), just to find that the absolute frame disappears and you get a total wavelength ratio of

$$\frac{\lambda_{\mathsf{Rx}}}{\lambda_{\mathsf{Tx}}} = \frac{1+\dot{\mathbf{x}}}{1-\dot{\mathbf{x}}}\,,$$

or a Doppler shift of

$$\frac{\Delta\lambda}{\lambda_{\mathsf{Tx}}} = \frac{1+\dot{\mathbf{x}}}{1-\dot{\mathbf{x}}} - 1. \qquad (2.15)$$

*By noting that for $\dot{\mathbf{x}} \ll 1$, $1/(1-\dot{\mathbf{x}}) \approx 1+\dot{\mathbf{x}}$.

*This is precisely the relativistic two way Doppler shift*, because there the wavelength is changed twice by the factor $\sqrt{\frac{1+\dot{\mathbf{x}}}{1-\dot{\mathbf{x}}}}$ (once each way). When $\dot{\mathbf{x}} \ll 1$, or in conventional units, $v \ll c$, the usual approximation yields

$$\frac{\Delta\lambda}{\lambda_{\mathsf{Tx}}} \approx (1+\dot{\mathbf{x}})^2 - 1 \approx 2\dot{\mathbf{x}}, \tag{2.16}$$

because $\dot{\mathbf{x}}^2$ is small enough to be neglected. This is the relationship that every radar engineer knows. For just about every radar measurement, even in spaceflight, this approximation is good enough.

Earth's orbital velocity is some 30 km/s. Then there are some meteorites that comes head-on towards Earth at near solar escape velocity, about 40 km/s relative to the Sun. This gives a velocity relative to Earth of near 70 km/s, or $2 \times 10^{-4}c$.

So $\dot{\mathbf{x}}^2 = 4 \times 10^{-8}$, which is still small enough for the errors to be small compared to the accuracy of current Doppler radars. It is unlikely that we will use a Doppler radar to measure anything much faster than that.

It is different in astronomy, where we measure one way Doppler shifts of astronomical objects, which may have radial speeds that are significant fractions of the speed of light.

## 2.5 Some relativistic Doppler calculations

This section may be skipped by those not interested in reading calculations. It contains nothing new, but working through multiple frames of reference gives some quite interesting insights.

For simplicity we shall use a transmitter moving at $\dot{\mathbf{x}}_{\mathsf{Tx}} = 0.6$ and a receiver moving at $\dot{\mathbf{x}}_{\mathsf{Rx}} = 0.8$, both relative to some arbitrary frame of reference (labeled with the subscript $ref$, as shown in figure 2.10). These values are 'friendly', since $\sqrt{1-0.6^2} = 0.8$ and $\sqrt{1-0.8^2} = 0.6$, nice round numbers!

Let the transmitted wavelength be $\lambda_{\mathsf{Tx}} = 0.1$ lightseconds (i.e., a frequency of 10 Hz, not quite an electromagnetic frequency, but then, a nice easy number to illustrate with). To set a baseline, we will first do the one way Doppler translation and do so through the reference frame.

The reference frame is effectively a 'receiver', which moves relative to the transmitter at -0.6.

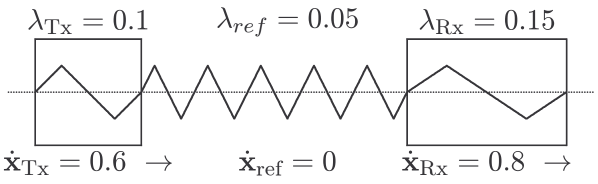

$\lambda_{\mathrm{Tx}} = 0.1$
$\lambda_{ref} = 0.05$
$\lambda_{\mathrm{Rx}} = 0.15$
$\dot{\mathbf{x}}_{\mathrm{Tx}} = 0.6$ →
$\dot{\mathbf{x}}_{\mathrm{ref}} = 0$
$\dot{\mathbf{x}}_{\mathrm{Rx}} = 0.8$ →

**Figure 2.10:** Relativistic Doppler shifts in wavelength, from the transmitter frame (left), through an arbitrary reference frame, to the receiver frame (right).

The transmitter wavelength translates to a reference frame wavelength of

$$\lambda_{ref} = 0.1\sqrt{\frac{1-0.6}{1+0.6}} = 0.05.$$

Then we translate from the reference frame to the actual receiver frame, which moves relative to the reference frame at speed 0.8c, giving

$$\lambda_{\mathsf{Rx}} = 0.05\sqrt{\frac{1+0.8}{1-0.8}} = 0.15.$$

This gives a relativistic Doppler shift from transmitter to receiver of

$$\frac{\Delta\lambda}{\lambda_{\mathsf{Tx}}} = \frac{0.15}{0.1} - 1 = 0.5,$$

a positive Doppler shift (i.e., a redshift) of half the transmitter wavelength.

Now, to compare, we will do a one-step transformation from transmitter to receiver, using the full relativistic equation. For this we need the relative speed of the receiver to the transmitter (relativistic subtraction of speed), which is

$$\dot{\mathbf{x}} = \frac{0.8-0.6}{1-0.8\times 0.6} \cong 0.3846,$$

giving a Doppler shift of

$$\frac{\Delta\lambda}{\lambda_{\mathsf{Tx}}} = \sqrt{\frac{1+0.3846}{1-0.3846}} - 1 = 0.5,$$

the same as for the two step transformation.

The two-step transformation may look like a way of determining the speed of the transmitter and the receiver relative to Newton's absolute frame of reference. Not so, because we chose an arbitrary reference frame, of which there are an infinite number.

There is also an infinite number of combinations of $\dot{\mathbf{x}}_{\mathsf{Tx}}$ and $\dot{\mathbf{x}}_{\mathsf{Rx}}$ that will give exactly the same Newtonian relative velocity and the same Newtonian Doppler shift. Special relativity dictates that a specific Doppler shift always implies one specific relative velocity.

Finally, let us look at the two way, radar type Doppler shift. We can do this in a one step calculation and find, quite simply

$$\frac{\Delta\lambda}{\lambda_{\mathsf{Tx}}} = \frac{1+0.3846}{1-0.3846} - 1 = 1.25,$$

meaning that the radar receiver will receive a wavelength of 1.25 times the transmitted frequency.

As a check on this value, we can work through the reference frame again, which gives a more intuitive 'feel' to the numbers. To this end, we have already done the first half of the calculations, i.e., up to the receiver (the $\lambda_{\mathsf{Rx}} = 0.15$ obtained above).

Now the ‘target’, reflecting the radar signal, is equivalent to a transmitter, moving at a speed 0.8 relative to the reference frame, which becomes the ‘receiver’, giving

$$\lambda_{ref_2} = 0.15\sqrt{\frac{1+0.8}{1-0.8}} = 0.45.$$

After this, we must go from the reference frame (which is equivalent to a transmitter) to the actual radar receiver. The reference frame is moving at speed -0.6 relative to the radar receiver, giving

$$\lambda_{\mathsf{Tx}} = 0.45\sqrt{\frac{1-0.6}{1+0.6}} = 0.225.$$

Since the original transmitted wavelength was 0.1, the total Doppler shift is

$$\frac{\Delta\lambda}{\lambda_{\mathsf{Tx}}} = \frac{0.225}{0.1} - 1 = 1.25,$$

the same as obtained through the one step calculation.

## 2.6 Summary

We have seen a fairly elaborate discussion of the synchronization of clocks, where we have seen how the absolute time frame was discarded through the relativity of simultaneity. This led to the Minkowski spacetime diagram where it was shown that the speed of light is indeed isotropic for all inertial observers in this representation.

The correlations between mass, energy and momentum from a special relativity viewpoint was discussed next. We saw that the rest mass of an object is essentially constant (unless it loses energy through radiation) and that one can view the moving mass to be more than the rest mass, because mass and energy is equivalent to each other.

The very important phenomenom Doppler shift was treated and the differences between Newtonian and relativistic Doppler shifts explained. We learned that one-way Newtonian Doppler shift depends on a static aether and that a specific Doppler shift can mean an infinite number of relative radial speeds.

Relativistic one-way Doppler shift is purely dependant upon the relative speed between the transmitter and receiver. We also learned that while one-way Doppler shifts differ markedly between the Newtonian and relativistic cases, the two-way Doppler shift shows less differences. The usual radar distance equation is simply an approximation for the low speed case.

We will now step off the purely inertial restriction and investigate how acceleration can be brought into special relativity. This is the topic of the next chapter.

# Chapter 3

# Linear acceleration and Relativity

when world-lines
are
not straight

So far we have only considered objects moving uniformly relative to some inertial reference frame. Such objects have straight world-lines in Minkowski spacetime. It then follows that objects undergoing acceleration should have curved world-lines. Although accelerated frames of reference are not strictly part of special relativity, we will investigate them as a link between special an general relativity. Before we go there, it will be useful to consider a different way of picturing spacetime—the so called space-propertime diagram.

## 3.1 Space-propertime diagrams

The usual way of drawing spacetime diagrams is to use coordinate time versus coordinate space. Some interesting insights can be gained by drawing propertime intervals against coordinate space intervals. Such diagrams are also known as Brehme diagrams.

In simple terms it means that if Jim is stationary in the inertial coordinate system and Pam is moving uniformly relative to him, we plot Jim's space intervals and Pam's propertime intervals against each other. Such a combination has been described by some authers, e.g., [Thorne], as *"my space and your time"*.

One of the consequences of using this curious mix of coordinates is that one can divide 'my space' by 'your time' and get an answer that exceeds the speed of light. Never the less, if one use it with caution, space-propertime diagrams are very useful and sometimes surpass the insight that can be gained from normal spacetime diagrams.

If Pam is also moving uniformly, we can plot Pam's space intervals and Jim's propertime intervals; it gives an equivalent, yet mirror-imaged picture, as shown in figure 3.1.

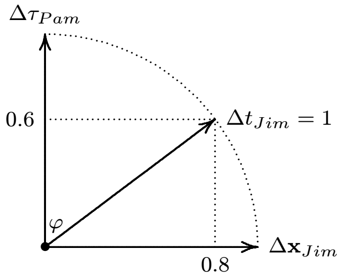


**Figure 3.1:** A space-propertime diagram where Pam moves uniformly at 0.8c relative to Jim (left) and where Jim moves uniformly at -0.8c relative to Pam (right). Note that the distances and times are intervals between events.

The figure is drawn for a very fast relative velocity (0.8c). What is immediately obvious, is that the slope of this coordinate time arrow ($t$) is less than 45 degrees (at least for positive velocities). Contrast this with Minkowski spacetime diagrams, where the slope of the (positive) speed of light is exactly 45 degrees.

What is the slope of the space-propertime arrow for light? The answer is zero or 180 degrees, depending on direction. This is so because a lightlike interval between two events represents a zero propertime interval.

The angle that the time arrow makes with the propertime axis ($\tau$) is

$$\varphi = -\arcsin(\dot{\mathbf{x}}) = -\arcsin(v/c),$$

in accordance with the answer to the question above, where positive angles are taken as per the usual convention.

We will use the space-propertime concept extensively in the discussion of accelerated objects that follows. We will see in later chapters that gravitational acceleration is also easily visualized in space-propertime diagrams.

## 3.2 Accelerated frames of reference

It is obvious that if objects are accelerated and their speeds change relative to the reference frame, they must have curved space-propertime arrows. Curves are normally characterized by curvature or radius of curvature and a centre of curvature, all of which may change along the curve.

A circle is the simplest case, where the curvature is constant and equals the inverse of the radius of the circle. The centre of curvature is simply the centre of the circle. Most other curves can be broken up into a large number of circle segments, each part of the so called *circle of curvature* for a point on the curve.

A straight line is an obvious exception, because the curvature is zero, with an infinite radius of curvature. We will use the circle of curvature to construct space-propertime diagrams for accelerating frames of reference, a simple example of which is shown in figure 3.2.

This is a circle segment and the curve has constant curvature $1/R$. Constant curvature does however not represent constant acceleration—not in the rest frame, neither in the accelerating frame. The reason for this will become clear later in this section.

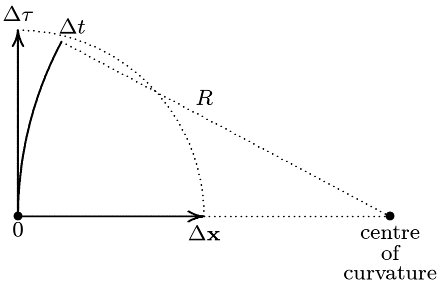


**Figure 3.2:** The wordline of an accelerating frame, shown here with a constant curvature around the centre of curvature and radius of curvature $R$. The curved worldline $(0, \Delta t)$ of the accelerating frame has the same length as the $\Delta\tau$ axis length $(0, \Delta\tau)$.

## 3.3 Transformation of acceleration

As a first step in studying accelerated frames, we need to know how to transform an acceleration measured inside a moving frame of reference back to the inertial reference frame. The acceleration $\ddot{\mathbf{x}}'$ measured inside a moving frame $\mathbf{x}'$ transforms to the rest frame as

$$\ddot{\mathbf{x}} = (1 - \dot{\mathbf{x}}^2)^{\frac{3}{2}}\ \ddot{\mathbf{x}}', \tag{3.1}$$

where $\dot{\mathbf{x}}$ is the instantaneous speed of the accelerating frame relative to the inertial frame of reference. This result is given as face value, but it is analyzed further in the box on page 62.

The value $(1 - \dot{\mathbf{x}}^2)^{\frac{1}{2}}$ is the velocity time dilation factor. Therefore, the rest frame measures an acceleration that is *three factors of velocity time dilation smaller* than what is measured inside the accelerating frame.

As an example, accelerate a test object at 1g (from rest) inside a spaceship that is moving uniformly at 0.6c relative to the inertial reference frame. Assume that the acceleration is in the direction of motion. An observer stationary in the *reference frame* will measure the test object's initial acceleration as

$$\ddot{\mathbf{x}} = (1 - 0.6^2)^{\frac{3}{2}} = 0.512\ \text{g}.$$

This will also hold for the case where the whole spaceship is accelerating relative to the reference frame. The velocity will then be changing continously, so the transformation only holds instantaneously.

## 3.4 The effect of acceleration on time

Let the spaceship accelerate uniformly at 1g as measured by some form of accelerometer on board of the spaceship. Let Pam ride the accelerating spaceship, while Jim remains stationary in the inertial reference frame.

Jim will measure Pam's acceleration as declining with her speed until, as she approaches the speed of light relative to him, he will observe her acceleration to approach zero.

In Pam's (accelerating) frame of reference, the measured acceleration will however remain at 1g for as long as her ship's propulsion system functions properly.

If Jim take Pam's constant acceleration ($\ddot{\mathbf{x}}_P$), he can transform it to his frame of reference as

$$\ddot{\mathbf{x}}_J = (1 - \dot{\mathbf{x}}^2)^{\frac{3}{2}}\ \ddot{\mathbf{x}},$$

where $\dot{\mathbf{x}}$ is Pam's changing relative speed at any moment.

Jim then integrates this expression twice with respect to time and finds the relationship between the space distance ($\mathbf{x}$) that Pam travel in his frame and the proper time ($\tau$) on board her ship for any given constant acceleration.

In other words, Jim can plot a space-propertime diagram for Pam's ship. The result of the double integration is the exponential function*

$$\mathbf{x}_J = \frac{e^{\ddot{\mathbf{x}}\tau} + e^{-\ddot{\mathbf{x}}\tau} - 2}{2\ddot{\mathbf{x}}}. \tag{3.2}$$

The function is shown graphically in Figure 3.3.

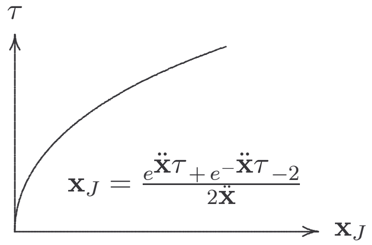


**Figure 3.3:** The space-propertime path for an object accelerating at a constant rate $\ddot{\mathbf{x}}$, as measured in the accelerating frame.

*Relativists call this ‘hyperbolic motion’ and writes it as $\mathbf{x} = \frac{cosh(\ddot{\mathbf{x}}'\tau')-1}{\ddot{\mathbf{x}}'}$, (e.g., [Thorne], notes section, referred to page 37), which boils down to the same thing.

It is not easy to extract Pam's propertime ($\tau$) out of the equation. Therefore it is left in the reciprocal form ($\mathbf{x}_J$ as a function of $\ddot{\mathbf{x}}$ and $\tau$).

For a very long space trip, even with moderate acceleration, $\ddot{\mathbf{x}}\tau \gg 1$, meaning $(e^{-\ddot{\mathbf{x}}\tau} - 2)$ becomes negligibly small compared to $e^{\ddot{\mathbf{x}}\tau}$.

The equation for the space interval in Jim's frame then reduces to

$$\mathbf{x}_J \approx \frac{e^{\ddot{\mathbf{x}}\tau}}{2\ddot{\mathbf{x}}}. \tag{3.3}$$

From this equation Pam's time interval $\tau$ can be easily extracted as

$$\tau \approx \frac{ln(2\ddot{\mathbf{x}}\mathbf{x}_J)}{\ddot{\mathbf{x}}}. \tag{3.4}$$

We will later use these approximations to analyze long duration space travel under constant acceleration.

Now back to the curve of figure 3.3. It is clear that the *curvature* is at a maximum at the origin ($\mathbf{x}_J, \tau = 0$) and at a minimum when $\mathbf{x}_J$ is large. At the origin, the *radius of curvature* can be shown to be $R_0 = 1/\ddot{\mathbf{x}}$.

At other points on the curve, the radius of curvature is enlarged by a factor $1/(1 - \dot{\mathbf{x}}_J^2)$, as discussed later in the chapter.

Of further interest is the fact that the length of the curve equals the coordinate time (Jim's time) that elapsed. This length is given by

$$t = \frac{(e^{\ddot{\mathbf{x}}\tau} - e^{-\ddot{\mathbf{x}}\tau})}{2}, \tag{3.5}$$

where $t$ is the time that Jim measures.

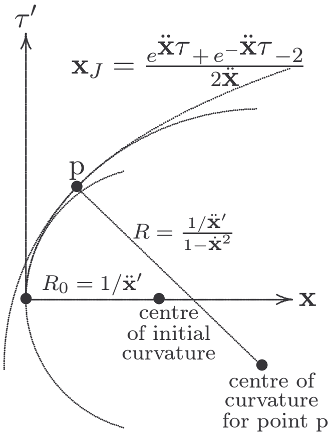

$\tau'$
$\mathbf{x}_J = \frac{e^{\ddot{\mathbf{x}}\tau} + e^{-\ddot{\mathbf{x}}\tau} - 2}{2\ddot{\mathbf{x}}}$
p
$R = \frac{1/\ddot{\mathbf{x}}'}{1-\dot{\mathbf{x}}^2}$
$R_0 = 1/\ddot{\mathbf{x}}'$
$\mathbf{x}$
centre of initial curvature
centre of curvature for point p

**Figure 3.4:** Circles of curvature along the acceleration curve can be obtained from the slope of the curve. The velocity $\dot{\mathbf{x}}$ is a function of the slope of the curve and with the velocity known, the radius of curvature $R$ and the centre of curvature can be found geometrically. From the equation for $R$ it is clear that as velocity $\dot{\mathbf{x}}$ tends to the speed of light, the radius of curvature $R$ will tend to infinity and the curvature will tend to zero, meaning acceleration relative to the rest frame ($\mathbf{x}\tau'$) will tend to zero.

The centre of curvature for any point is found by drawing a line normal to the curve in the direction that the line curves and with a length equal to the radius of curvature $R$. In order to find the radius of curvature, we need both the velocity ($\dot{\mathbf{x}}$) and the acceleration ($\ddot{\mathbf{x}}'$) at that point of the curve.

We know the acceleration and the velocity can be found from the slope of the curve, because the slope represents the velocity vector at that point. Once we have velocity and acceleration, the radius of curvature is known and we know where the centre of curvature is located, as shown in Figure 3.4.

When we do not have a curve and want to construct it from scratch, the method is similar. The simplest algorithm involves keeping track of the angle $\varphi$ through which the velocity vector has turned, as shown in figure 3.5. Decide on a time increment $\Delta t$ and from the old curve position (p) and the old centre of curvature, draw a circle segment $\Delta t$, giving $\Delta\varphi = \Delta t/R$.

You now have $\varphi_{new} = \varphi + \Delta\varphi$. For this new position, find the new radius of curvature, which together with $\varphi_{new}$ gives the new centre of curvature. The process can then be repeated for as many cycles as you like. By making $\Delta t$ very small, a highly accurate curve can be obtained. The box on page 64 gives a programming algorithm for constructing such a curve.

The algorithm is simplified by first doing a $\Delta\varphi$ rotation with distance $R$ on the x-axis of a 'dummy' coordinate system $\mathbf{x}, y$, obtaining $\Delta\mathbf{x}$ and $\Delta y$. Then a coordinate system rotation is performed through the angle $\varphi$, giving the space and propertime movements $\Delta\mathbf{x}$ and $\Delta\tau'$. The construction method of finding the space-propertime curve for an accelerating frame is useful when the acceleration changes with time in a complex way. If the acceleration is kept constant, it is easy to calculate values straight from the equations.

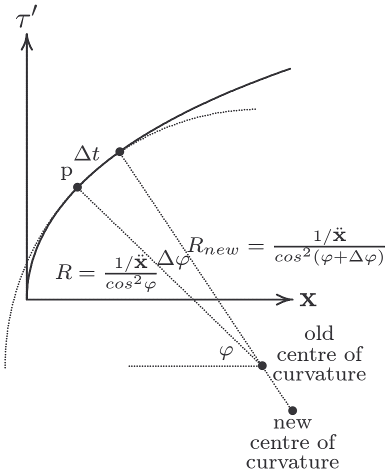

$\tau'$
$\Delta t$
p
$R_{new} = \frac{1/\ddot{\mathbf{x}}}{cos^2(\varphi+\Delta\varphi)}$
$\Delta\varphi$
$R = \frac{1/\ddot{\mathbf{x}}}{cos^2\varphi}$
$\mathbf{x}$
$\varphi$
old centre of curvature
new centre of curvature

**Figure 3.5:** Construction of the acceleration curve in space-propertime. The coordinate time interval $\Delta t$ together with the current radius of curvature $R$ determines the new position on the curve and also $\Delta\varphi$, giving the new $\varphi$, which determines a new radius of curvature and a new centre of curvature.

Using the approximate formulas, we will now proceed to do a few interesting calculations.

Firstly, we will work out how long (in spaceship propertime) it will take a spaceship, accelerating at a comfortable 'one earth gravity' $(1g)$, to travel a distance equal to the radius of the observable universe.

Secondly, we will determine how far the ship will go in the expected productive lifetime of a human spacefarer aboard.

If we work in geometric units of years, then 1g equals about 1 lightyear/year$^2$, so $\ddot{\mathbf{x}} \approx 1$. The radius of the observable universe, in light travel time, is presently $15 \times 10^9$ lightyears at most.

The proper time needed to travel a distance equal to the radius of the observable universe is

$$\tau' \;\approx\; ln(2\mathbf{x}\ddot{\mathbf{x}})/\ddot{\mathbf{x}} \;=\; ln(2 \times 15 \times 10^9) \;\approx\; 24 \text{ years},$$

well within a productive human lifetime.

In normal reference frame time though, the journey would take as long as the present age of the universe. Assuming a near constant expansion rate, the observable universe would by then have grown to about double it's present size.

Next, let us find out how far a human spacefarer, who keeps up 1g acceleration all the time, can go in about 30 years of 'productive' propertime.

$$\mathbf{x} \;\approx\; e^{\ddot{\mathbf{x}}\tau'}/2\ddot{\mathbf{x}} \;=\; e^{30}/2 \;\approx\; 5 \times 10^{12} \text{ lightyears},$$

about 300 times the radius of the observable universe. This means that if the universe were not expanding, our spacefarer could possibly have circumnavigated the entire universe many times!

But what is the use of space-travel at the nearly the speed of light? One will not be able to observe much! There is a space travel 'trick' for reaching a distant star system in relative comfort and then to stop there. This is so that one can observe what is going on.

The trick is to accelerate half the distance there at 1g and then reverse the engine's thrust. This way you can decelerate at 1g for the second half of the way and 'stop' near the star system. Well, approximately, at least. You should be able to observe, despite some residual velocity.

What will be the distance that you can reach if you still have say 30 years of expected life left? We shall work out how far you get in 15 years of constant acceleration and then double it, because the deceleration phase will take as long as the accelerating phase. The answer is

$$\mathbf{x} \approx 2 \times e^{15}/2 \approx 3 \text{ million lightyears},$$

which is not all that far on a cosmological scale. The Andromeda galaxy, part of our Local Group of galaxies, is some 2 million lightyears away. If you time your mid-point carefully at around 14.5 years, you can cruise to a 'stop' somewhere inside Andromeda.

An interesting question: what maximum speed, relative to the reference frame, would you have reached at the halfway mark? The speed after a long time of sustained 1g acceleration ($\ddot{\mathbf{x}} \approx 1$) can be closely approximated by

$$\dot{\mathbf{x}} \approx 1 - \frac{1}{2\mathbf{x}^2}\,, \tag{3.6}$$

where here $\mathbf{x}$ is the coordinate distance to the halfway point (about $10^6$ lightyears for the trip to the Andromeda galaxy).

The speed works out to $\dot{\mathbf{x}} \approx 0.9999999999995$. That's twelve nines after the decimal point—closer than 1 part in $10^{12}$ from the speed of light!

All the above calculations are in the realm of science fiction. Why? Because of (amongst other things) the following problem - where do we find a drive for the spaceship that can keep up even the modest acceleration of 1g for tens of years?

The energy required for this sort of mission cannot be carried on board the spaceship, because to get that close to the speed of light, we will have to convert just about every gram of mass of the entire spaceship into energy at 100% efficiency - which is an impossibility.

So the energy must come from some external source. How about tapping radiation coming from the stars that we pass on our way? The problem is that radiation coming from the stars behind the ship will be Doppler-shifted virtually out of existence. The same will happen to any energy that we try to beam to the ship from Earth.

Radiation from the stars ahead of the ship will be Doppler-shifted to white-hot energy, but it comes from the wrong direction. Radiation pressure on the front of the spaceship will be enormous.

There is (perhaps) one 'plausible' energy source though - the energy of the vacuum. If we can borrow some energy from the vacuum during the acceleration phase and give it back to the vacuum during the deceleration phase, we might be in business.

How such a 'vacuum drive' might work, nobody knows. If only scientists and engineers can come up with some plan! "Free energy", unfortunately, seems to be an illusion.*

Back to reality. We have seen how continuous linear acceleration slows down the clock on board a spaceship. Inevitably, the question must come up: is it the acceleration itself that does the trick, or is it the very high relative velocity that results from the continious acceleration?

The answer to this question is not an easy one. Acceleration is an absolute thing - it can be measured on board the spaceship with no reference to the outside world. Velocity is a relative thing that can only be measured with reference to the outside world.

The next section attempts to answer this question.

*The Internet is 'flooded' with free energy schemes, some of them proposing the energy of the vacuum. So far, no credible schemes have surfaced.

## 3.5 Desynchronization revisited

Looking at the space-propertime diagram, e.g., figure 3.4, it appears as if it is the velocity that causes the relative slowing down of the clocks. The acceleration is just the agent that creates the relative velocity.

On the other hand, looking at the approximate equation for propertime, $\tau' \approx ln(2\mathbf{x}\ddot{\mathbf{x}})/\ddot{\mathbf{x}}$, it appears as if the acceleration $\ddot{\mathbf{x}}$ is the decisive factor. But then, speed is the first time derivative of acceleration, so in a way we can use them 'interchangeably'.

By far the simplest way to comprehend the situation is to assume that linear acceleration itself has no effect on the rate of clocks and that it is the resulting relative velocity that causes relative time dilation and a desynchronization of clocks.

The following scenario illustrates the point. Assume that we have a small, incompressible laboratory floating in free space with a master clock on the floor of the laboratory.

Now set up a 'repeater' for the ground clock' against the ceiling, which we slave to the floor clock as follows: the floor clock transmits laser light pulses to the repeater, with a pulse repetition frequency (PRF) derived from the floor clock's stable frequency source.

The repeater uses these pulses to increment counters that drive the repeater display. Provided that, in setting the initial reading of the repeater, we use the normal method of subtracting the light travel time. We are therefore continously synchronizing the repeater with the floor clock.

Now accelerate the laboratory uniformly at a constant acceleration $\ddot{\mathbf{x}}$, in a direction from the floor to the ceiling. We take this as the positive $\mathbf{x}$ direction.

In the time that each light pulse is in transit from the floor to the ceiling, the ceiling picks up some extra speed due to the acceleration. The reciever is therefore moving at a higher velocity than the transmitter. One can expect a Doppler shift of the laser light frequency and also of the PRF.

For a short distance of mild acceleration, we can take the travel time of the pulse as approximately $h$ in geometric units (e.g. metres), where $h$ is the height of the laboratory in metres.

So, during the travel time of each laser pulse, the ceiling has picked up additional speed $\ddot{\mathbf{x}}h$. From the previous chapter, we know that this gives a Doppler shift of received PRF relative to transmitted PRF of

$$\frac{\Delta\lambda}{\lambda} = \ddot{\mathbf{x}}h,$$

which is the fractional rate at which the ceiling repeater will be 'loosing time' against the floor clock.

After a time $\Delta t$ of acceleration, the ceiling repeater will have lost $\Delta t\ \ddot{\mathbf{x}}\ h$ time units. Since $\Delta t\ \ddot{\mathbf{x}} = \dot{\mathbf{x}}$, the speed change of the laboratory in the time $\Delta t$, we can say that the ceiling receiver lost $\dot{\mathbf{x}}h$ units of time, i.e., the product of the speed change and the height of the laboratory.

If we now stop the acceleration, the repeater will start to 'tick' at the same rate as the floor clock again (because there will be no more Doppler shift). The repeater will however be $\dot{\mathbf{x}}h$ units of time behind the floor clock, as viewed by the inertial reference frame in which the laboratory was stationary before the acceleration happened.

In the laboratory frame, however, everybody will insist that the repeater is still synchronized with the floor clock—after all, the whole setup was designed to keep it synchronized!

The analysis used so far was fairly loose, because of the simplification that the light travel time is equal to the height of the laboratory. This will certainly not hold for large acceleration over long distances. Further, the accelerating laboratory is not an inertial frame and the Doppler shift calculation is not strictly valid.

Amazingly enough, when a full relativistic analysis is done, we get the same answer in a rigorous way and it is valid for any velocity change and any laboratory height. Figure 3.6 illustrates this point on a space-propertime diagram.

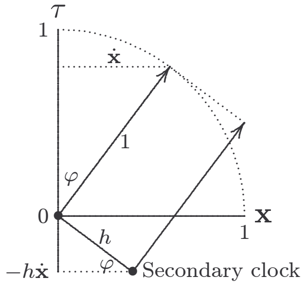


**Figure 3.6:** The unity space-propertime vector $\nu$ represents a clock moving at velocity $\dot{\mathbf{x}}$ relative to coordinate system $\mathbf{x}, \tau$. A secondary (synchronized) clock, stationary relative to the primary, is located at a distance $(h)$ from the origin, as measured in the moving frame. Because $\dot{\mathbf{x}} = \sin\varphi$, the desynchronization offset (secondary - primary) $= -h\sin(\varphi) = -h\dot{\mathbf{x}}$.

Although desynchronization must be viewed as a relative velocity effect, there is a subtle second order effect caused directly by the acceleration.

Relative to the inertial reference frame, the laboratory suffers increasing Lorentz contraction as it's velocity in direction of movement increases. Relative to an inertial frame, the ceiling clock thus always travels at a lower velocity than the floor clock and suffers less time dilation.

The effect is loosely illustrated in figure 3.7. Because acceleration has an absolute nature, this difference can be measured absolutely.

An 'experiment' to test this can be done as follows: put two identical clocks on the floor of the accelerating spaceship and synchronize them. Then move one clock slowly to the ceiling, leave it there for a long time without any attempt to synchronize it to the floor clock. Finally, bring the ceiling clock slowly back to the floor again.

When the two clocks are directly compared, the clock that spent time at the ceiling of the accelerating laboratory will be ahead of the floor clock. This effect is absolute as compared to the effect of desynchronization. The latter is relative in the sense that it can only be measured over a distance, using light or other signals that travel at the speed of light.

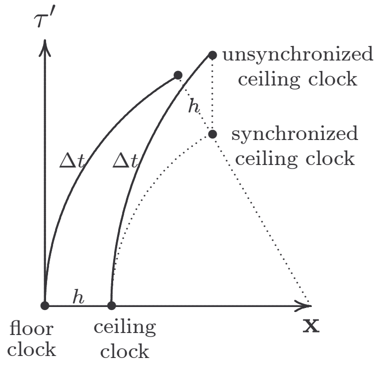


**Figure 3.7:** The $\mathbf{x}\tau'$ paths of the floor and ceiling clocks, separated by height $h$, being accelerated from rest in an inertial frame of reference, where the ceiling clock is not being synchronized to the floor clock. The ceiling clock suffers less acceleration because of accumulating length contraction due to the increasing velocity. However, the curved space-propertime pathlength of both clocks must be the same, i.e. $= \Delta t$. The only solution possible is where the unsynchronized ceiling clock's propertime is ahead of that of the floor clock, as shown. The apparent $\mathbf{x}\tau'$ position of a synchronized ceiling clock is also shown.

This effect is directly related, but not quite equivalent to *gravitational time dilation*, as will be discussed fully in a later chapter.

The described absolute effect of acceleration on the ceiling clock is much smaller than the 'apparent running fast' of the ceiling clock caused by the changing desynchronization. In just about all practical situations it can be ignored as insignificant.

This just about wraps up acceleration in a gravity free environment. In the next chapters, we will look intensively at gravitational acceleration.

## 3.6 Summary of Special Relativity

We have learned that the best way to view special relativity is by means of the space interval and the time interval between events. This removes (and prevents) paradoxical conclusions to be drawn. An observer that is present at two events always measures a time interval that is shorter than the time interval measured by any inertial observer not present at both events.

The Lorentz transformation is the 'tool' for transforming time and space intervals from one inertial reference frame to another. This is possible because the spacetime interval is a constant for any two events. The spacetime interval is a function of the time interval and the space interval.

If the space interval exceeds the distance that light can travel in the corresponding time interval, the spacetime interval is called 'space-like'. If the time interval exceeds the time that light will take to travel the corresponding space interval, the spacetime interval is called 'time-like'. The borderline between the two is called 'light-like' (what else?).

Momentum and energy in relativistic dynamics differ from the same concepts in Newtonian dynamics. In special relativity, the difference is coupled to the velocity time dilation factor. If any material object is moving at a speed approaching that of light in a reference frame, its momentum and energy approaches infinity in that reference frame.

In special relativity, Doppler shift of electromagnetic signals also differs from the equivalent effect in Newtonian dynamics. We have seen that the difference only shows up in one-way Doppler shifts. this is due to the 'absolute rest frame' effects of Newtonian dynamics. In two-way, round trip signals, the rest frame effects cancel out and the two theories mentioned gives the same result.

Lastly, we have examined linear acceleration in this chapter. Acceleration does not affect atomic clocks directly. However, the fact that the speed of the accelerated clock changes relative to whichever inertial reference one chooses, causes an effect on clocks.

We have seen that the one-way distance that could be accomplished in a realistic human spacefarer's lifetime is extraordinary... provided that the traveler's spaceship is linearly accelerated to very close to the speed of light.

If the ship is continously being accelerated, its average speed will be very close to the speed of light in virtually every inertial frame of reference. Hence the 'longevity' of the spacefarer as calculated in an inertial reference frame.

Having had a taste of acceleration in the gravity-free environment, it is now time to move on to general relativity, the realm of gravity and gravitational acceleration. The next seven chapters are devoted to the this very important part of physics—important because we are all experiencing the effect of gravity every moment of our lives... well just about!

## Transformation of acceleration details

Set up an inertial reference frame $(\mathbf{x}, t)$ and another inertial frame $(\mathbf{x}', \tau')$, moving at a speed $\dot{\mathbf{x}}$ relative to the reference frame. Now let an observer inside the moving frame accelerate a test object with a constant acceleration $\ddot{\mathbf{x}}'$, for a time $\Delta\tau'$, starting from rest in the moving frame. If the time interval is small, the test object would have acquired a speed $\Delta\dot{\mathbf{x}}' = \ddot{\mathbf{x}}'\Delta\tau'$ relative to the moving frame. How will the reference frame $\mathbf{x}, t$ measure this acceleration?
Firstly, the additional speed $(\Delta\dot{\mathbf{x}})$, *as measured by the reference frame*, can be obtained from the law for relativistic summation of velocities, as

$$\Delta\dot{\mathbf{x}} = \frac{\dot{\mathbf{x}} + \Delta\dot{\mathbf{x}}'}{1 + \dot{\mathbf{x}}\Delta\dot{\mathbf{x}}'} - \dot{\mathbf{x}}.$$

Secondly, to find the acceleration measured by the reference frame, we also need to transform the time interval $\Delta\tau'$ to the rest frame, using the time dilation factor, obtaining

$$\Delta t = \frac{\Delta\tau'}{\sqrt{1 - \dot{\mathbf{x}}^2}}.$$

Now we can find the acceleration measured by the reference frame through $\ddot{\mathbf{x}} = \Delta\dot{\mathbf{x}}/\Delta t$. After a bit of algebraic juggling, we obtain the acceleration relative to the reference frame as

$$\ddot{\mathbf{x}} = \frac{(1 - \dot{\mathbf{x}}^2)^{\frac{3}{2}}}{1 + \dot{\mathbf{x}}\Delta\dot{\mathbf{x}}'}\ \ddot{\mathbf{x}}'.$$

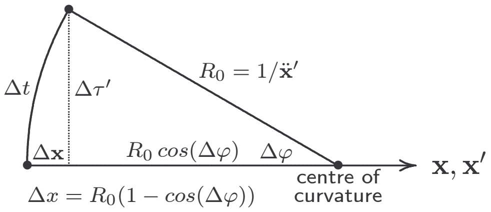
In the limit, where the time interval is so short that the change in velocity becomes negligible $(\Delta\dot{\mathbf{x}}' \to 0)$, the denominator term approaches 1 and the transformation equation approaches

$$\ddot{\mathbf{x}} = (1 - \dot{\mathbf{x}}^2)^{\frac{3}{2}}\ \ddot{\mathbf{x}}'.$$

So the acceleration transformation can be viewed as three 'velocity time dilation' transformations, i.e., $(\sqrt{1 - \dot{\mathbf{x}}^2}\,)^3$.

### About acceleration and radius of curvature

The first figure below shows a circular arc $\Delta t$, defined by radius of curvature $R_0$, rotated through angle $\Delta\varphi$ off the $\mathbf{x}$-axis. In the limit $\Delta t, \Delta\varphi \rightarrow 0$, the results are as shown below. It is the same as the Newtonian acceleration for an object moving at the speed of light around a circle with radius $R_0$.


$\Delta t$
$\Delta\tau'$
$R_0 = 1/\ddot{\mathbf{x}}'$
$\Delta\mathbf{x}$
$R_0\, cos(\Delta\varphi)$
$\Delta\varphi$
$\mathbf{x}, \mathbf{x}'$
centre of curvature

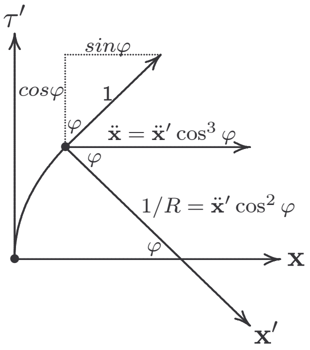
$\Delta x = R_0(1 - cos(\Delta\varphi))$

$\Delta t = R_0\ \Delta\varphi$

If $\Delta\varphi \rightarrow 0$, then $[1 - cos(\Delta\varphi)] \rightarrow (\Delta\varphi)^2/2$, and

$$\mathbf{x} = \frac{e^{\ddot{\mathbf{x}}'\tau'} + e^{-\ddot{\mathbf{x}}'\tau'} - 2}{2\ddot{\mathbf{x}}'}$$

The above is valid for acceleration from a position of rest relative to the reference frame. Once the object has picked up some speed relative to the reference frame, the velocity vector makes an angle $\varphi$ with the $\tau'$-axis and the line to the centre of curvature makes the same angle with the $\mathbf{x}$-axis. We know that $\ddot{\mathbf{x}} = \ddot{\mathbf{x}}'(\sqrt{1 - \dot{\mathbf{x}}^2})^3 = \ddot{\mathbf{x}}' \cos^3\varphi$, so the radius of curvature at angle $\varphi$ must be $R = 1/(\ddot{\mathbf{x}}' \cos^3\varphi) \times \cos\varphi = 1/(\ddot{\mathbf{x}}' \cos^2\varphi)$, shown below in inverse (acceleration) form.


$\tau'$
$sin\varphi$
$cos\varphi$
1
$\varphi$
$\ddot{\mathbf{x}} = \ddot{\mathbf{x}}' \cos^3\varphi$
$\varphi$
$1/R = \ddot{\mathbf{x}}' \cos^2\varphi$
$\varphi$
$\mathbf{x}$
$\mathbf{x}'$

### Algorithm for programming a relativistic acceleration curve by using radius of curvature $R$

Understanding of the algorithm requires some programming experience, as it is written in a form of pseudo code. Text starting with “ is a comment and not part of the algorithm.

```
Start Algorithm
  Initialize the following variables as double precision floating point:
    $acc = value\ of\ choice$     “acceleration $\ddot{\mathbf{x}}'$
    $dt = value\ of\ choice$      “time interval $\Delta t$
    $R = 1/acc$                   “initial radius of curvature
    $phi = 0$                     “initial velocity vector rotation angle $\varphi$
    $dphi = dt/R$                 “incremental rotation angle $\Delta\varphi$
    $\mathbf{x} = 0$              “initial space displacement
    $tau = 0$                     “initial propertime displacement $\tau'$
    $d\mathbf{x} = 0$             “initial space increment $\Delta\mathbf{x}$
    $dtau = 0$                    “initial propertime increment $\Delta\tau'$
    $d\mathbf{x}a = 0$            “intermediate value $\Delta\mathbf{x}$
    $dy = 0$                      “intermediate value $\Delta y$
  Repeat from here
    Output $\mathbf{x}$ and $tau$ to text or graphics device
    $d\mathbf{x} = R \times (1 - COS(dphi))$                           “$\Delta\mathbf{x}$ after rotation
    $dy = R \times SIN(dphi)$                                          “$\Delta y$ after rotation
    $d\mathbf{x}a = d\mathbf{x} \times COS(phi) + dy \times SIN(phi)$  “$\Delta\mathbf{x}'$ after rotation
    $dtau = dy \times COS(phi) - d\mathbf{x} \times SIN(phi)$          “$\Delta\tau'$ after rotation
    $\mathbf{x} = \mathbf{x} + d\mathbf{x}a$                           “new $\mathbf{x}$
    $tau = tau + dtau$                                                 “new $\tau$
    $phi = phi + dphi$                                                 “new $\varphi$
    $R = 1/(acc \times (COS(phi))^2)$                                  “new $R$
    $dphi = dt/R$                                                      “new $\Delta\varphi$
  Repeat until terminating condition is satisfied
End Algorithm
```

With a suitably small value for $\Delta t$, this algorithm can be used to calculate the space-propertime curve for an accelerating object to any accuracy you want. Further, a new acceleration can be calculated for every cycle of the programmed loop to simulate a rocket with changing acceleration as it's fuel burns up.

# Chapter 4

# Static gravity as geometry

about
geodesics in
space-propertime

When relativists talk about 'gravity as geometry', they usually refer to *differential geometry*, a branch of mathematics dealing with curves on curved surfaces (although the 'curve' can be a straight line and the 'curved surface' can be flat!). It is a rather difficult subject: Misner, Thorne and Wheeler's benchmark book [MTW] devotes eight chapters, in fact the whole part III of almost 90 pages to it. The title page of Part III of [MTW] says it all:

> **THE MATHEMATICS OF
> CURVED SPACETIME**
>
> *Wherein the reader is exposed to the charms of a new temptress—
> Modern Differential Geometry—and makes a decision:
> to embrace her for eight full chapters; or,
> having drunk his fill, to escape after one.*

The 'escape' is possible due to [MTW]'s brilliant scheme of having a 'track 1' and a 'track 2' path through the quite formidable book. Since the text you are reading is aimed mainly at engineers, it avoids the 'new temptress' as best it can and rather 'tempt' the reader with more 'engineering-like' language.

Geometric gravity is all about geodesics in spacetime. Geodesics are all about taking the shortest possible path through spacetime. This text exploits the concept of *space-propertime* that we have used in chapter 3. Recall that it is about 'your time' plotted against 'my space', where I am stationary in the inertial reference frame and you are moving relative to me.

It is showed that geodesics are relatively simple paths in space-propertime, at least if the situation is restricted to simple scenarios. The simplest scenario in relativistic gravity occurs in Schwarzschild geometry. It is about gravity outside of an isolated spherically symmetrical mass that is also uncharged, non-rotating and permanently at rest at the origin of the coordinate system.

The resulting geodesics can be translated into relatively simple accelerations of test objects or 'test particles'. The term 'test particle' is used to emphasize the fact that the mass of the test object is negligible compared to the mass of the gravitational source.

As discussed in chapter 3, linear acceleration in gravity-free space can be viewed as a space-propertime rotation about a point we called the centre of curvature. It is reasonable to assume that gravitational acceleration can also be viewed as space-propertime rotation about some centre of curvature. A space-propertime geodesic is just the path created by such a rotation, at least when viewed for a very short time.

In relativity, the geodesic centre of curvature lies somewhere in four dimensional spacetime and the path must generally be thought of as movement through space and time. We will first explore the elements of the geodesic for test objects that are, at least momentarily, stationary in the gravitational field.

By 'stationary' we mean that the test object has no *spatial* movement relative to the source of gravity. In four dimensional spacetime, 'stationary' still means movement in the time dimension. The space-propertime paths of moving and stationary particles have the same length, if measured over a short time and they are more or less in the same location in space.

We will later take it step by step to the more general situation where the object is not spatially stationary, but rather moving in some arbitrary spatial direction in the spherical symmetrical gravitational field.

## 4.1 Newtonian acceleration from rest

'From rest' in this subtitle means a test object that is free-falling, but is momentarily stationary relative to a gravitational field. This is like when a ball that you throw vertically upwards, reaches it's highest point relative to the centre of Earth, where it is momentarily stationary relative to Earth's centre.

If you could throw the ball in a vacuum, and accurately measure it's position against time for a reasonable period around it's highest point, you can in principle obtain the centre of geodesic curvature ($cc$) and the radius of geodesic curvature ($R$) for the ball at it's highest point, at least approximately (see figure 4.1).

The geodesic centre lies (space wise) in a direction through the centre of Earth and at a distance of just under one lightyear, or roughly $10^{13}$ km. Since we are working with 'acceleration from rest', we ignore the Earth's

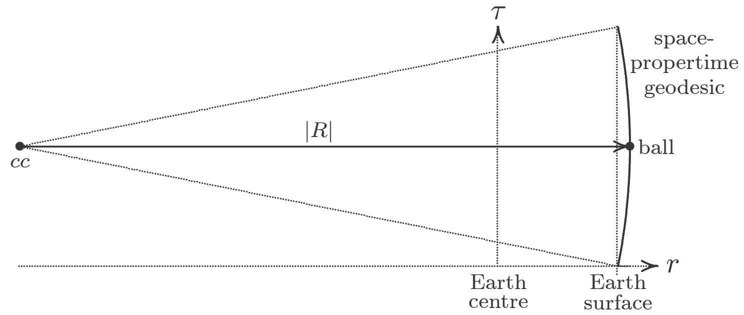


**Figure 4.1:** A highly exaggerated space-propertime diagram of the ball thrown spatially upwards (here 'upwards' means along the horizontal space axis), showing the geodesic as a space-propertime rotation around the centre of curvature ($cc$).

rotation here.

Due to the relatively small gravitational field on Earth's surface, the radius of geodesic curvature is obtained, quite simply, as the inverse of the Newtonian gravitational acceleration in geometric units. Divide 1g $= -9.8$ m/s$^2$ by $c^2$ to get about $-10^{-16}$ m$^{-1}$, and invert to get $|R| \approx 10^{16}$ metres, a distance very close to one lightyear.

In Newton's gravitation, the acceleration is quantitatively equal to the negative of the slope of the Newton potential function $\Phi = -GM/r$, i.e.,

$$a_r = -\frac{d}{dr}\left(\frac{-GM}{rc^2}\right) = -\frac{GM}{r^2c^2}. \tag{4.1}$$

We can check our previous calculation by noting that $GM/c^2$ for Earth is about $4.5 \times 10^{-3}$m and the radius of Earth is about $6.3 \times 10^6$m, giving

$$a_r \approx -\frac{4.5 \times 10^{-3}}{6.3^2 \times 10^{12}} \approx -10^{-16} \text{ m}^{-1},$$

in agreement with the rough $c^2/g$ calculation above.

## 4.2 Quasi-Newtonian acceleration

Newton's theory of gravity is not a 'geometric spacetime theory'. It can however be partially 'geometrized' by adding a little of Einstein's theory of relativity to it.

Newton defined the potential energy of a test object with mass $m$, at radial distance $r$ from a central mass $M$ as

$$E_p = -GmM/r. \tag{4.2}$$

Far from mass $M$, when $r \to \infty$, $E_p \to 0$.

Einstein convinced us that the total energy of a test object must include it's rest energy by his $E = mc^2$. If we add this rest energy of the test object to the Newtonian potential energy, we have a 'quasi-Newtonian' potential energy

$$E_p = mc^2 - GmM/r = mc^2(1 - GM/(rc^2)). \tag{4.3}$$

So the potential* of mass $M$ is simply $\Phi = 1 - M/(rc^2)$.

*Potential is the function by which the rest energy of a test object must be multiplied to get potential energy.

The slope of $\Phi$ can be used to find the radius of curvature ($|R|$) for the free fall of the test object from rest, as is shown in figure 4.2. The gravitational centre of curvature $cc$ is where the tangent line to the curve intersects the $r$ axis (i.e., the line $\Phi = 0$).

The radius of curvature is no longer the simple inverse of the slope of the curve, but is shorter by the factor $1 - GM/(rc^2)$. This implies that the gravitational acceleration is no longer $a = -GM/r^2$, but rather

$$a' = \frac{-GM}{r^2(1 - GM/(rc^2))}, \tag{4.4}$$

which has a larger modulus than the classical Newtonian acceleration. We will later see that this is a good approximation for relativistic gravity in all but the most extreme conditions.

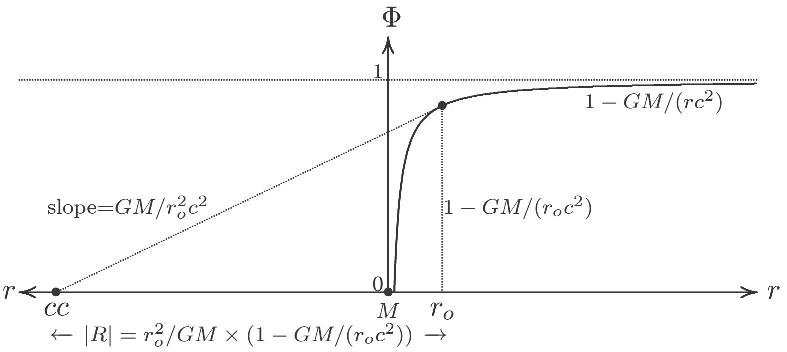


**Figure 4.2:** Curve of the quasi-Newtonian potential $(1 - GM/(rc^2))$ and the construction to find the radius of curvature ($|R|$) from the slope of the potential curve.

The potential at any point is proportional to the 'rate of time' at that point. Taking this statement at face value for now, we can exchange the $\Phi$-axis with a time axis ($t$) and plot the curves for a number of different times, as is shown in figure 4.3.

The curves represent 'equally spaced' consecutive times read off a stationary clock as a function of the distance from mass $\bar{M}$. To put it more formally,

$$\Delta t' = (1 - GM/(r_oc^2))\Delta t, \tag{4.5}$$

where $\Delta t'$ is the time increment read off the local clock and $\Delta t$ the corresponding time increment read off a distant clock, where the gravitational field has no effect.

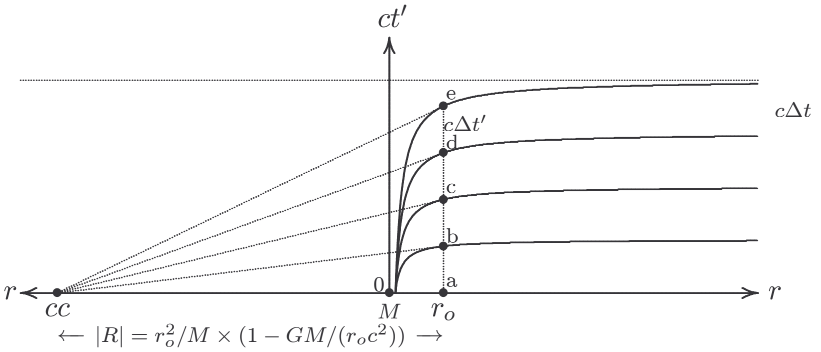

$ct'$
e
d
c
b
a
$c\Delta t'$
$c\Delta t$
0
$r$
$r$
$cc$
$M$
$r_o$
$\leftarrow |R| = r_o^2/M \times (1 - GM/(r_o c^2)) \rightarrow$

**Figure 4.3:** 'Quasi-Newtonian time dilation', where each of the points $a$, $b$, $c$, $d$ and $e$ represents consecutive time readings on a local stationary clock, situated at radial distance $r_o$ form $M$. It is clear from the curves that the local clock records less elapsed time than a distant clock, far to the right. The tangents to the curves at distance $r_o$ from mass $M$ all meet at the centre of curvature $cc$.

When we draw a tangent to each of the curves at distance $r_o$, it is fairly obvious that they will all intersect at the same point, $cc$ in the figure. This point is the centre of curvature for the space-propertime movement of a test object starting to free-fall from a state of rest towards mass $M$.

This is true in this 'geometrized' Newtonian gravity and approximately true for all but very extreme gravitational fields. The inverse of the radius of curvature thus gives a geometrized gravitational acceleration.

Figure 4.4 shows a section of the space-propertime geodesic of the test object as it 'rotates' around the centre of geodesic curvature. The arclength $c\Delta t'$ has been highly exaggerated for clarity, because the curvature centre $cc$ will move as soon as $r_o$ decreases.

The radius of curvature shown is only valid when $c\Delta t' \to 0$, but this is exactly what we are after: *the acceleration from rest at distance* $r_o$.

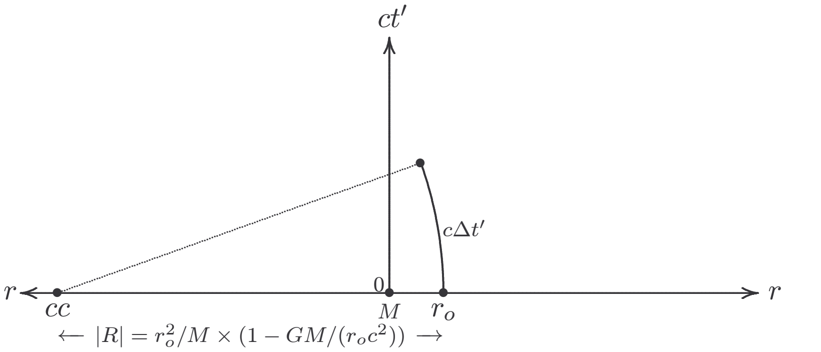

$ct'$
$c\Delta t'$
0
$r$
$r$
$cc$
$M$
$r_o$
$\leftarrow |R| = r_o^2/M \times (1 - GM/(r_o c^2)) \rightarrow$

**Figure 4.4:** A 'quasi-Newtonian geodesic' for an object free-falling from rest at distance $r_o$ from $M$, during an exaggerated time interval $\Delta t'$. To be valid, $\Delta t'$ should approach zero, because $cc$ will shift as soon as the object moves closer to $M$.

The normal Newtonian acceleration is an approximation of the 'geometrized' Newtonian acceleration when the field is very weak.*

*A very interesting observation is that the 'geometrized' Newtonian acceleration will also reduce to the normal Newtonian acceleration if the speed of light was infinite ($c \to \infty$).

When $r$ becomes very much larger than $GM/c^2$, then $GM/(rc^2) \to 0$ and the factor $(1 - GM/(rc^2)) \to 1$. This reduces the quasi-Newtonian acceleration to the classical Newtonian gravitational acceleration $-GM/r^2$.

## 4.3 Relativistic gravitational acceleration

Since we have seen that

$$\Delta t' = (1 - GM/(r_o c^2))\Delta t,$$

where $\Delta t'$ is local time and $\Delta t$ the corresponding coordinate time, we can guess a timelike spacetime metric for our 'quasi-Newtonian' scenario as follows:

$$dt'^2 = \left(1 - \frac{GM}{rc^2}\right)^2 dt^2 - \frac{dr^2}{c^2} - \frac{r^2 d\psi^2}{c^2} \tag{4.6}$$

where $dt$ is coordinate time differential and $dt'$ a corresponding time differential in the local observer's frame. Recall that $dr$ is a radial displacement and $r d\psi$ a transverse displacement in the distant observer's frame.* This metric indicates 'time dilation' and 'flat space', since only the time-time coefficient ($g_{tt}$) deviates from the flat spacetime of the Lorentz metric.

*The terms $r^2 d\phi^2 + r^2 \sin^2 d\theta^2$ of Chapter 1 are combined into $r^2 d\psi^2$ for economy.

Further, since we are working with acceleration of a test object that is momentarily stationary in the gravitational field, it implies that $dr = d\psi = 0$, or at least that they are of negligible magnitude. The metric then simplifies to

$$dt'^2 = \left(1 - \frac{GM}{rc^2}\right)^2 dt^2. \tag{4.7}$$

The equivalent Schwarzschild timelike metric (from chapter 1) is

$$d\tau^2 = \left(1 - \frac{2GM}{rc^2}\right) dt^2. \tag{4.8}$$

Carefully note the subtle difference between the last two equations. Since $(1 - 2GM/(rc^2)) \approx (1 - GM/(rc^2))^2$ when $r \gg GM/c^2$, 'relativistic gravity' approximates to 'quasi-Newtonian gravity' in a weak field.*

*This is at least true for a momentarily stationary test object in a weak field.

It is obvious from the last equation that for the relativistic case

$$d\tau/dt = \pm\left(1 - \frac{2GM}{rc^2}\right)^{\frac{1}{2}}. \tag{4.9}$$

The negative root has no physical meaning, but the positive root is the 'gravitational time dilation factor' or the 'gravitational redshift factor' of general relativity. It tells us how much slower stationary clocks near a gravitational source 'tick' than clocks far away from any such source.

This also means that physical processes near a gravitational source emit radiation at a lower frequency than the same processes far from such a source. Hence the term 'gravitational redshift'.

For weak fields, $(1-2GM/(rc^2))^{\frac{1}{2}} \approx 1 - GM/(rc^2)$, the quasi-Newtonian potential. Does this mean that we can get the relativistic acceleration from rest by taking the slope of the curve $d\tau = (1-2GM/(rc^2))^{\frac{1}{2}}dt$? Almost, but not quite.

The quasi-Newtonian treatment allowed time dilation, but only 'flat space'. General relativity sports both time dilation and 'curved space'. If we compare the timelike metric of our quasi-Newtonian spacetime with the Schwarzschild metric, this becomes clear:

$$dt'^2 \approx g_{tt}dt^2 - \frac{dr^2}{c^2} - \frac{r^2 d\psi^2}{c^2} \quad \text{(quasi-Newtonian)}, \tag{4.10}$$

and:

$$d\tau^2 = g_{tt}dt^2 - g_{rr}\frac{dr^2}{c^2} - \frac{r^2 d\psi^2}{c^2} \quad \text{(Schwarzschild)}. \tag{4.11}$$

Note that, unlike the quasi-Newtonian metric, the spacial part in the radial direction $(dr^2/c^2)$ of the Schwarzschild metric is modified by $g_{rr}$. This is space curvature at work.

The Schwarzschild metric can be visualized (more or less) as in figure 4.5, were a fictitious 'hyperspace' dimension is introduced, forming a space-hyperspace plane together with the space-propertime plane.

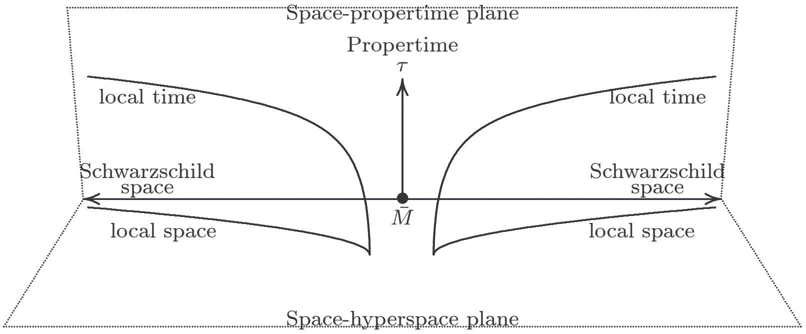


**Figure 4.5:** Curved spacetime, where the local time curve has curvature into both the propertime and the hyperspace directions. This spacetime curve intersects the space-hyperspace plane at a radial distance $r = 2GM/c^2$, the event horizon of a mass singularity (black hole).

The idea is to show that the local time curve is not only curved in the propertime direction, but also into the hyperspace direction; hence we have 'curved spacetime'. The local time curve must satisfy $d\tau/dt = \sqrt{g_{tt}} = [1-2GM/(rc^2)]^{0.5}$. Differentiating this function against $r$, we get the spacetime slope in the propertime direction:

$$\frac{d\tau/dt}{dr} = \sqrt{g_{rr}}\frac{GM}{r^2c^2}. \tag{4.12}$$

The local space curve is conformal to the space-hyperspace plane and needs to be constructed so that a Schwarzschild radial distance differential $dr$ corresponds to a local space distance differential of $dr/(1 - 2GM/(rc^2))^{0.5}$, thus making the curve conform to the Schwarzschild metric. Simple Pythagoras then gives the slope of spacetime in the hyperspace direction as:

$$\frac{d_h}{dr} = \left(g_{rr}\frac{2GM}{rc^2}\right)^{0.5}, \tag{4.13}$$

where $d_h$ refers to a 'hyperspace distance' differential. Using the two slopes above, the tangent to the spacetime curve intersects the space-hyperspace plane at point $cc$, as illustrated in figure 4.6.

When you have one side of a right triangle and the slopes in two directions, it is a relatively straightforward geometry exercise to obtain the distance between $r_o$ and $cc$ as

$$|R_o| = \frac{\sqrt{g_{tt}}\, r_o^2c^2}{GM}, \tag{4.14}$$

the (local) radius of geodesic curvature for an object starting to free-fall from rest at distance $r_o$ from $M$.

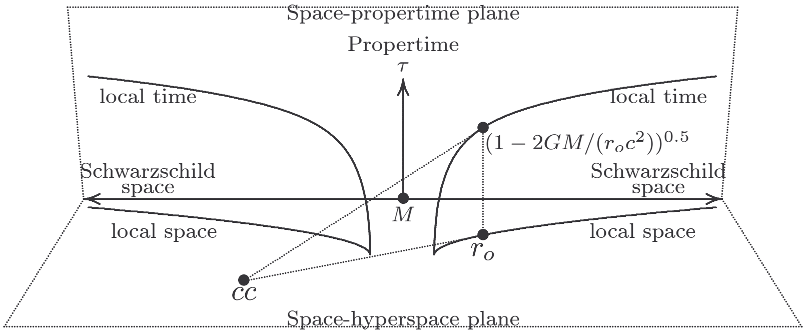


**Figure 4.6:** The centre of curvature $cc$ is where the tangent to the local time and local space curves intersect. The radius of curvature in the coordinate system is the projection of interval $r_o \rightarrow cc$ onto the Schwarzschild space axis.

Like before, the radial gravitational acceleration of a test object momentarily at rest at distance $r_o$, as measured by a local observer, is obtained by taking the square of the speed $c$, divided by the radius $|R_o|$, giving

$$a_{r(local)} = -\sqrt{g_{rr}}\,\frac{GM}{r_o^2}. \tag{4.15}$$

This is larger in magnitude than the quasi-Newtonian acceleration that we have dealt with before (recall that $g_{rr} = 1/g_{tt} >= 1$). It is considerably larger than standard Newtonian acceleration for strong gravitational fields.

In fact, at the event horizon ($r = 2GM/c^2$), where $g_{rr} \rightarrow \infty$, the local stationary acceleration diverges to infinity. This tells us that no material object can be locally stationary at the event horizon. It can only be falling inwards.

This is how an observer that is on the spot (a local observer) will measure the acceleration of an object, free-falling from rest. An observer that is a large distance away from the source of gravity (a distant observer), hypothetically measuring the same object starting to free-fall at distance $r_o$ from the mass, will not obtain the same acceleration.

Due to the spacetime curvature between the two observers, the distant stationary observer measures an acceleration that is three factors of *gravitational time dilation* smaller than what the local observer measures, as will be shown below.

Since the gravitational time dilation (or redshift) factor is $\sqrt{g_{tt}}$, the measured static acceleration of the local observer must be multiplied by $g_{tt}^{1.5}$, giving

$$a_{r(distant)} = -g_{tt}\,\frac{GM}{r_o^2}. \tag{4.16}$$

Amazingly, this makes the magnitude of the relativistic acceleration measured by the distant observer smaller (in magnitude) than the $-GM/r_o^2$ of classic Newton mechanics, at least for momentarily stationary test objects.

At the event horizon, where $r_o = 2GM/c^2$ and $g_{tt} = 0$, the acceleration of the object as measured by the distant observer becomes zero. We will later see that, according to a distant observer, the test object's movement seems to 'freeze' there.

It is again interesting to note that if $c$ was infinite, then both local and distant accelerations would have reduced to the Newtonian acceleration.

Further, with infinite light speed, the event horizon would shrink to zero radius. This is where Newton's gravitational acceleration "blows up" to infinity, so in a way, Newton's theory is an infinite light speed theory of gravity—stated in different terms, it is an "action at a distance" theory. More about that later.

The conversion factor for acceleration between the local and the distant observers is best understood when we consider how gravitational acceleration will be measured by the respective observers.

A simple way of understanding the gravitational acceleration conversion factor is to obtain the time ($dt$) it takes an object that starts to free-fall from a state of rest in a gravitational field, over an infinitesimal distance ($dr$).

The acceleration is then easily calculated as

$$a = 2\frac{dr}{dt^2},$$

from the well known expression for acceleration from rest, $s = 0.5at^2$, where $s$ is the distance traveled, $a$ the constant acceleration and $t$ the travel time.

This helps to make the 'three factors of gravitational time dilation' more palatable. Firstly, the distant observer's observation of the distance traveled is

$$dr_{distant} = (g_{tt})^{0.5}\ dr_{local},$$

due to space curvature, as we have seen before (from the Schwarzschild metric).

Secondly, the distant observer's observation of the elapsed time can be expressed as

$$dt^2_{distant} = \frac{dt^2_{local}}{g_{tt}},$$

due to the gravitational time dilation—the distant observer's clock runs faster then the local observer's, hence the longer time interval measured.

Since distant acceleration is proportional to $dr_{distant}/dt^2_{distant}$, the conversion factor is

$$a_{distant} = (g_{tt})^{1.5}\ a_{local}. \tag{4.17}$$

## 4.4 Opposing acceleration components

There is a very interesting way of looking at the gravitational acceleration measured by a distant observer. Let an observer at a large distance (approaching infinity) from an isolated mass $M$, measure the radial acceleration of an object initially stationary at distance $r$ from the mass.

It would appear as if there is a component of relativistic acceleration working against the expected Newtonian acceleration. Write the radial acceleration as (we write $g_{tt}$ out to illustrate a point)

$$a_r = -\left(1 - \frac{2GM}{rc^2}\right)\frac{GM}{r^2} = -\frac{GM}{r^2} + \frac{2(GM)^2}{r^3c^2}. \tag{4.18}$$

It appears as if there is an 'opposing acceleration' of

$$a_{r(opp)} = \frac{2(GM)^2}{r^3c^2} \tag{4.19}$$

working against the expected Newtonian radial acceleration $-GM/r^2$.

This opposing acceleration is tiny when $GM/r \ll c^2$, but precisely cancels out the Newtonian acceleration when $r = 2GM/c^2$, i.e., at the Schwarzschild radius or event horizon of a static black hole.

As an example of how tiny the opposing component is in normal circumstances, let us view Earth as sitting in empty space. Now let a distant observer (hypothetically) measure the static radial acceleration at Earth's surface.

The mass of Earth is roughly $6 \times 10^{24}$ kg and the radius roughly $6.3 \times 10^6$ metres. With $G = 6.67 \times 10^{-11}$ m$^3$/(kg s$^2$) and $c = 3 \times 10^8$ m/s, the opposing acceleration works out to be in the order of 1 nano-g.

Another way of putting it is that the static relativistic acceleration at Earth's surface, as hypothetically measured from afar, is about one nano-g less than the static gravitational acceleration predicted by Newton's theory.

For us, the local observers on Earth, the static acceleration is about half a nano-g *stronger* than what Newton would have predicted. Can you recall why? Look at the local static acceleration equations again.

In the next chapter, we will show that there are a few other 'opposing' and also 'additive' acceleration components making their appearance in relativity. This happens when the movement of the test object is taken into account.

## 4.5 Summary of static gravity

It was shown that normal Newtonian gravity can be pictured as a rotation around a centre of curvature in spacetime. The radius of curvature for Earth is about one light-year. When an object is rotated around this radius with a velocity equal to the speed of light, the centripetal acceleration is about 1g.

This lead to the concept of 'quasi-Newtonian' gravity, where $e = mc^2$ were added to the potential energy of an object stationary in a gravitational field. A 'spacetime metric' for Newtonian gravity was then 'guessed' and an equation for quasi-Newtonian acceleration in a gravitational field derived. It was shown that in low fields, the quasi-Newtonian field reduces to standard Newtonian gravity.

Similarly, the quasi-Newtonian gravity is a weak field approximation of relativistic gravity. The latter involves both time dilation and 'curved space', i.e. curved spacetime. This causes the same acceleration to be measured differently by a local and a distant observer.

A local observer measures a gravitational acceleration that is of higher magnitude than what Newton's theory predicts. A distant observer, however, measures the same acceleration to be of smaller magnitude than what Newton would have predicted.

Finally, it was shown that for a distant observer, the difference between a Newtonian and a relativistic calculation boils down to an 'opposing acceleration' component in general relativity. This is caused by the curvature of spacetime.

We have now treated two special cases: a) where there is relative velocity, but no gravity—special relativity, and b) where there is gravity, but no velocity—not quite general relativity. When the two is combined, we have case c), general relativity. How this combination of a) and b) is done is the subject of the next chapter.

# Chapter 5

# Gravity and Moving Objects

fancy
geodesics in
space-propertime

Having studied 'static' acceleration to some detail, the step towards acceleration experienced by moving objects seems to be relatively easy. However, we will soon see that space curvature complicates things considerably. Fortunately the results are relatively simple in the case of Schwarzschild coordinates. We will consider pure radial movement in this coordinate system first.

## 5.1 Gravity and radial movement

As has been seen many times before, radial movement means that the space-propertime path is not perpendicular to the space axis. It makes an angle

$$\varphi = \arcsin(v/c)$$

with the *time axis*, where $v$ is the locally measured radial velocity.

In the static case of the previous chapter, the centre of curvature for the geodesic was situated on the space-hyperspace plane.indexspace-hyperspace plane This is not so for the case with a radial starting velocity. The line from the origin to the centre of curvature now makes the same angle as above, i.e.

$$\varphi = \arcsin(v/c)$$

with the *space axis*.

As shown in figure 5.1, the original (static) geodesic centre of curvature ($cc$) is 'raised' perpendicular to the space-hyperspace plane, to a new position $cc'$. This action increases the radius of curvature of the geodesic and thereby decreases the radial acceleration.

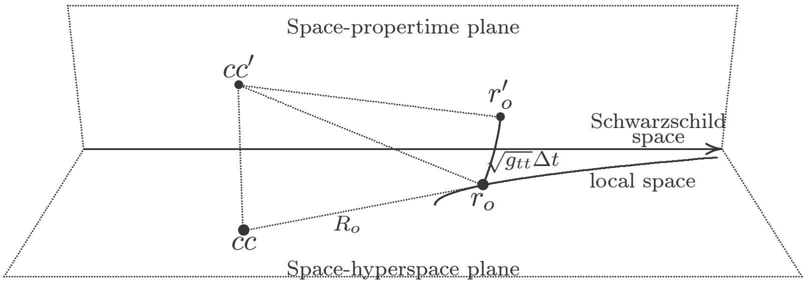


**Figure 5.1:** Geodesic movement through space-hyperspace-propertime for an object with a positive radial velocity. The centre of geodesic curvature ($gc$) is 'raised' normal to the space-hyperspace plane to the point $gc'$, to give a larger effective radius of curvature. This reduces the effective radial acceleration.

The locally measured radial acceleration is shown in appendix C to be

$$a = -g_{rr}\frac{GM}{r^2}(1 - \frac{v_\varphi^2}{c^2}), \tag{5.1}$$

where $v_\varphi$ is the locally measured radial velocity.

This is a quite understandable combination of the static acceleration $(-g_{rr}\frac{GM}{r^2})$ of the previous chapter and the (square of) the velocity time dilation of special relativity, $\sqrt{1 - v^2/c^2}$. It is clear that the local radial velocity diminishes the static acceleration.

This is easy to comprehend, in the sense that when an object is accelerated to the local speed of light, it cannot be accelerated any further. Hence radial acceleration becomes zero.

Due to space curvature, the coordinate radial acceleration requires a pretty tricky transformation of the local acceleration. This is shown in appendix C to result in the following expression:

$$a_r = -\frac{GM}{r^2}(g_{tt} - 3g_{rr}\frac{v_r^2}{c^2}), \tag{5.2}$$

where $v_r$ is the coordinate radial velocity. Note that $-GM/r^2$ is the Newton gravitational acceleration, which is reduced in magnitude firstly by $g_{tt}$ (which is less than unity) and secondly by an effective positive acceleration component

$$\frac{GM}{r^2} \times 3g_{rr}\frac{v_r^2}{c^2}$$

that is dependant upon space curvature $g_{rr}$ and radial velocity $v_r$.

It is somewhat like when the radial speed of an aircraft is measured by a Doppler-radar, while the aircraft is flying at a constant ground speed and

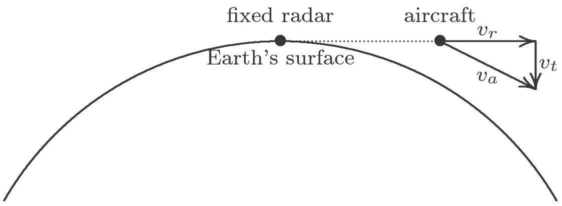


**Figure 5.2:** Doppler-radar measurement of the radial speed of an aircraft, flying at constant altitude and ground speed, measured near the horizon. If the true ground speed of the aircraft is represented by vector $v_a$, the radar will measure a radial speed of $v_r = \sqrt{v_a^2 - v_t^2}$. For illustrative purposes, the refraction of the radar signal through the atmosphere is ignored.

altitude, in a direction directly away from the radar. Due to the curvature of Earth, the measured radial velocity will decrease (see figure 5.2).

So judged purely by the measured radial speed, the aircraft seems to be decelerating. In the case of our distant observer, the same type of effect is operating due to space curvature. The effect is discussed in detail in appendix C.

In order to get a feeling for the magnitude of this opposing acceleration, let us consider a meteorite hitting the Earth's surface radially at escape velocity, about 11 km/s (ignoring the atmosphere).

For Earth, we can simply work with $-GM/r^2 = 1\text{g}$ and express our results in g's. The speed of 11 km/s translates to $v_r \approx 3.7 \times 10^{-5}\ c$. This gives an opposing acceleration

$$a_{r(opp)} \approx -1 \times 3 \times (3.7 \times 10^{-5})^2 \approx -4 \times 10^{-9}\ \text{g}.$$

The negative sign means 'upwards', since 'g' works downwards. This acceleration of -4 nano-g is four times the opposing acceleration caused by the gravitational redshift alone (the static relativistic effect).

However, if that 'meteorite' is a muon particle hitting the Earth's surface at $v_r = 0.99c$, the opposing acceleration works out to

$$a_{r(opp)} \approx -1 \times 3 \times 0.99^2 \approx -3\text{g}.$$

So even ignoring atmospheric drag and other effects, the muon will be *decelerating* at 1 - 3g = -2g just before it hits the surface. Yes, -2g is correct! This is some serious braking, especially for a particle that is being 'pulled' by Earth's gravity.

But remember that it is the acceleration that is (hypothetically) measured in the coordinate system, i.e. by a distant observer. Due to the space curvature between the muon and the observer, the apparent opposing acceleration simply 'overwhelms' the normal Newtonian acceleration.

Engineers are trained to question "funny" results. So how do we get some confidence in the "funny" result of the calculation in question? The answer lurks in the Schwarzschild coordinate velocity of light. In a purely radial direction, the speed of light is

$$c_{rad} = (1 - \frac{2GM}{rc^2})\ c. \tag{5.3}$$

This means it is slower near a massive body than far from it. If, using this equation, the path of a photon is modeled in a radial direction and the acceleration extracted from the data, the result is an apparent coordinate acceleration of

$$a = 2GM/r^2$$

near the surface of Earth. Contrast this with the normal Newtonian acceleration of

$$a = -GM/r^2 = 1\text{g}.$$

So for Earth's surface, radial light seems to accelerate at -2g in the coordinate system. There is a -3g acceleration opposing the Newtonian value, as measured in the coordinate system.

The muon, moving at very close to the speed of light will suffer roughly the same effect. Otherwise it might eventually have exceeded the speed of light.

One must however remember that for the local observer, momentarily stationary as the test object passes, the observed radial acceleration will be quite different:

$$a_{loc} = \frac{-GM}{r^2}\sqrt{g_{rr}}\left(1 - \frac{v_{loc}^2}{c^2}\right), \tag{5.4}$$

where $v_{loc}$ is locally measured radial velocity.

The factor $\sqrt{g_{rr}}$ (space curvature) enhances local radial acceleration and the radial velocity factor diminishes local radial acceleration. The local radial acceleration can never become positive, because $v_{loc} < c$. The local acceleration approaches zero if the local radial velocity approaches the speed of light.

For the case of the muon hitting Earth's surface, we are approximately local observers* and if we assume that there is no space curvature between us and the muon's point of impact, we will measure the muon's acceleration as

$$a_{loc} \approx 1 \times (1 - 0.99^2) \approx 0.02\ \text{g}.$$

This means that we will observe the muon (atmospheric drag and other effects ignored) to accelerate at only 0.02g (downwards) when it hits the surface. Muons travel at so close to the speed of light that Earth's gravity can hardly accelerate them any more.

Figure 5.3 shows a plot of the gravitational acceleration measured by a distant and a local observer respectively, for an object falling to the event horizon of black hole of a million solar masses, starting from rest at a large

*Strictly, we should be free-falling and momentarily stationary.

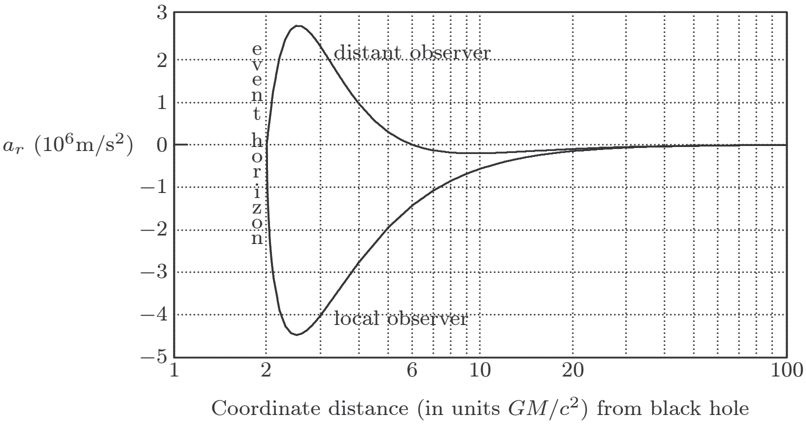


**Figure 5.3:** Radial acceleration of an object free-falling at (minus) escape velocity towards a black hole of a million solar masses, as measured by a distant observer (top curve) and the transformation of that acceleration to the reference frame of a local observer (bottom curve).

distance and moving towards the hole at very close to (negative) escape velocity.

The acceleration in the distant observer's frame peaks (negatively) at $a_r \approx -2 \times 10^5$ m/s$^2$ around $r = 9.5GM/c^2$, becomes zero at $r = 6GM/c^2$ and reaches a maximum positive value of $a_r \approx 2.7$ million m$/s^2$ around $r = 2.5GM/c^2$. Near the event horizon the observed acceleration tends to zero, due to both $g_{tt}$ and $v_r$ tending to zero.*

*Radial escape velocity in the distant frame is given by $v^2_{r(esc)} = g^2_{tt} \times 2GM/r$.

To the locally stationary observer however, the acceleration stays negative and peaks at $a_r \approx -4.4$ million m$/s^2$ around $r = 2.5GM/c^2$. When approaching the event horizon, the local acceleration also tends to zero as the local velocity approaches the speed of light.

Precisely at the event horizon, the equation for local radial acceleration diverges due to $g_{rr} \to 1/0$. However, $(1 - v^2/c^2) \to 0$ at a faster rate, so the acceleration tends to zero.

## 5.2 Gravity and transverse movement

In order to picture transverse movement in the space-hyperspace-propertime domain, one needs at least 4 dimensions in your diagram, because you need 2 normal space dimensions—one space dimension for the transverse movement and one space dimension for gravity to work in.

Since 4 dimensional drawings are not possible, it is usual to leave out either the propertime dimension or the hyperspace dimension. Fortunately, for visualizing purely transverse geodesics, the hyperspace dimension is not all that important, because the test object remains at a constant distance from the primary mass and experiences constant space curvature.*

*The space curvature is only precisely constant if the primary mass is spherically symmetrical and non-rotating.

The circular orbit is not the spacetime geodesic; one must view the time component of the geodesic as well. The circular orbit of a massive object transcribes a helix in spacetime. Figure 5.4 pictures a small segment of such a geodesic of a circular orbit.

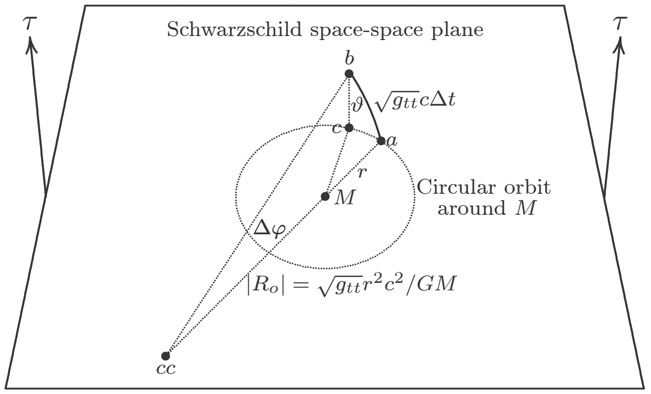

$\tau$
$\tau$
Schwarzschild space-space plane
$b$
$\vartheta$
$\sqrt{g_{tt}}c\Delta t$
$c$
$a$
$r$
$M$
Circular orbit around $M$
$\Delta\varphi$
$|R_o| = \sqrt{g_{tt}}r^2c^2/GM$
$cc$

**Figure 5.4:** Geodesic movement through space-propertime for an object in a circular orbit around a mass $M$. The circle segment *cc-a-b* is tilted by an angle $\vartheta = \arcsin(v_o/c)$ relative to the normal to the space-space plane, where $v_o$ is the locally measured circular orbital velocity. The angular movement of the geodesic is $\Delta\varphi = \sqrt{g_{tt}}c\Delta t/|R_o|$. The projection $c$ (of point $b$) onto the space-space plane is on the circular orbit.

It is not a trivial exercise to rigorously find relativistic gravity's centripetal acceleration as measured relative to a straight line in Euclidean space, so the details will be discussed in Appendix C. Here it will only be loosely discussed as the centrifugal acceleration measured by a local observer.

First, we look at the Newtonian centrifugal acceleration $v_o^2/r$, where $v_o$ is the circular orbital velocity. This is the velocity where the centrifugal acceleration precisely balances the Newtonian gravitational acceleration $-GM/r^2$, or

$$\frac{v_o^2}{r} - \frac{GM}{r^2} = 0. \tag{5.5}$$

It can be shown that the locally stationary observer will, for the same circular orbit, obtain a relativistic balance of a accelerations

$$\frac{g_{tt}v_o^2}{r} - \frac{GM}{r^2} = 0, \tag{5.6}$$

where $v_o$ is the circular orbit velocity as measured by the local observer. After a bit of algebraic juggling,* this can be shown to be equivalent to

*Replace $g_{tt}$ with $1 - 2GM/(rc^2)$, multiply out and collect $GM/r^2$ terms on the right.

$$\frac{v_o^2}{r} = \frac{GM}{r^2}(1 + 2v_o^2/c^2), \tag{5.7}$$

which, after being transformed to the frame of the distant observer gives the relativistic centripetal acceleration for a circular orbit as

$$a_{cp} = \frac{-GM}{r^2}(g_{tt} + 2v_t^2/c^2), \tag{5.8}$$

where $v_t^2 = g_{tt}v_o^2$ is the transverse (orbital) velocity as measured by a distant observer.

The above treatment is only valid for circular orbits, but the result happens to be valid for any value of purely transverse velocity, as is proven in Appendix C, where the centripetal acceleration at the periapsis or apoapsis of an elliptical orbit is derived.

An interesting and important observation from the last equation is the following: since $g_{tt} = 1 - 2GM/(rc^2)$, it is clear that for $v_t^2 > GM/r$, $a_{cp}$ is larger than standard Newtonian acceleration $(-GM/r^2)$. For $v_t^2 < GM/r$, $a_{cp}$ is smaller than Newtonian acceleration.

At $v_t^2 = GM/r$ the opposing and additive components cancel each other, representing a circular orbit, and the centripetal acceleration is Newtonian $(-GM/r^2)$.

This tells us that a perfectly circular orbit 'looks' Newtonian when measured by a distant observer. Relativistic effects do not show up in perfectly circular orbits measured form afar (i.e. in Schwarzschild coordinates).

Non-circular orbits, on the other hand, suffer increased centripetal acceleration when they are inside the equivalent circular orbit radius (i.e., near the perihelion of a planet), where the transverse velocity is greater than the circular orbital velocity.

The opposite happens when the orbit goes outside of the equivalent circular orbital radius. These are contributing factors to the perihelion shift of a planet with a fairly eccentric orbit, like Mercury, as will be discussed in the next chapter.

Looking at the centripetal acceleration equation, one can say that we have an opposing acceleration due to the gravitational redshift alone, of

$$a_{redsh} = \frac{2(GM)^2}{r^3c^2}. \tag{5.9}$$

For Earth's surface, we know this from a previous calculation as about -1 nano-g .

Then there is an 'additive' acceleration of

$$a_{add} = \frac{-2GM}{r^2}\frac{v_t^2}{c^2}, \tag{5.10}$$

due to the transverse velocity alone, working with the Newtonian acceleration.

If we have a high speed particle, say the muon again, this time momentarily moving purely transverse (i.e., horizontally) close to Earth's surface at $v_t =$

$0.99c$, we can calculate the additive acceleration as follows: since $-GM/r^2$ equals 1g, the additive acceleration must be:

$$a_{add} = 2 \times 0.99^2 \text{ g} \approx 2 \text{ g},$$

again ignoring atmospheric drag and other effects.

This means that the muon will appear to be accelerated towards Earth's centre at about 3g (the normal +1g plus the 2g additive). This approximate value is equally valid for local and distant observers. As an exercise, the reader may perhaps contemplate on why this is so.*

*Hint: Purely transverse velocity does not involve spacetime curvature.

## 5.3 Radial and transverse velocities combined

In the previous sections, we have treated purely radial and purely transverse movement only. So what happens if both radial and transverse movements are present? The geodesic movement will then cause a non-circular orbit of some sorts.

It is possible, by the same type of arguments that we used for purely transverse or purely radial movement, to calculate the spacetime geodesic in four dimensions. The projection of the geodesic onto a flat 2-dimensional space surface then gives the relativistic orbit.

It is however quite a messy operation, involving tricky tensor algebra. A more 'engineering-like' method is to try and work out the equivalent 'Newtonian' acceleration by combining the two accelerations that we obtained before (for radial and transverse movements respectively).

This is a relatively simple operation. The accelerations that we have worked out so far all works in the radial direction and we can add them up in a Newtonian manner. The resultant radial acceleration, as measured in the distant (coordinate) frame, is then

$$a_r = \frac{-GM}{r^2}\left(g_{tt} - 3g_{rr}\frac{v_r^2}{c^2} + 2\frac{v_t^2}{c^2}\right), \tag{5.11}$$

where $v_r$ and $v_t$ are respectively the radial and transverse velocity components measured in the distant frame.

So can we go and calculate the trajectory that a particle will follow in space by using this radial acceleration in a Newtonian manner? Not quite, because we are not done yet.

Rather surprisingly, there are additional transverse acceleration components in relativistic gravity. It should not be surprising, though, because Newton's trajectories in gravitational field undergo transverse accelerations in most cases.

**An aside...** When the author reached this point in developing the quasi-Newtonian interpretation, he promptly plugged the radial acceleration equation into a program for calculating and plotting orbits. In strong gravitational fields the orbits were way off the mark, when compared to standard relativistic solutions (discussed in the next chapter). So back to the drawing board...

Think about the ball rolling down a slope. The gravitational force works vertically downwards, yet the ball picks up speed down the slope. That acceleration contains both radial and transverse components.

Both these components are modified in curved spacetime. Hence, there is at least one transverse component to factor into a quasi-Newtonian calculation. It will become clear later that this must be an opposing acceleration in the transverse direction.

To return to the ball down the slope example: the ball will pick up transverse speed slower than expected. If the ball is rolled up the slope, it will lose less speed than expected. In the low speed, weak gravity case, the effect will be negligible, but not so at relativistic velocities.

The magnitude of the apparent opposing transverse acceleration can be found in a similar way as for purely radial movement. The apparent opposing radial acceleration was previously found to be

$$a_{r(opp)} = \frac{3GM}{r^2} g_{rr} \frac{v_r^2}{c^2}.$$

As discussed in appendix C, the transverse component has the same form, but with a factor 2 instead of 3 and with $v_r^2$ replaced by the product $v_r v_t$, i.e.,

$$a_{t(opp)} = \frac{2GM}{r^2} g_{rr} \frac{v_r v_t}{c^2}. \tag{5.12}$$

If we are close to a black hole and the product $v_r v_t \neq 0$, then $g_{rr}$ plays a significant role. In the low gravity of Earth, where $g_{rr} \approx 1$, significant velocities tend to dominate the action.

As an example, let our muon particle again momentarily move at a speed of about 0.99$c$, but this time at an angle of 45 degrees off the radial towards Earth, so that $v_r \approx -0.7c$ and $v_t \approx 0.7c$. In the absence of atmospheric drag and other effects, the opposing transverse acceleration will be

$$a_{t(opp)} \approx -2 \times -0.7 \times 0.7 \approx 1\text{g}.$$

The starting negative sign is because g is a negative (-9.8 m/s$^2$) acceleration.

This is a very significant acceleration in the negative transverse direction, i.e., against the direction of transverse movement. For comparison, let us also calculate the total radial acceleration of the muon for this scenario, i.e.,

$$a_r \approx [1 - 3 \times (-0.7)^2 + 2 \times 0.7^2] \approx 0.6\text{g}.$$

In Newtonian dynamics, $a_{t(opp)} = 0$ and $a_r = 1\text{g}$, so one can expect that there will be a significant difference between the relativistic track and the Newtonian track of the muon.

The relativistic radial acceleration is smaller than the Newtonian value, but the opposing transverse acceleration will cause the muon to curve a little more strongly towards the centre of Earth than what Newton would have predicted.

> To understand why, think about what would happen if the muon's transverse velocity approaches zero (which is unlikely, but it illustrates the point).

So much for the rather weak gravitational field of planet Earth. Next, we will work through an example in a strong gravitational field. Let a distant observer measure the accelerations of our muon, instantaneously at one Earth radius from a black hole with a mass of a hundred million ($10^8$) Earths.

This gives $g_{tt} = 0.861$ and the Newtonian radial acceleration will by definition be $10^8\text{g}$.

Assume instantaneous (local) velocity components of $v_r = 0.7c$ and $v_t = 0.7c$ (both positive in this case). The velocities transform to distant (coordinate) velocity components

$$v_r = 0.7c \times 0.861 \approx 0.6c$$

and

$$v_t = 0.7c \times \sqrt{0.861} \approx 0.65c.$$

> If the coordinate velocity components were higher by as much as 0.1%, the local velocity would have exceeded the speed of light, which is not allowed.

The radial acceleration comes out as

$$a_r \approx 10^8 \left( 0.861 - 3\frac{(0.6)^2}{0.861} + 2 \times 0.65^2 \right) \approx 4.4 \times 10^7\text{g}.$$

The opposing transverse acceleration works out to be

$$a_{t(opp)} \approx -2 \times 10^8 \, \frac{0.6 \times 0.65}{0.861} \approx -9.1 \times 10^7\text{g}.$$

The negative result means it is a positive acceleration, in-line with the transverse movement (again because we work with g = -9.8 m/s$^2$).

One can expect that the lower radial acceleration, coupled with a significant transverse acceleration will result in a trajectory dramatically different from a Newtonian trajectory. Just how dramatic will become clear in the next chapter, on orbits.

The typical engineer's question to the above is: how can a purely radial gravitational 'force' produce an acceleration that is normal to the force? It was loosely answered before. Appendix C attempts to answer this question in a rigorous way. The following is another 'loose' view of the effect.

Newton trajectories also cause (non-zero) transverse speed to either increase or decrease, depending whether the trajectory is in-falling or out-falling. The conservation of energy demands it. In the curved spacetime world of general relativity, energy conservation works slightly differently and this causes additional accelerations.

A second common question is: if an observer could ride with the muon, would the opposing accelerations be perceived as 'opposing forces'? The answer is no; in the absence of other external forces, the muon (and hypothetical observer) follows a spacetime geodesic and would not experience any forces.

> An observer might perhaps experience tidal gravity forces, which will be a squeeze and a stretch. More about that in a later chapter.

Now for a crucial question: are these opposing and additive accelerations real, or are they just artefacts of measurement? The simplest, yet useless answer is that they are a bit of both. They are real in the sense that they can be measured—the paths of fast moving objects in a gravitational field is different from Newtonian predictions.

On the other hand, the relativistic accelerations differ, depending on who is making the measurements. The local and distant observers get different answers when they measure the acceleration of the same object moving in the same gravitational field.

This is because their measurement rods are of different length (or of different orientation in curved space) and their clocks tick at different rates. All these are most notable when strong gravitational fields are involved.

In the next chapter it will be shown that the relativistic gravitational accelerations can be used in a quasi-Newtonian way to construct orbits. Such orbits are virtually indistinguishable from the orbits produced by the more formal relativistic orbital equations.

## 5.4 Newton's gravity and the speed of light

Although Newton surely never contemplated that the speed of light is infinite, it is interesting to note that if we take relativistic gravity and set $c \rightarrow \infty$, it reduces to Newton's gravity. First, look at the metric coefficients $g_{tt}$ and $g_{rr}$

$$g_{tt} = 1 - \frac{2GM}{rc^2}.$$

It reduces to unity when $c \rightarrow \infty$, and since $g_{rr} = 1/g_{tt}$, so does $g_{rr}$. Taken only this into account, the gravitational accelerations reduce to:

$$a_r \quad \rightarrow \quad -\frac{GM}{r^2}\,(1 - 3\frac{v_r^2}{c^2} + 2\frac{v_t^2}{c^2})$$

$$a_t \rightarrow 2\frac{GM}{r^2}\frac{v_r v_t}{c^2}.$$

It is fairly obvious that all the speed ratios $(v/c)$ approach zero when $c \rightarrow \infty$, (for all speeds less than infinite, that is) so that what we have left is the Newtonian acceleration

$$\begin{aligned} a_r &\rightarrow -\frac{GM}{r^2} \\ a_t &\rightarrow 0. \end{aligned}$$

It is also interesting to note that the Schwarzschild radius $r_S = 2GM/c^2$ reduces to zero when $c \rightarrow \infty$, so no black holes, only (naked)singularities at the centre. Since the acceleration would be Newtonian, one can expect that all orbits would be Newtonian.

All the above is roughly the same thing as leaving the value of $c$ at its normal value and stating that Newton's gravity holds in the weak field $(GM \ll rc^2)$, low speed $(v \ll c)$ domain.

If Einstein was right, it seems that should $c \rightarrow \infty$, the weak field, low velocity domain rules everywhere. But then, as far as we know, there was never a time when the speed of light approached infinity.

## 5.5 Summary of relativistic acceleration

In this chapter, the static relativistic acceleration has been developed further into firstly, acceleration with pure radial movement, secondly, acceleration with pure transverse movement and then radial acceleration with both radial and transverse movements.

It was shown that radial movement decreases radial acceleration in both the local reference frame and the coordinate reference frame. For the local frame, it is just a velocity time dilation type effect. For the coordinate frame, the velocity time dilation is enhanced by a space curvature effect.

The most interesting effect of this chapter is likely to be the additional (and conditional*) component of acceleration in the transverse direction. We have seen how high velocities with both radial and transverse components influence the acceleration that particles experience. This is most severe in strong gravity fields, but it is even significant in the weak field of Earth.

> *Recall that the opposing transverse acceleration is a function of the product of radial and transverse velocity. If either is zero, it vanishes.

Finally, the fact that if the speed of light was infinite, relativistic gravity would reduce to Newton's gravity was loosely discussed. This could make one wonder if the 'speed of gravity' is not infinite. We will discuss this at the end of the next chapter (orbital dynamics).

# Chapter 6

# Orbital dynamics - an Introduction

From Newtonian
to
'Einsteinian' orbits

The orbit is an extremely important aspect of gravitational theory, because just about every object that is observable from Earth is in some form of orbit relative to something. The proverbial 'Newton apple' that falls from the tree is attempting to go into orbit around the centre of the Earth. It is however prevented from doing so because the surface of the Earth intervenes.

Any object that is in free-fall in friction-free space obeys two fundamental laws of physics: the conservation of total energy and of total angular momentum. In Newton mechanics, total energy is the sum of potential energy and the kinetic energy of movement and the two types of energy can be exchanged, as long as the sum remains constant.

Angular momentum is the sum of any rotational momentum that the object may have and the angular momentum relative to the origin of whatever coordinate system is used for the measurement. In the case of an apple hanging from a tree, it has angular momentum proportional to the product of it's mass, the speed at which the rotation of the Earth carries it along (relative to the centre of the Earth) and it's distance from the centre of the Earth.

As the apple falls from the tree, it's distance from the centre of the Earth decreases and so does it's potential energy. To compensate, the apple must pick up more speed so that an increase in kinetic energy can make up for the loss of potential energy—this is high school stuff.

Not so well known is the fact that the transverse speed of the apple must also increase, in order to maintain angular momentum as the distance from the centre of the Earth decreases. Surprisingly, this means that the apple does not fall precisely vertically, but will hit the ground a minute distance to the east of the point vertically below it's starting point!*

*Everywhere except at the poles of course!

If the Earth could be reduced to a point mass, the apple would have gone into an elliptical orbit around that point mass—at least if we ignore atmospheric effects. That is how Newton would have argued, based upon his theory of gravity.

We will now use the two conserved quantities to analyse orbits around an isolated, spherically symmetrical and homogeneous massive object, which simplify things considerably. To lay a foundation, we will first consider orbits in Newton mechanics and then show the essential elements of relativistic orbits.

Since many parameters of Newtonian orbits differ from the more general parameters of general relativity, the subscripted $N$ will be used to indicate Newtonian parameters, e.g., $E_N$ indicates Newtonian energy. The unsubscripted parameters with the same name will indicate general relativistic parameters.

## 6.1 Newtonian orbits

In Newton dynamics, the total orbital energy of a small mass $m$ (the 'test object'), orbiting at a (variable) distance $r$ from the centre of a large mass $GM$ is simply the kinetic energy plus the (negative) potential energy:

$$E_N = \tfrac{1}{2}m(v_r^2 + v_t^2) - GmM/r,$$

where $v_r$ and $v_t$ are the radial and transverse components of orbital velocity, as we used them before. In general, $r$, $v_r$ and $v_t$ will be constantly changing, but $E_N$ will remain constant.

The second conserved quantity, the angular momentum of the orbit, is given by

$$L_N = m\ r\ v_t,$$

dependant only on the transverse speed component of the orbit (and of course the distance and the mass of the test object).

Engineers will probably more readily recognize the angular momentum as $L = I\ \omega = m\ r^2\ \omega$, where $I$ is the moment of inertia and $\omega$ the angular speed. It is the same thing as $L_N = m\ r\ v_t$.

We will later show that this is the equivalent of Kepler's second law, stating that the orbit sweeps out equal areas in equal intervals of time. If the total energy and the angular momentum of a Newton orbit are known, the orbit can be solved.

In orbital analysis it is convenient to work as "normalized as possible". The trick is to get rid of the mass of the orbiting body ($m$) and work with the total energy parameter (total energy per unit orbiting mass) of the orbiting object, $\tilde{E}_N = E_N/m$, and likewise with the angular momentum parameter, $\tilde{L}_N = L_N/m$, i.e.

$$\tilde{E}_N = \tfrac{1}{2}(v_r^2 + v_t^2) - GM/r \quad \text{and} \quad \tilde{L}_N = r\ v_t, \tag{6.1}$$

where $\tilde{E}_N$ has the dimensions m$^2$/s$^2$ and $\tilde{L}_N$ has the units of m$^2$/s.

Some textbooks, e.g., [MTW], sometimes normalizes the angular momentum parameter further: $L^\dagger = \tilde{L}/M = L/(mM)$, i.e., the angular momentum per unit orbiting mass per unit primary mass. While it makes many orbital equation a bit more compact, it is somewhat 'ugly' looking and some of the intuitive meaning of the equations are lost.

Too much normalization may not always be a good thing, so we will continue to use $\tilde{L}$. It does mean that the fraction $\tilde{L}/M$ occurs frequently in the equations that follow.

Another very useful orbit parameter is the *effective potential*, which is the sum of the *transverse* kinetic energy plus the potential energy, both per unit orbiting mass, i.e.

$$\tilde{V}_N = \tfrac{1}{2}v_t^2 - GM/r,$$

with the obvious units of speed squared, m$^2$/s$^2$. By substituting $v_t = \tilde{L}_N/r$, we can write the equation as

$$\tilde{V}_N = \tfrac{1}{2}\tilde{L}_N^2/r^2 - GM/r, \tag{6.2}$$

i.e., a function of the constant $\tilde{L}_N$ and the variable radial distance $r$.

The physical meaning of $\tilde{V}_N$ is the total energy (parameter) at the 'turning points' of the orbit radius $r$, i.e., the periapsis and apoapsis. At these points, the radial component of the orbit velocity ($v_r$) is zero and the total energy is made up only of transverse kinetic energy and potential energy.

By plotting $\tilde{V}_N$ for a specific $\tilde{L}_N$, the turning points (if any) for a given total orbital energy $\tilde{E}$ can be easily found, as illustrated in figure 6.1.

The curve can be understood by noting that at values of $r \gg \tilde{L}_N$, the term $\frac{1}{2}\tilde{L}_N^2/r^2$ becomes negligible and the term $-GM/r$ dominates. At $r = \tilde{L}_N^2/GM$, the curve reaches a minimum and for smaller distances the term $\frac{1}{2}\tilde{L}_N^2/r^2$ starts to become dominant.

The distance $r = \tilde{L}_N^2/GM$ is also where circular orbits will occur (point $\mathbf{A}$ in the figure). Circular orbits represent the minimum total energy for a given angular momentum $\tilde{L}_N$. If the total energy is increased (with constant $\tilde{L}_N$), the orbit becomes elliptical and has a minimum radius at $\mathbf{B}$, with a maximum radius at $\mathbf{B}'$.

A total energy of precisely zero represents escape energy and the orbit will become open (a parabola). All positive total energy levels represent open hyperbolic orbits, at least in Newtonian dynamics.

Another way to look at the effective potential for an elliptical orbit is as follows. For any point along the orbit, the (constant) total orbital energy

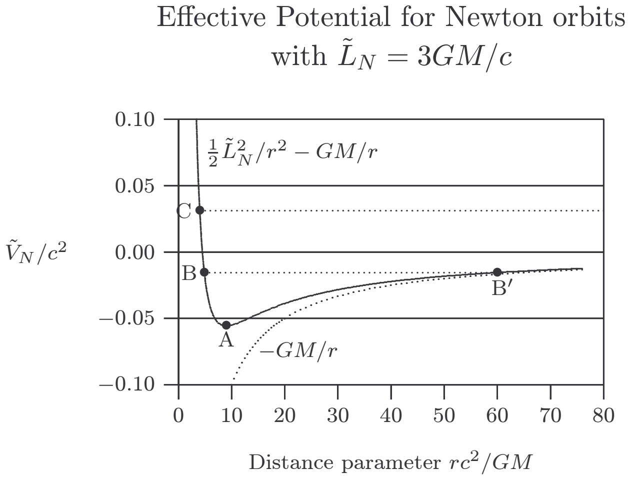


**Figure 6.1:** The effective potential curve (top, solid) and the potential energy curve (bottom, dotted), drawn in geometric (dimensionless) units. Point **A** represents a circular orbit. Point **B** is the closest approach of a closed, elliptical orbit and **B**′ the furthest point, at the same energy level value as point **B**. Point **C** is at positive total energy, which represents an open orbit that will let the orbiting object escape.

is made up of three components: the negative potential energy (dotted curve), the transverse kinetic energy (difference between the sold curve and the dotted curve) and the radial kinetic energy (difference between line **B** - **B**′ and the solid curve).

## 6.2 Kepler's laws for planetary motion

Johannes Kepler established his three laws of planetary orbits in the early 1600s. His first and second laws (elliptic orbits and 'sweeping equal areas in equal time') was established in 1609 and his third law (the correlation between the planetary period and the semi-major axes of the ellipse), followed about ten years later.

**Kepler's first law** states that a planetary orbit is an ellipse, with the Sun at one of the two foci. The eccentricity ($e$) of the orbit is related to the orbital constants $\tilde{E}_N$, $\tilde{L}_N$ and the mass $M$ by

$$e = \sqrt{1 + 2\tilde{E}_N \frac{\tilde{L}_N^2}{(GM)^2}}. \tag{6.3}$$

The eccentricity $e$ is less than unity for negative $\tilde{E}_N$, representing closed (circular or elliptical) orbits. When $\tilde{E}_N = 0$, $e = 1$, representing a parabola. For positive $\tilde{E}_N$, $e > 1$ and the orbit is hyperbolic. The orbital equation for (Keplerian) planetary orbits is

$$\frac{r}{GM} = \frac{\tilde{L}_N^2}{(GM)^2} \frac{1}{1 + e\cos\phi}, \tag{6.4}$$

where $r$ is the radial distance from the Sun and $\phi$ the angular displacement from the point of closest approach along the planar orbit (the perihelion).

By setting $\phi = 0$ and $\phi = \pi$ respectively, the two turning points of the elliptical orbit (the perihelion $r_p$ and aphelion $r_a$) can easily be found (see figure 6.2).

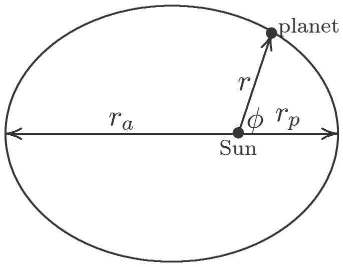


**Figure 6.2:** An elliptical orbit, showing the relationship between $r$, $r_p$, $r_a$ and $\phi$. In the case of planetary orbits around the Sun, the perihelion distance $r_p$ occurs when $\phi = 0$ and the aphelion distance $r_a$ when $\phi = \pi$.

**Kepler's second law** states that a planetary orbit sweeps out equal areas around the Sun in equal time intervals. Let the orbital angle change by $d\phi$, forming an approximate triangle with a 'base' of $rd\phi$ and a 'height' of $r$. Since $d\phi \to 0$, the area of the triangle (half base times height) is

$$dA = \frac{1}{2}r^2 d\phi = \frac{1}{2}\tilde{L}_N dt.$$

Because $\tilde{L}_N$ is a constant by definition,

$$dA/dt = \frac{1}{2}\tilde{L}_N$$

is also a constant, in agreement with Kepler's second law.

**Kepler's third law** states that the orbital period of a planet is proportional to the three-halves power of the semi-major axis of the orbit. This can be shown by utilizing Kepler's second law.

The total area of an ellipse is $A = \pi a^2\sqrt{1-e^2}$, where $a$ is the semi-major axis. Divide this area $A$ by Kepler's constant rate $dA/dt = \frac{1}{2}\tilde{L}_N$ and we have the planet's period as

$$T = \frac{2\pi a^2\sqrt{1-e^2}}{\tilde{L}_N}.$$

It can be shown (e.g., [Faber]) that $\sqrt{1-e^2} = \tilde{L}_N/\sqrt{GMa}$, so that

$$T = \frac{2\pi a^{3/2}}{\sqrt{GM}}, \tag{6.5}$$

which agrees with Kepler's third law.

This relationship is often used to find the mass of the primary object, where a secondary body (planet or a moon) with a known elliptical orbit and a known period is observed. $M$ is easily extracted in terms of the known parameters. It is just as easy to find the semi-major axis $a$ from a known primary mass and a known period.

If the ellipticity is also known, the perihelion and aphelion are found from

$$r_p = a(1-e), \quad \text{and likewise,} \quad r_a = a(1+e).$$

The value of $\tilde{L}_N$ can then be obtained from the orbital equation (6.4), by making $\phi = 0$, i.e.,

$$\tilde{L}_N = \sqrt{r_p(1+e)GM}. \tag{6.6}$$

Then, knowing $\tilde{L}_N$, the effective potential $\tilde{V}_N$ is obtained from equation 6.2, which equals the total energy $\tilde{E}$ for that point (the perihelion).

This method was used to calculate some of the orbital parameters for planet Mercury, with only the period $T$ and the eccentricity $e$ known (together with the mass of the Sun, of course). The results are shown in table 6.1. To get the absolute values of the orbital energy and angular momentum,

| **Data for planet Mercury** | | | | |
|---|---|---|---|---|
| Parameter | geometric | | SI | |
| **Inputs** | | | | |
| mass of Sun $M$ | $1.4765 \times 10^3$ | m | $1.9891 \times 10^{30}$ | kg |
| mass of Mercury $m$ | $2.4518 \times 10^{-4}$ | m | $3.3030 \times 10^{23}$ | kg |
| period $T$ | $2.2786 \times 10^{15}$ | m | $7.6005 \times 10^{6}$ | s |
| eccentricity $e$ | 0.2056 | | 0.2056 | |
| **Calculated** | | | | |
| semi-major axis $a$ | $5.7907 \times 10^{10}$ | m | $5.7907 \times 10^{10}$ | m |
| perihelion distance $r_p$ | $4.6002 \times 10^{10}$ | m | $4.6002 \times 10^{10}$ | m |
| aphelion distance $r_a$ | $6.9813 \times 10^{10}$ | m | $6.9813 \times 10^{10}$ | m |
| angular momentum $L_N$ | $2.2186 \times 10^3$ | m$^2$ | $8.9605 \times 10^{38}$ | kg m$^2$/s |
| angular mom. $\tilde{L}_N = L_N/m$ | $9.0491 \times 10^6$ | m | $2.7128 \times 10^{15}$ | m$^2$/s |
| total energy $E_N$ | $-3.1257 \times 10^{-12}$ | m | $-3.7846 \times 10^{32}$ | kg m$^2$/s$^2$ |
| total energy $\tilde{E}_N = E_N/m$ | $-1.2749 \times 10^{-08}$ | | $-1.1458 \times 10^{9}$ | m$^2$/s$^2$ |
| velocity at perihelion $v_p$ | $1.9671 \times 10^{-4}$ | | 58,973 | m/s |
| velocity at aphelion $v_a$ | $1.2962 \times 10^{-4}$ | | 38,859 | m/s |

**Table 6.1:** Some parameters for the planet Mercury, as calculated from the formulae given in the text. The input values were taken from [Mitton], using constants $G = 6.6714 \times 10^{-11}$ m$^3$kg$^{-1}$s$^{-2}$ and $c = 2.9979 \times 10^8$ m/s.

($E_N = \tilde{E}_N \times m$ and $L_N = \tilde{L}_N \times m$), the mass of Mercury is also needed.

Armed with this information (on what make Newtonian orbits 'tick'), it is time to enter the more formidable arena of relativistic orbits.

## 6.3 Relativistic orbits

The orbits of general relativity can be solved via the Schwarzschild metric and Einstein's equations for geodesics in spacetime. The actual geodesic equations fall outside the scope of this book, because one needs MTW's 'temptress', differential geometry! The results distilled out of the geodesic equations for Schwarzschild geometry is however quite simple.

Like in the case of Newtonian orbits, the total energy and total angular momentum are also conserved quantities, but they are quite different from the values in Newton orbits. We will first examine these differences and then turn to the orbital equations.

### 6.3.1 Conservation of energy and angular momentum

The usual way to find the values of the conserved quantities is to extract the rates of change of the coordinates $t$ and $\phi$ in terms of the proper time ($\tau$) from Einstein's spacetime geodesic equations (e.g., [MTW] eqs. 25.17 and 25.18). Only the results will be given here. Firstly

$$c^2\frac{dt}{d\tau} = \frac{\tilde{E}}{g_{tt}},$$

where $\tau$ represents proper time, $\tilde{E}$ is the relativistic total energy parameter and $g_{tt} = 1 - 2GM/(rc^2)$, the time-time coefficient of the Schwarzscild metric, giving

$$\tilde{E} = c^2 g_{tt}\frac{dt}{d\tau} = \frac{c^2 g_{tt}}{\sqrt{g_{tt} - g_{rr}v_r^2/c^2 - v_t^2/c^2}}. \tag{6.7}$$

The expansion of $dt/d\tau$ comes from the Schwarzschild metric, $v_r$, $v_t$ are the radial and transverse velocity components and $g_{rr} = 1/g_{tt}$, as before.

It can be shown that when $v_r, v_t \ll c$ and $r \gg 2GM/(rc^2)$, the expression approximates to the Newton energy parameter, plus the constant $c^2$, i.e. $\tilde{E} \approx \frac{1}{2}(v_r^2 + v_t^2) - GM/r + c^2$. The added $c^2$ corresponds to the fact that Einstein adds the rest energy ($E = mc^2$) of an object to it's total energy , while Newton does not. Recall that $\tilde{E}$ represents energy per unit rest mass.

The second conserved quantity, angular momentum, is also obtained from the geodesic equations, i.e.,

$$\frac{d\phi}{d\tau} = \frac{\tilde{L}}{r^2},$$

where $\tilde{L}$ is the relativistic angular momentum parameter, giving

$$\tilde{L} = r^2\frac{d\phi}{dt}\frac{dt}{d\tau} = \frac{r\ v_t}{\sqrt{g_{tt} - g_{rr}v_r^2/c^2 - v_t^2/c^2}}, \tag{6.8}$$

which approximates to the Newton case in the low gravity, low velocity limit, where $g_{tt}, g_{rr} \to 1$ and $v_r, v_t \ll c$. Recall that the $\tilde{L}$ has the unit m$^2$/s.

To wrap up the orbital parameters, we can get the relativistic effective potential energy parameter by extracting $v_t$ from eq. 6.8 and substitute it into eq. 6.7 (with $v_r = 0$), giving

$$\tilde{V} = c^2\sqrt{g_{tt}(1 + \tilde{L}^2/(r^2c^2))}, \tag{6.9}$$

which is equal to $\tilde{E}$ at the turning points of the orbit. In the low gravity, low angular momentum limit ($g_{tt} \approx 1$ and $\tilde{L}/c \ll r$), the equation approximates to the Newton case (plus $c^2$), i.e., $\tilde{V} \approx \frac{1}{2}\tilde{L}^2/r^2 - GM/r + c^2$. Figure 6.3 shows the relativistic and the Newton effective potentials on the same scale, with the $c^2$ added to the Newton case to make it easy to compare.

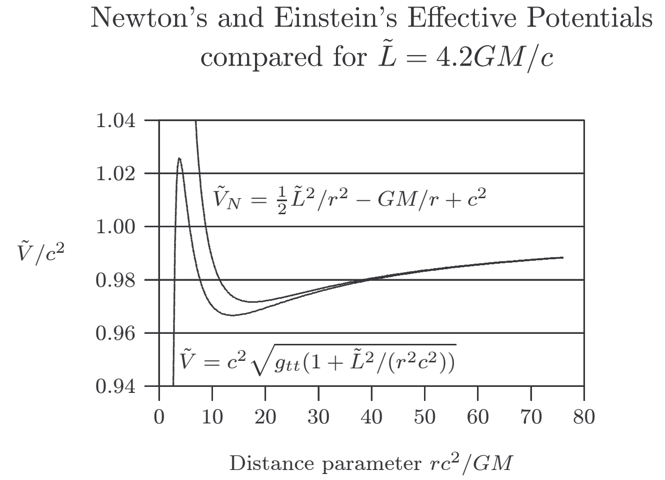


**Figure 6.3:** The top curve is the Newton effective potential (with $c^2$ added to make it easy to compare) and the bottom curve the general relativistic effective potential, both drawn for the same $\tilde{L}$. Note how closely the two curves match for distances $rc^2/GM > 50$.

These two curves show the dramatic differences between the effective potentials given by the two theories, at least for the 'near' region. For $rc^2/GM > 50$, the differences get vanishingly smaller (always remembering the $c^2$ that was added for easy comparison, of course).

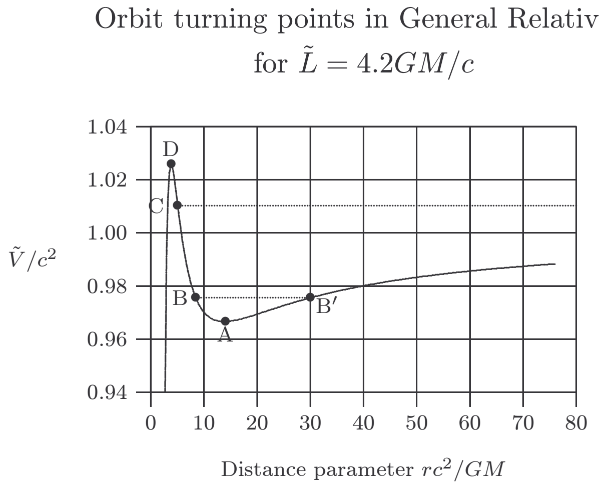
In the near region it is clear that, for the same $\tilde{L}$ and $\tilde{V}$, a relativistic orbit's periapsis will generally be closer and the apoapsis (slightly) farther from the black hole than for the Newton case.*

> *The fact that we have added $c^2$ to the Newton effective potential, makes no difference to the turning points.

Generally, one can say that the eccentricity of a relativistic orbit is larger than the equivalent Newton orbit. Further, for a given $\tilde{L}$, circular orbits (at the minima of the curves) are closer to the black hole for relativistic orbits than for Newton's. This comes from the larger gravitational acceleration in relativistic dynamics.

Just like in the case of the Newton effective potential, the relativistic curve can be used to determine broad characteristics of various orbits. The three usual types of orbits for objects with mass, i.e., circular, non-circular (closed) and open, are shown in figure 6.4.

Points **A**, **B** and **C** are the equivalents of the corresponding points in Newton dynamics. At point **D** (the peak of the curve), a circular orbit is theoretically possible, but unlike point **A**, it is unstable and any perturbation will cause the object to either escape or spiral into the black hole.



Orbit turning points in General Relativity for $\tilde{L} = 4.2GM/c$

**Figure 6.4:** Points **A, B** and **C** represent the energies for circular, closed and open orbits respectively, around a Schwarzschild black hole. Point **D**, the maximum of effective potential (for $\tilde{L} = 4.2GM/c$), represents an unstable circular orbit that will either escape or spiral into the hole.

At energy levels larger than point **D**, there is no orbit solution possible and objects will either escape or be 'swallowed' by the hole, depending on the geometry of the orbit. Contrast this with the Newtonian case, where such an object could in principle enter the black hole, swing around the central singularity and then escape again.

In general relativity, no orbiting object without some means of propulsion, can venture as close as $r = 3GM/c^2$ to a Schwarzschild black hole, without being dragged in and eventually being crushed out of existence in the central singularity.

**Solution of the orbit's turning points** The shape of the effective potential energy curve changes for different values of the angular momentum parameter $\tilde{L}$. Generally speaking, the trough and the peak shifts further apart when $\tilde{L}$ increases and move closer to each other when $\tilde{L}$ decreases—closer both along the energy and the distance axis.

As is clear from figure 6.4, the turning points of an orbit are where $\tilde{V}^2 = \tilde{E}^2$, i.e., where the constant energy line cuts the effective potential curve. For any closed orbit, the line $\tilde{E}^2 =$ constant cuts the effective potential curve in three places.

For open orbits that is either escaping, or falling into the mass, the constant energy line cuts the curve in only two places. The radial distances of the turning points are given by the following third order (or cubic) equation*

*From eq. 7.8 in the next main section, with the radial rate $du/d\phi = 0$.

$$u^3 - u^2/2 + \bar{M}^2 u/\tilde{L}^2 + (\tilde{E}^2 - 1)\bar{M}/(2\tilde{L}^2) = 0,$$

where $u = \bar{M}/r$, $\bar{M} = GM/c^2$ (i.e., geometric units where $G = c = 1$, so that $\bar{M}$ has the units metres and $u$ is dimensionless), all for mathematical convenience and clarity. To find the values, we have to solve a third order equation of the form

$$u^3 + au^2 + bu + k = 0.$$

The algorithm for solving this form is given in most books on mathematical formulae, and it is a bit of a thriller! For those interested, a tailored version is shown in the box at the end of this chapter (page 113).

The important thing is however that relativistic orbits are 'solvable' to some degree. In the latter part of this chapter, we will have a closer look at the general solution of relativistic orbital equations.

**The minimum value of effective potential** At $\tilde{L} \cong 3.4651GM/c$, the peak coincides with the trough, precisely at $r = 6GM/c^2$. This represents the minimum angular momentum that will allow a stable orbit around a Schwarzschild black hole, as illustrated in figure 6.5.

Since there is no longer a 'hump', more total energy will not save an object from falling into the hole, unless it has enough energy to escape *and* that energy is directed outwards. Otherwise, it will eventually pass too close to the hole and spiral into it. An orbiter must have an angular momentum parameter $\tilde{L} > 3.4651GM/c$ in order to remain in a stable, closed orbit around a black hole.

This is not quite 'rocket science', because an Earth orbiter also needs a minimum angular momentum in order to stay above the atmosphere. If the launch vehicle fails to deliver this angular momentum, e.g., if it delivers too much radial velocity instead of transverse velocity, the satellite will eventually fall back and enter the atmosphere.*

*Unless the total energy is enough to escape, which is not normally the case with satellite launches.

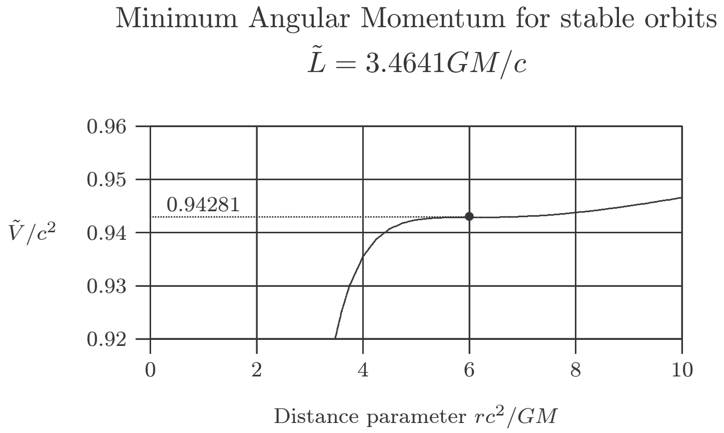


**Figure 6.5:** With angular momentum parameter $\tilde{L} < 3.4641GM/c$, no circular orbit around a Schwarzschild black hole can be stable, but will spiral into the hole. This also means that circular orbits with $r < 6GM/c^2$, or with total energy $\tilde{E} < 0.94281c^2$ are unstable for such a black hole.

Elliptical orbits can venture closer to the hole than $r = 6GM/c^2$, provided that they have enough angular momentum to create a 'hump' in the effective potential curve—and provided that their total energy parameter $\tilde{E}/c^2$ is not larger than the amplitude of the 'hump'.

It seems that a reasonably 'safe' angular momentum is one that creates a 'hump' amplitude $\tilde{V}/c^2 = 1$; at least an orbiting object will then escape if the total energy parameter is larger than the hump. This is (again) provided that the radial component of the energy is directed outwards.

It is fairly easy to show that $\tilde{L} = 4GM/c$ provides just that—a 'hump' amplitude equal to the escape energy. This $\tilde{L}$ curve has a trough at $r = 12GM/c^2$, which can be considered the 'ultra safe' circular orbit radius around a Schwarzschild black hole.

If something large hit's your ship and imparts a lot of energy on it, chances are pretty good that the ship will escape from the hole. It is only if the collision imparts a lot of radially inwards energy on your ship that you are still in trouble. But then, if you had the propulsion system to take you to a black hole in the first place, that system should be good enough to save your skin.

A spaceship with a serious propulsion system can venture quite close to a Schwarzschild black hole. This is because it can manage quite different dynamics than what a purely free-falling object has at it's disposal.

However, near a black hole, the dynamics become very interesting. Suppose you have a spaceship in circular orbit at a reasonably safe distance (say at $r = 7GM/c^2$) from a rather large black hole. By activating a series of short reverse thrust 'burns', you can reduce the ship's orbital energy and momentum with each burn and settle into a stable, yet slightly smaller orbit after each burn.

With a bit of juggling, you can circularize your orbit at each step. That is until you reach the *marginally stable orbit radius* of $r = 6GM/c^2$. You have reached the minimum angular momentum for stable circular orbits, $\tilde{L} = 3.4641GM/c$.

Any further reverse thrust burns will cause the spaceship to spiral inwards, rather than settle into a slightly smaller stable orbit. You will have to continously juggle between reverse thrust and forward thrust burns, in order to slow down the inspiral, because circular orbits are no longer stable.

As the ship nears the limit for *marginally bound orbits*, i.e., $r = 4GM/c^2$, the burns become very critical in terms of exactly how much total energy the ship retains after the burn. A bit too much energy and the ship will rapidly spiral outwards and tend to escape from the hole.

At $r = 4GM/c^2$, the circular orbit velocity equals the escape velocity—hence a pretty unstable situation arises. A bit too little energy and the ship will tend to plunge dangerously inwards. Suppose that, with the aid of the on-board computer, you succeed in achieving a slow inspiral by a precise combination of reverse and forward thrust burns.

The speed of the ship will build up rapidly and as the orbit radius nears the value $r = 3GM/c^2$, the ship will approach the local speed of light. Since it cannot quite reach the speed of light, even forward thrust burns cannot save you and the ship from being dragged into the hole.

The only way out is to arrange the burns so that there is a radially outwards component in the thrust. In fact, any forward thrust component will be a waste of energy, because the orbital speed will remain more or less constant, near the local speed of light. You should rather orientate the ship so that all the thrust is directed radially outward. In fact, it will be better to thrust somewhat backwards and outwards, because the velocity vector of the ship will never exceed the speed of light.

What is more, the engine should be kept running constantly at just the right thrust so that a slow inspiral is achieved. Even this scheme is full of danger, because controlling a spaceship orbiting at practically the speed of light, so close to a black hole ($r < 3GM/c^2$), is a horrendously complex task.

Before crossing the 'extreme danger radius', you should rather abandon the exercise and blast the ship to a safe orbital radius again. The only 'safe' procedure for getting close to the event horizon of a Schwarzschild black hole is to slowly descent radially from a safe distance and then hover just above the event horizon with the engine blazing.

One complication of this scheme is that, unlike in a free-falling orbit, you and your crew might be subjected to horrendously high 'g-forces', depending on the size of the black hole. Interestingly enough, the more massive the hole, the easier it will be on your bodies, because you will be more distant from it's centre.

There will also be tidal forces that tend to stretch you in the radial direction and squeeze you in transverse directions relative to the hole. We will defer discussion of these forces until chapter 9, where they will be analyzed.

### 6.3.2 Effective potential of light

The above analysis of orbits does not hold for light, because the 'rest mass' of a photon is zero and the angular momentum parameter is undetermined ($L/m = r^2 d\phi/d\tau$, where both $m$ and $d\tau$ are zero). There are however ways to determine an equivalent 'effective potential' for light.

Both $\tilde{E}$ and $\tilde{L}$ goes to infinity, but the ratio of the two, $b = \tilde{L}/\tilde{E}$ remains finite and is called the *impact parameter* for light. This is however not the effective potential for light, which is defined as (e.g., [MTW])

$$B^{-2} = \frac{g_{tt}}{r^2} c^2, \tag{6.10}$$

plotted in figure 6.6 in normalized form.

One can think of the effective potential of light as the transverse velocity of light divided by the distance. The transverse speed of light ($c_{tr}$) at Schwarzschild radial distance $r$ is $c_{tr} = \sqrt{g_{tt}}$.

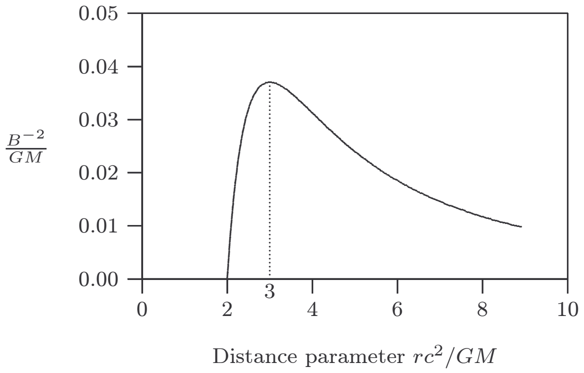


**Figure 6.6:** The effective potential for light. When the closest approach is $r > 3GM/c^2$, the light will be deflected, but will always escape the black hole. At $r = 3GM/c^2$, the light can be in an unstable circular orbit around the hole, while with $r < 3GM/c^2$, the light will always spiral into the hole.

The angular movement in time $dt$ is $d\theta = c_{tr}dt/r$ and therefore the instantaneous angular velocity at closest approach (when the movement is purely transverse) is

$$\frac{d\theta}{dt} = \frac{\sqrt{g_{tt}}}{r}c = B^{-1}.$$

The effective potential for light, as defined above, is just the square of the angular velocity at closest approach, as plotted in the figure.

The curve has no local minimum value, but tends to zero when either $r \to 2GM/c^2$, or $r \to \infty$. There is only one bound orbit possible, and that is at the maximum, when $r = 3GM/c^2$, but even that is an unstable situation, because any perturbation will cause the light to either escape or fall into the black hole.

Light rays that pass a Schwarzschild black hole at a closest distance of $r > 3GM/c^2$ will be deflected, but cannot be captured. Passing light rays that venture inside $r = 3GM/c^2$ will always spiral into the hole. At the borderline case between the two, there is a change that the light may be captured into a circular orbit.

It is however possible for light rays generated (or reflected from) inside the $2GM/c^2 < r < 3GM/c^2$ zone to escape from a Schwarzschild black hole, but then the light needs a positive radial velocity component. From just outside the Schwarzschild radius ($r = 2GM/c^2$), light only escapes if the initial direction is very close (or equal) to precisely radially outward.

It is shown in [MTW] that, in order for light to escape from $r < 3GM/c^2$, the angle between the propagation direction of the light and the radial direction must satisfy

$$\sin\delta < 3\sqrt{3}\,\frac{GM}{c^2}B^{-1} = 3\sqrt{3}\,\frac{GM}{c^2}\frac{\sqrt{g_{tt}}}{r}.$$

The inverse is also true: all infalling light rays will tend towards precisely radially inward as they near the event horizon. If you hover your spaceship (with engines blazing) just outside the event horizon, all incoming light will come from a narrow cone directly 'overhead'.

The whole 360 degrees of visible universe will be compressed into that cone and the rest will be just black, creating the illusion that the black hole is busy engulfing you—not a nice place to be!

Kip Thorne describes this vividly in [Thorne]. To avoid capture, the condition for incoming light rays that are still outside $r = 3GM/c^2$, is the opposite of the above, i.e.,

$$\sin\delta > 3\sqrt{3}\,\frac{GM}{c^2}\frac{\sqrt{g_{tt}}}{r},$$

meaning that the (inwards) propagation direction must be deviate from the radial by more than $\delta$.

Now that we have a fair idea of the mechanisms of relativistic orbits, we can attempt to also understand the trajectories that material objects and light will follow outside of a Schwarzschild black hole. To understand, we have to turn to the actual orbital equations, which is the subject of the next chapter.

# Chapter 7

# Orbital equations

from geodesics
to
real orbits

So far, we have used only the effective potential to establish the broad relativistic orbital behavior. In this section we will discuss the general orbital equations for Newtonian and relativistic orbits around Schwarzschild black holes.

Circular orbits are still relatively simple, since they have a constant radius and are therefore not influenced by space curvature. Non-circular orbits are pretty complex, because the object moves through curved space for most of the time.

With the equations for $\tilde{L}$, $\tilde{E}$ and $\tilde{V}$, as derived in a previous section, it is only possible to directly solve for $r$ at the turning points of orbits. In this section, we will first look at Newton orbits again and then show how the relativistic solution differs from Newton's.

The approach followed here is a bit of a mixture between the approaches of [Faber] and [MTW]. See the box *Comparison to Faber and MTW orbital equations* on page 116 for some clarification.

**Newtonian orbits** As we have shown before, Newtonian orbits can be solved by using Kepler's laws. Alternatively, the solution can be found by using Newton's law of gravity and his laws of motion.

By writing down the Newtonian equation of motion for an object moving in a gravitational field and solving it, the following differential equation can be obtained, e.g., [Faber]:

$$\frac{d^2r}{dt^2} = r\frac{d\phi^2}{dt^2} - \frac{GM}{r^2}, \tag{7.1}$$

where $\phi$ is the orbital angle as measured from the point of closest approach. This gives the rate of change of the radial velocity ($d^2r/dt^2 = v_r/dt$). It equals, as expected, the sum of the positive centrifugal acceleration and the negative gravitational acceleration.

Note that $rd\phi^2/dt^2 = v_t^2/r$, the more well known expression for centrifugal acceleration. It is easy to see that for circular orbits, $d^2r/dt^2 = 0$ and one can directly solve the for constant circular orbital speed, i.e.,

$$v_o = \sqrt{GM/r} \tag{7.2}$$

For purposes of comparison with general relativity later, equation 7.1 can be (laboriously) reworked into

$$\left(\frac{du}{d\phi}\right)^2 = 2\frac{(GM)^2}{\tilde{L}_N^2 c^2}\frac{\tilde{E}_N}{c^2} + 2\frac{(GM)^2}{\tilde{L}_N^2 c^2}u - u^2,$$

where again, $u = GM/(rc^2)$, the dimensionless (inverse) distance parameter that simplifies the mathematics somewhat. Now differentiate $u$ with respect to $\phi$ and divide by $2\frac{du}{d\phi}$, to get a second order differential equation*

$$\frac{d^2u}{d\phi^2} = \frac{(GM)^2}{\tilde{L}_N^2 c^2} - u, \tag{7.3}$$

which can be solved analytically, e.g., [Faber, p. 193]. The result is the equation for a conic section

$$u = \frac{(GM)^2}{\tilde{L}_N^2 c^2}(1 + e\cos\phi)\ , \tag{7.4}$$

or, restoring $r$:

$$r = \frac{\tilde{L}_N^2}{GM(1 + e\cos\phi)}\ , \tag{7.5}$$

the same as we had before, when we obtained the orbital equation through the Kepler laws for planetary motion (equation 6.4). We will now investigate how the relativistic orbital equations differ from the Newtonian ones above.

**Relativistic orbits** Although relativistic orbits can be solved by the same procedure as above, this time through the relativistic equations of motion, it is very complex process. A slightly simpler procedure is through the geodesic equations, where the rate of change of $r$ is obtained relative to proper time $\tau$, e.g., [MTW] eq. (25.16):

$$\left(\frac{dr}{d\tau}\right)^2 + \tilde{V}^2 = \tilde{E}^2, \tag{7.6}$$

where $\tilde{V}$ and $\tilde{E}$ are the relativistic effective potential and the energy parameters. From the definition of the angular momentum parameter ($\tilde{L} = r^2 d\phi/d\tau$), $d\tau$ can be extracted and replaced in equation (7.6), resulting in

$$\left(\frac{dr}{d\phi}\right)^2 = \frac{r^4}{\tilde{L}^2 c^2}(\tilde{E}^2 - \tilde{V}^2). \tag{7.7}$$

*Note that with $u = \frac{GM}{rc^2}$, $\frac{dr}{d\phi} = \frac{-r^2c^2}{GM}\frac{du}{d\phi}$.

Because $\tilde{V}$ is a function of $r$, this equation is not analytically solvable. It is however very easy to solve numerically (point by point), as will be shown later.

Since we would like to compare the relativistic case with the Newtonian case, it is helpful to differentiate this equation to obtain the equivalent second order differential equation. Again employ $u = GM/(rc^2)$ and writing out $\tilde{V}^2 = (1 - 2GM/(rc^2))(1 + \tilde{L}^2/(r^2c^2))$ in equation (7.7) one gets*

*Note that with $u = \frac{GM}{rc^2}$, $\frac{dr}{d\phi} = \frac{-r^2c^2}{GM}\frac{du}{d\phi}$.

$$\left(\frac{du}{d\phi}\right)^2 = \frac{(GM)^2}{\tilde{L}^2c^2}(\frac{\tilde{E}^2}{c^4} - 1) + 2\frac{(GM)^2}{\tilde{L}^2c^2}u - u^2 + 2u^3, \tag{7.8}$$

an easily differentiable form. Following the recipe used in the Newtonian analysis, differentiate $u$ with respect to $\phi$ and divide by $2\frac{du}{d\phi}$, to get the second order orbital equation

$$\frac{d^2u}{d\phi^2} = \frac{(GM)^2}{\tilde{L}^2c^2} - u + 3u^2. \tag{7.9}$$

When compared to the Newtonian orbital equation (7.3), we see that apart from the constants $\tilde{L}_N$ and $\tilde{L}$ that differ, there is an additional 'relativistic' term, $3u^2$. It is immediately obvious that, far from mass $M$, where $3u^2$ is small compared to the other terms, the equation approaches the Newtonian solution. The extra term makes the equation analytically unsolvable though.

This fact becomes apparent when, after solving any one of equations (7.7), (7.8) or (7.9) numerically, one finds that closed orbits do not repeat themselves—they precess in the direction of normal orbital movement.

This precession becomes increasingly more complex as the orbit comes closer to the black hole. This precession was first observed as the perihelion shift of Mercury, as will be discussed in the next chapter.

**Numerical solution of orbital equations** By the theorem of differentials, equation (7.7) may be written as

$$dr = \pm\frac{dr}{d\phi}d\phi = \pm\frac{r^2}{c\tilde{L}}\sqrt{\tilde{E}^2 - \tilde{V}^2}\ d\phi, \tag{7.10}$$

still in geometric units, where $c, G = 1$. Choosing a small constant positive angular increment $\Delta\phi$, we can approximate the orbital equation by

$$\Delta r \cong \pm\frac{r^2}{c\tilde{L}}\sqrt{\tilde{E}^2 - \tilde{V}^2}\ \Delta\phi, \tag{7.11}$$

and then numerically integrate to find the radius $r_j$ for any orbital angle $\phi_j = \sum \Delta\phi$.

This equation does cause a few hassles though. Firstly, it does not 'know' in which direction $r$ is changing, so one must find a way of telling the algorithm. More seriously, at the periastron and apastron, $\tilde{E} = \tilde{V}$ so that

$dr/d\phi = 0$, i.e., $\Delta r = 0$ for any $\Delta\phi$. Straight forward numerical integration then 'locks up' and $r$ remains constant for varying $\phi$ (an anomalous circular orbit).

There are various ways to overcome these problems, but it is easier to use the second order differential equation (7.9). It overcomes the problems more readily, because there is no square root taken and $\frac{d^2u}{d\phi^2}$ is non-zero at the periastron and apastron. A simple method of numerically solving equation (7.9) is to recognize that

$$\Delta u_n \cong \Delta u_{n-1} + \frac{d^2u}{d\phi^2}\Delta\phi^2, \tag{7.12}$$

where $\Delta u_{n-1}$ is the previously calculated $\Delta u$ in the numerical integration. Fortunately, $\Delta u_{n-1}$ and $\frac{d^2u}{d\phi^2}$ cannot become zero at the same time, so the numerical integration will not 'lock up', except in the highly unlikely event that $\Delta u_{n-1} = -\frac{d^2u}{d\phi^2}\Delta\phi^2$ (identically) occurs. To guard against this, one must check for the event and give the orbiter the tiniest of 'nudges' in the right direction.

One remaining hassle is that the algorithm does require knowledge of the 'previous value of $\Delta u_n$' before the integration loop starts. This can be overcome if the orbit is started at periapsis or apoapsis, where $\Delta u_{n-1} \cong -\Delta u_n$, so that

$$\Delta u_{n-1} \cong \tfrac{1}{2}\frac{d^2u}{d\phi^2}\Delta\phi^2. \tag{7.13}$$

Hence, for the next round, one knows the problematic first $\Delta u_{n-1}$.

This method was used to plot the orbit in figure 7.1, using the algorithm shown in box *Numerical solution of the second order orbital equation* on page 114.

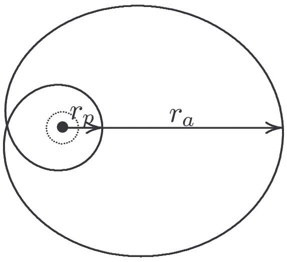


**Figure 7.1:** A relativistic orbit quite close to a non-rotating black hole. The dotted circle represents the event horizon and the orbit's closest approach is at $r_p/\bar{M} = 5$, or $2\frac{1}{2}$ times the event horizon radius. The orbit parameters $\tilde{E}, \tilde{L}$ have been chosen so that the periastron shift is $2\pi$ radians per orbit, causing the orbit to repeat itself after making a double loop. For the vast majority of values $\tilde{E}, \tilde{L}$, relativistic orbits will not repeat themselves, due to periastron shifts other than $2\pi$, or multiples thereof.

There are more sophisticated methods for solving the equation numerically, but this one is particularly simple, although it is mainly of pedagogical value.

In real life, orbits tend to be even more complex, with multiple sources of gravitation acting upon the orbiting body.

Further, most gravitational sources will be moving relative to whatever reference frame one chooses. Such cases must be solved by perturbation methods or point by point approximations.

**Orbital equation for light** The Schwarzschild metric, arranged as a light-like interval ($d\tau = 0$), can be used to find the orbital equation for light. It is however very easy to see how equation 7.9 must be modified for light, where $\tilde{L} \to \infty$.

It means that the term $(GM)^2/(\tilde{L}^2c^2)$ in the equation for material bodies falls away and the orbital equation for light becomes simply

$$\frac{d^2u}{d\phi^2} = -u + 3u^2. \tag{7.14}$$

This equation can be used to obtain the gravitational deflection of light when it passes a massive body, as will be discussed in the next chapter. It is still not a solvable differential equation and successive approximations or numerical methods must be used, depending on the purpose of the 'solution'.

## 7.1 A quasi-Newtonian, 'poor man's orbit'

It is possible to get a pretty accurate orbital plot from the radial and transverse quasi-Newtonian relativistic acceleration components of chapter 4, i.e.,

$$a_r = \frac{-GM}{r^2}\left(g_{tt} - 3g_{rr}\frac{v_r^2}{c^2} + 2\frac{v_t^2}{c^2}\right)$$

and

$$a_t = \frac{2GM}{r^2}g_{rr}\frac{v_r v_t}{c^2},$$

Now the usual methods employed by engineers when simulating dynamic systems can be used, i.e., determining accelerations along the $x$, $y$ and $z$ axes of a chosen coordinate system and then numerically integrating for velocities and displacements. For a planar orbit, one needs only a two-dimensional coordinate system.

Working with a small positive angular increment $\Delta\phi$, one implements the following broad algorithm (a detailed algorithm is shown in the box on page 115):

```
Begin algorithm
For a specific orbital position:
  Obtain $g_{tt}$
```

```
    Calculate the radial and transverse accelerations
    Rotate these accelerations to the coordinate axis through the orbital angle
    Integrate for the new $x$ and $y$ velocities and positions
    Obtain the new radial distance and new orbital angle
    Obtain the new radial and transverse velocities by rotating
      through the negative of the new orbital angle
Repeat for the new orbital position
End Algorithm
```

The algorithm does not fully conserve total energy and angular momentum along an elliptical orbit, i.e., it is not fully 'conservative'. Over a full orbit, the errors tend to mostly cancel out.

The reason for the small error is that $a_r$ and $a_t$ are kept constant over a time interval $\Delta t$, while in elliptical orbits, they are continously changing. The errors are positive for half of a full orbit and negative for the other half. The 'non-conservative' nature of the algorithm will however cause the orbit to eventually diverge away from a fully 'conservative' orbit.

This aside, the 'poor man's orbit' does remarkably well. In comparison to the more accurate numerical solution of equation 7.9, the error per revolution is very small, visible as a slight anomalous peripsis shift over many revolutions. The error is a function of the time increment per cycle—the smaller the more accurate, at the expense of simulation time, of course.

The test orbit's starting parameters were as follows: $r = 5GM/c^2$, $\phi = 0$, $\Delta\phi = \pi/100000 \approx 30\mu$rad, $v_t = 0.4675c$, $v_r = 0$, giving $\tilde{L} = 3.785GM/c$, and $r_{max} = 26.787GM/c^2$.

The orbit is the same one as pictured in figure 7.1 on page 105, with 360 degrees periapsis shift per orbit. Apart from the slight periapsis shift error, the shape of the orbit is the same as obtained through the more accurate algorithm. The author likes to think that engineers will relate more readily to this algorithm than to the more formal one.

For one thing, it is hard to think of a better way of illustrating the apparent opposing acceleration in the transverse direction. Switch that acceleration off in the simulation and the orbit shape changes dramatically.

There is just no way that a purely radial acceleration can produce an orbit that even approaches the 'real thing', at least not in the strong field, high velocity environment close to a black hole.

For the reader who might want to test out the algorithm on a computer, a few notes. $GM/c^2$, $\Delta t$ and $r$ must all be in metres. The author used a constant angular movement per cycle, simply because it speeds up the plot when far from the mass, where the errors are smaller due to the lower velocities. Lastly, $GM/c^2$ is actually just a scale factor in the plot and can just as well be chosen as unity.

The algorithm is very useful in studying the order of magnitude of the opposing acceleration effects. Apart from being able to view the orbital changes, it is relatively easy to (numerically) extract the values of $v_r$ and $v_t$ anywhere along the orbit.

In the orbit of planet Mercury, the maximum value of $v_r v_t$ is of the order $10^{-8}$. With $2GM/c^2/r^2$ of the order $10^{-18}$ m$^{-1}$, the opposing transverse acceleration is of the order $10^{-26}$ m$^{-1}$ maximum, and when multiplied by $c^2$, it translates to only $10^{-9}$ m/s$^2$.

This is a really tiny opposing transverse acceleration (some 0.1 nano-g). It can however be shown that it is the main contributor to the perihelion shift, as will be illustrated in the next paragraph.

Use the test orbit above, giving a periapsis shift of $2\pi$ radians and then suppress the opposing transverse acceleration. The algorithm then produces a periapsis shift of about $\pi/2$ radians. One can argue that the other $3\pi/2$ radians must have come from the opposing transverse acceleration.

The values of the radial and transverse acceleration components for the test orbit can also be easily found from the simulation. The maximum opposing transverse acceleration is $a_{t(max)} \approx \pm 2.7 \times 10^{-3} GM/c^2$ m$^{-1}$ and the equivalent total radial acceleration at that time is $a_{r(eq)} \approx -2.7 \times 10^{-2} GM/c^2$ m$^{-1}$.

So the magnitude of the opposing transverse acceleration component reaches some 10% of the magnitude of the radial acceleration component. This ratio gets higher the closer the periapsis is to the gravity generating mass. This 'explains' the huge periapsis shifts and also the high eccentricity of highly relativistic orbits.

One must however always remember that the transverse acceleration is not 'real', in the sense that there is no 'force' in that direction, or in any other direction for that matter. The orbiter simply moves along a spacetime geodesic, following a 'straight' path in curved spacetime.

The distant observer, essentially sitting in (locally) flat spacetime, observes the orbit and may conclude that there is acceleration, both radial and transverse, as shown in the above equations for $a_r$ and $a_t$.

To conclude this chapter on orbital dynamics, we will briefly examine one of the interesting mysteries of relativistic orbits—how can they work if gravity propagates at the speed of light, as all effects in relativity are forced to do?

## 7.2 The 'Speed' of Gravity

Up to now, we have worked with small objects orbiting a large mass, permanently at rest at the origin of an inertial coordinate system. The gravitational field was therefore static in the coordinate system and the 'test object' followed a spacetime geodesic that was determined by the static curvature of spacetime at it's location.

What happens if the test particle is replaced by a massive object of comparable size to the mass at the origin? Such situations are common in the universe, particularly in binary star systems, where the two stars orbit around their common centre of gravity.

The easiest way to analyse binary systems is to choose the coordinate system so that the common centre of gravity is permanently at rest at the origin. Now both masses will be moving (and accelerating),* relative to the chosen coordinate system.

*Accelerating in a Newtonian sense; following spacetime geodesics in a relativistic sense.

So what happens to the gravitational field? It is reasonable to expect that the field must be varying continuously at every point in the coordinate system. According to Newton's orbital theory, the variations in the gravitational field must occur *simultaneously* with the change of the positions of the two bodies.

This means that the effect of the change in position of a mass must 'propagate' at infinite speed to every point in the coordinate system—an effect called 'action at a distance'. Without this assumption, Newton orbits cannot work as they do, meaning they will not be stable.

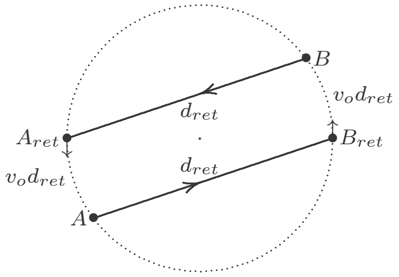


**Figure 7.2:** Two identical stars in circular orbits around their common centre of gravity (the centre point of the circle). If gravitational forces propagate at the speed of light, then star $A$ will experience a gravitational force towards the 'retarded' position of star $B$ (position $B_{ret}$) and visa versa for star $B$. The 'retarded' distance between the stars is indicated by $d_{ret}$. In the time it takes light to propagate distance $d_{ret}$, the stars move a distance $v_o d_{ret}$ along the circular orbit, where $v_o$ is the orbital velocity relative to the centre of the circle.

If no effect can propagate faster than the speed of light, as relativity demands, how does a Newton orbit remain stable? Figure 7.2 illustrates the case where the gravitational forces do not point towards the centre of gravity of the two stars, but rather towards the 'retarded' positions of the stars.

The retarded positions is where an observer on each star would have observed the other star, taking into account the propagation delay of light over the distance $d_{ret}$. This situation does not allow stable orbits in Newton mechanics—there will be residual transverse forces that tend to speed up the orbits and make the two bodies drift apart.

Since we observe stable orbits, we may conclude that nature does not work this way—gravitational effects apparently propagate at a speed of near infinity. The problem is that Einstein's relativity theory forbids any effect to propagate faster than the speed of light—and that includes gravity.

So how can Einstein's theory of gravity produce stable orbits? Further, since Einstein's gravity reduces to Newton's gravity in the limiting case of low fields and low velocity, how can the two theories be reconciled?

The answer lurks in the depth's of solutions to Einstein's field equations,* which are unfortunately outside of the scope of this book. Loosely stated, it tells us that for any mass that moves uniformly relative to an inertial frame, (i.e., a non-accelerated mass), it's gravitational field appears static relative to the mass itself—i.e., it moves as if attached to the mass.

A test particle that is kept stationary in the inertial frame and than released, will immediately start to fall towards the proper position of the moving mass and not towards it's retarded position. Now this is totally unremarkable, because we could just as well have viewed the mass as stationary and the test particle as moving relative to the mass. The gravitational acceleration of such a particle would surely be towards the proper position of the stationary mass.

Further, general relativity also tells us that if we could somehow abruptly stop the movement of the mass relative to the inertial frame, the moving gravitational field will also stop moving, but the 'stop effect' will propagate at the speed of light from the mass outwards.

This means that the field will be deformed for a period of time, with some (outer) parts of the field still moving at the original speed and some (inner) parts having already stopped moving. A test particle at some distance from the mass will continue to be curved towards the "extrapolated" position of the moving mass until such time as the change in the gravitational field reaches the particle.

At that time the particle will start to curve towards the proper position of the now stationary mass. Figure 7.3 illustrates the situation in a simplified way.

If the above still sounds weird, take comfort from the fact that Maxwell's equations for electromagnetic radiation predicts exactly this behavior for the field around a moving charge. If the moving charge is abruptly stopped, the field at some measurement point keeps pointing towards the extrapolated position of the charge until the effect of the change in the velocity of the charge has had time to propagate towards the measurement point. This effect has been measured in the laboratory.

The foregoing is readily comprehensible for a uniformly moving mass, but what about the two stars in orbit around their common centre of gravity? Surely, in the inertial reference frame, both masses are constantly being

*For a brief, but excellent technical treatment, search the Internet for 'The speed of gravity revisited' [Ibison].

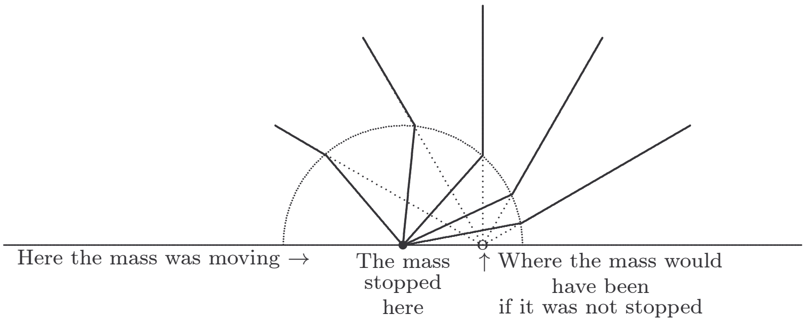

Here the mass was moving →
The mass stopped here
↑ Where the mass would have been if it was not stopped

**Figure 7.3:** A mass was moving uniformly relative to an inertial frame, but is abruptly stopped at time $t = 0$. The effect of the change in velocity of the mass propagates radially outwards at the speed of light (a sphere with radius $c \times t$). Particles outside of the sphere at any given time experience gravity towards the position where the mass would have been if it was not stopped. Particles inside the sphere will experience gravity directly towards the now stationary mass.

accelerated towards their common centre of gravity.

But do the stars experience acceleration in the sense that a force is acting upon them? Not quite.* The two stars are in free-fall and are for all practical purposes inertially moving masses. More technically correct, they are moving along space-time geodesics, which can be thought of as moving in straight lines through curved space-time.

Their respective gravitational fields are moving with them, just as in the inertially moving mass in the absence of gravity. Relative to their space-time geodesics they experience no acceleration, so their gravitational fields are not deformed, except for small higher order effects that will be discussed below.

If we consider one mass as the 'test particle', that mass will always curve directly towards the proper position of the other mass and not towards it's retarded position. If the stars have sufficient transverse velocity, they will stay in stable orbits around each other.

The above is an oversimplification of the real situation, because general relativity also tells us that there is a higher order interaction between the gravitational fields of the two stars. This interaction causes some deformation of the gravitational fields of both stars and the deformation continuously robs them of a little orbital energy.

For normal binary stars the effect is negligible because of the relatively large distance between them. For binary neutron stars that are very massive and very close to each other, the effect becomes appreciable.

This loss of orbital energy is radiated as gravitational waves, which are the subject of a later chapter. The effect is observable when the binary neutron stars also happen to be pulsars. The orbital periods of the binary pulsars

*The only external force that the stars will experience is tidal gravity, which is the subject of a later chapter.

can be deduced very accurately from their highly stable 'light house' effect.

The rate of orbital decay agrees with Einstein's theory to within 1%, which is the experimental uncertainty. Alternatively stated, the binary pulsars tell us that the 'speed of gravity' is within 1% of the speed of light. Einstein's theory of gravity says that gravitational waves and other gravitational disturbances propagate at exactly the speed of light. So it seems that Einstein was right again!

In summary then, because of the 'static' nature of the gravitational field of an inertially moving mass, gravity has the *appearance* of instantaneous propagation. However, static gravitational fields do not propagate, just as static electromagnetic fields do not propagate. Any change to the static gravitational field that is caused by the acceleration* or the deformation of a gravity generating mass, propagates at the speed of light and so re-establish the changed field.

*Acceleration here means being forced out of it's geodesic path through spacetime.

Orbiting bodies tend to follow space-time geodesics and therefore do not suffer acceleration in the usual sense of the word. Their gravitational fields follow them in their orbits as if they were rigid extensions of the orbiting body.

Where there are more than one mass in a system that generate gravitational fields, the fields do interfere with each other in a way that radiates some of the orbital energy of the system away as gravitational waves.

### Algorithm for finding the periapsis and apoapsis for a closed relativistic orbit around a black hole

We have to find the solution to the equation $u^3 + au^2 + bu + k = 0$ Declare all parameters used as double precision floating point values, give conventional (SI) values to $G,\ M,\ c,\ \tilde{E},\ \tilde{L}$ and calculate the following:

$$
\begin{aligned}
\bar{M} &= GM/c^2 && \text{`geometrize (to units metres)} \\
\tilde{L} &= \tilde{L}/c && \text{`geometrize (to units metres)} \\
\tilde{E} &= \tilde{E}/c^2 && \text{`geometrize (dimensionless)} \\
a &= -0.5 \\
b &= \bar{M}^2/\tilde{L}^2 \\
k &= 0.5(\tilde{E}^2 - 1)\ b
\end{aligned}
$$

Begin algorithm

```
    $Q = (3b - a^2)/9$
    $R = (9ab - 27k - 2a^3)/54$
    $D = Q^3 + R^2$   'note the $Q$-cubed, not squared!
If  $D < 0$ Then   'solution is possible
    $\theta = \cos^{-1}(R/\sqrt{-Q^3})$
    $u_1 = 2\sqrt{-Q}\ \cos\left(\frac{\theta}{3}\right) - \frac{a}{3}$
    $u_2 = 2\sqrt{-Q}\ \cos\left(\frac{\theta + 2\pi}{3}\right) - \frac{a}{3}$
    $u_3 = 2\sqrt{-Q}\ \cos\left(\frac{\theta + 4\pi}{3}\right) - \frac{a}{3}$
    $r_1 = \bar{M}/u_1$
    $r_2 = \bar{M}/u_2$
    $r_3 = \bar{M}/u_3$
Else   'no solution is possible
End If
```

End algorithm

If a solution is found, the distance $r_3$ represents the periapsis. If the line $\tilde{E} =$ constant cuts the effective potential curve in three places (i.e., a closed orbit), $r_2$ represents the apoapsis and $r_1$ the point to the left of the 'hump'. If $\tilde{E} = 1$, i.e., escape energy, $r_2$ will be infinite ($u_2 = 0$) and care must be taken if the algorithm is run on a computer.

### Numerical solution of the second order orbital equation

The algorithm is shown in a sort of ‘pseudo-BASIC’, where $x_n = x_{n-1} + \Delta x$ is written simply as $x = x + \Delta x$. We start the orbit at the periastron or apastron of the orbit, where we set $\phi = 0$.

Declare all parameters used as double precision floating point values Give conventional (SI) values to $r,\ G,\ M,\ c,\ \Delta\phi,\ \tilde{L}$ and then precalculate the following:

```
$\bar{M} = GM/c^2$    ‘geometrize (to units metres)
$\tilde{L} = \tilde{L}/c$    ‘geometrize (to units metres)
$u = \bar{M}/r$
$\frac{d^2u}{d\phi^2} = \frac{\bar{M}^2}{\tilde{L}^2} - u + 3u^2$
$\Delta u = \frac{1}{2}\frac{d^2u}{d\phi^2}$    ‘the starting value $\Delta u_{n-1}$
```

Do Until some condition or user interaction ends loop

```
    $x = r\cos\phi,\quad y = r\sin\phi$,   plot $x$ and $y$
    If $\Delta u <> \frac{-d^2u}{d\phi^2}\Delta\phi^2$ Then   ‘normal procedure
        $\Delta u = \Delta u + \frac{d^2u}{d\phi^2}\Delta\phi^2$
    Else   ‘guard agains lock up
        $\Delta u = 1E-99 \times \mathrm{Sgn}(\Delta u)$   ‘a tiny nudge
    End If
    $u = u + \Delta u$
    $\phi = \phi + \Delta\phi$
    $\frac{d^2u}{d\phi^2} = \frac{\bar{M}^2}{\tilde{L}^2} - u + 3u^2$
    $r = \bar{M}/u$
    If $r <= 2\bar{M}$ Then Exit Do ‘falling into black hole
```

Loop

When the term $\frac{\bar{M}^2}{\tilde{L}^2}$ is left out in both calculations of $d^2u/d\phi^2$ above, the ‘orbit’ of light is obtained.

This algorithm will produce a Newtonian orbit if the term $3u^2$ is left out in both calculations of $d^2u/d\phi^2$ above and $\tilde{L}$ is chosen appropriately (Newtonian).

If both $\frac{\bar{M}^2}{\tilde{L}^2}$ and $3u^2$ are left out, the result is light moving in a straight line, which is what pure Newton mechanics predicts.

### Quasi-Newtonian relativistic orbit algorithm

The algorithm is shown in a sort of 'pseudo-BASIC', where $x_n = x_{n-1} + \Delta x$ is written simply as $x = x + \Delta x$.

Declare all parameters used as double precision floating point values. Initialize $\Delta\phi$, $G$, $M$, $c$, $r$, $\phi$, $v_r$, $v_t$ with starting values and then pre-calculate:

```
$v_x = v_r \cos\phi - v_t \sin\phi$
$v_y = v_r \sin\phi + v_t \cos\phi$
Do Until some condition or user interaction ends loop
    $\Delta t = r\Delta\phi/vt$   'constant angular rate
    $g_{tt} = \sqrt{1 - 2GM/(rc^2)}$
    $g_{rr} = 1/g_{tt}$
    $x = r\cos\phi$,   $y = r\sin\phi$
        (plot $x$ and $y$)
    $a_r = -GM/r^2\ (g_{tt} - 3g_{rr}v_r^2/c^2 + 2v_t^2/c^2)$
    $a_t = 2GM/r^2\ g_{rr}v_r v_t/c^2$
    $a_x = a_r \cos\phi - a_t \sin\phi$
    $a_y = a_r \sin\phi + a_t \cos\phi$
    $x = x + v_x\Delta t + 0.5a_x\Delta t^2$
    $y = y + v_y\Delta t + 0.5a_y\Delta t^2$
    $v_x = v_x + a_x\Delta t$
    $v_y = v_y + a_y\Delta t$
    $r = \sqrt{x^2 + y^2}$
    If $r <= 2GM/c^2$ Then Exit Do   'falling through event horizon
    $\phi = \phi + \Delta\phi$
    $v_r = v_x \cos\phi + v_y \sin\phi$
    $v_t = -v_x \sin\phi + v_y \cos\phi$

Loop
```

### Comparison to Faber and MTW orbital equations

For readers familiar with [Faber] and [MTW], a few words of comparison (note that in contrast with this text, both use geometric units throughout). Faber defines $u = 1/r$ (units $\mathrm{m}^{-1}$), so that $du/d\phi = -(1/r^2)dr/d\phi$, also $\mathrm{m}^{-1}$. Just before Faber's equation (163), the following first order differential equation is derived:

$$\left(\frac{du}{d\phi}\right)^2 + u^2 = \frac{b^2-1}{h^2} + 2\frac{Mu}{h^2} + 2Mu^3, \qquad (7.15)$$

with $M$ being geometrized as $GM/c^2$ (units metres), $b = \tilde{E}$ and $h = \tilde{L}$, as symbolized in this text (the MTW symbols).

MTW uses the same definition for $M$ and defines $u = M/r$ (dimensionless), so that $du/d\phi = -(M/r^2)dr/d\phi$, also dimensionless and hence obtains the equivalent equation

$$\left(\frac{du}{d\phi}\right)^2 = \frac{\tilde{E}^2 - (1-2u)(1+L^{\dagger 2}u^2)}{L^{\dagger 2}}, \qquad (7.16)$$

where $L^\dagger = \tilde{L}/M$, a dimensionless momentum parameter. If $L^\dagger = \tilde{L}/M$ is substituted and the equation expanded, the result is

$$\left(\frac{du}{d\phi}\right)^2 = \frac{M^2}{\tilde{L}^2}(\tilde{E}^2-1) + 2\frac{M^2}{\tilde{L}^2}u - u^2 + 2u^3, \qquad (7.17)$$

similar to eq 7.8 of this text (just the units differ). This equation is readily differentiable, giving the second order equation

$$\frac{d^2u}{d\phi^2} = \frac{M^2}{\tilde{L}^2} - u + 3u^2, \quad \text{(dimensionless)}, \qquad (7.18)$$

MTW does not differentiate equation 7.16, hence the more compact form that it was left in. The Faber equivalent of equation 7.18 is (eq. 163)

$$\frac{d^2u}{d\phi^2} + u = \frac{M}{h^2} + 3Mu^2, \quad (\text{units m}^{-1}), \qquad (7.19)$$

where $u = 1/r$, $M = GM/c^2$ and $h = \tilde{L}$ (unit metres). The differences caused by the alternative definitions of $u$ are small, yet subtle. It makes no difference in any calculations, but may be confusing at times. A good question is: why bother with $u$ at all, i.e., why not use $r$ straightaway? The answer is that while eq. (7.17) is third order in $u$, reducing to second order after differentiation, it would have started as fourth order in $r$, making life more difficult, mathematically speaking.

# Chapter 8

# Some tests of general relativity

perihelion shift, deflection,
time delay, geodetic effect,
dragging of inertial frames

In this chapter, we will discuss the three most well known tests of Einstein's general theory of relativity, i.e., the perihelion shift of Mercury, the bending of light by the Sun's gravity and the time delay of light passing close to the Sun. There have been others, which will be mentioned in later chapters, but the three discussed here are all relatively easily understood, using the knowledge gained in the previous chapter.

If you have not yet read Clifford Will's bestseller on tests of general relativity, "Was einstein Right?" [Will(b)], it is highly recommended. It is an excellent nontechnical description of most of the tests performed up to the mid 1980s, and some that are yet to be performed.

In this chapter, the reader will be filled in with some of the technical details that professor Will had to leave out, probably on the insistence of his publisher, who most likely wanted a popular bestseller!*

*The heavy technical details are found in his book [Will(a)].

## 8.1 The perihelion shift of Mercury

By the time Einstein was working on his general theory, the anomalous shift of the perihelion of Mercury was well known. It was anomalous, because when astronomers carefully applied Newton's theory of gravity to the orbit of Mercury, and they compensated for all the known effects of perturbation caused by the other planets on the orbit, there was a 'residue' of about 43 arcseconds per century.

Einstein was well aware of this and has used it to some degree to 'test' his theory. From the full relativistic orbital equation

$$\frac{d^2u}{d\theta^2} = \bar{M}/\tilde{L}^2 - u + 3\bar{M}u^2, \tag{8.1}$$

it takes just a bit of mathematical sweat to prove that for nearly circular orbits, i.e., $e \approx 0$, the perihelion shift per orbit is

$$\Delta\theta_p \approx \frac{2\pi}{\sqrt{1 - 6\bar{M}/r_o}} - 2\pi, \tag{8.2}$$

where $r_o$ is the average orbital radius around mass $\bar{M}$. This says that instead of sweeping out an angle $2\pi$ between two successive points of closest approach (the perihelion of a planetary orbit), the orbit sweeps out an angle $2\pi$ divided by a 'modified time dilation' $\sqrt{1 - 6\bar{M}/r_o}$, making $\Delta\theta_p > 0$, always.

Let us plug actual values into the equation. The mass of the Sun is about 1477m (geometrically) and the average orbital distance of Mercury from the Sun is about 5.79 $\times 10^{10}$m, giving a perihelion shift per orbit of

$$\Delta\theta_p \approx \frac{2\pi}{\sqrt{1 - 6 \times 1477/(5.79 \times 10^{10})}} - 2\pi \approx 4.88 \times 10^{-7} \text{ radians.}$$

Mercury completes about 415 orbits every Earth century, giving the answer per century as about 0.2 mrad., or about 41 arcseconds.

This falls short from the actual measurement by 2 arcseconds, which is due to the fact that Mercury has a pretty elliptical orbit ($e = 0.2056$), making the approximate equation slightly inaccurate. It does however illustrate the point. In the later chapter, about the Post-Newtonian Formalism, we will discuss a more accurate approximation.

## 8.2 The deflection of light by the Sun

Einstein completed his general theory during World War I, so there was not much change to quickly verify his prediction that the Sun will bend the light rays passing close to it's surface (called a 'Sun-grazing ray') by 1.75 arcseconds. Shortly after the war, in May 1919, British astronomer Sir Arthur Eddington succeeded to measure the deflection of light from stars during a total eclipse of the Sun, on an island off the coast of (then) Spanish Guinea.

This particular test was not very accurate, although Sir Arthur claimed a value of $1.60 \pm 0.31$ arcseconds. This was at the time close enough to Einstein's prediction, to show that there is more to light deflection than what any 'quasi-Newtonian' theory could come up with. If one assume the corpuscular theory for light, then those particles, traveling at the speed of light, would according to Newton's gravity, suffer a deflection of 0.875 arcseconds if they graze the surface of the Sun—precisely half of what Einstein predicted![Faber].

It was later discovered by theorists that any theory of gravity that is compatible with equivalence principle (equivalence of acceleration and gravity), will automatically predict a light deflection of 0.875 arcseconds for a Sun-grazing ray. The other 0.875 arcseconds comes from the curvature of space [Will(b)].

So what does the deflection formula look like? Starting with the orbital equation for light,

$$\frac{d^2u}{d\theta^2} = -u + 3\bar{M}u^2, \tag{8.3}$$

and using successive approximations, one obtains a rather simple formula for a small deflection of light in the relatively weak gravitational field at the surface of the Sun:*

> *The approximation does not hold for large deflections in strong gravitational fields. It is then better to solve the equation numerically.

$$\Delta\phi \approx \frac{4\bar{M}_{Sun}}{r_{Sun}}. \tag{8.4}$$

The radius of the Sun is about $6.96 \times 10^8$m, so that the deflection works out to [Faber]

$$\Delta\phi \approx \frac{4 \times 1477}{6.96 \times 10^8} \; \approx \; 8.49 \times 10^{-6} \text{ radians } \approx 1.75 \text{ arcseconds.}$$

This prediction has later (in the 1970s) been confirmed to the 1% level, using long baseline radio interferometry [Will(a)]. The radio waves were coming from quasars that, as the Earth orbits the Sun, pass close to the Sun once a year.

## 8.3 The Shapiro time delay of light

As far as we can ascertain, the speed of light in free space is one of the universal constants. It does not necessarily mean that the speed of light is the same everywhere—we know that it is slower in air than it is in space and also slower in glass than it is in air.

Gravity also has an effect on the speed of light in a certain way. That effect is not locally measurable though. If an inertial observer, momentarily stationary near a massive object, makes a speed of light measurement, the standard value of $c = 2.99792458 \times 10^8 \; m/s$ will be obtained. In geometric units the locally measured speed of light is unity of course. So what is the effect we are after?

### 8.3.1 The effect of gravitational time dilation

The closer a clock is to a gravity generating mass, the slower it runs. How does that affect the measurement of the speed of light? If distances remained the same for local and distant observers, then the local observer must get a speed of light larger than $c$. We have seen before that gravity contracts space by the same factor that it dilates time.

If we could observe a pulse of light from a large distance while it moves close to and transversely relative to a massive object, we would perceive that it has slowed down by a factor $\sqrt{g_{tt}}$, due to gravitational time dilation (or gravitational redshift, if you like). The best way to put it is that the local transverse speed of light transforms to the distant reference frame by a factor $\sqrt{g_{tt}}$, just like any other transverse velocity.

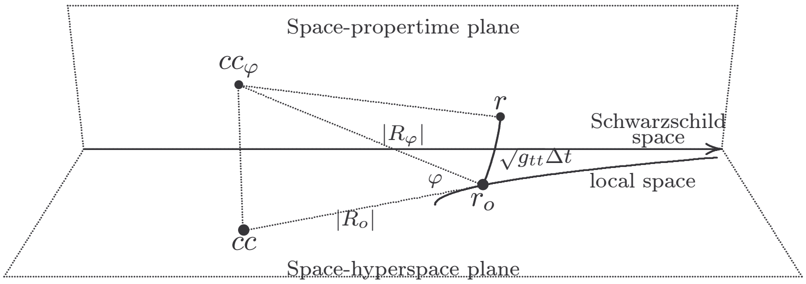


**Figure 8.1:** A repeat of figure 5.1 (geodesic movement) of the previous chapter. If $\varphi \to \pi/2$, i.e. the radial velocity $\dot{r} \to 1$, then $R_\varphi$ tends to infinity and the space-propertime path ($r_o - r$, length $\sqrt{g_{tt}}\Delta t$) of a radially moving object will tend to coincide with the local space curve. The projection of $r_o - r$ onto Schwarzschild space will then have a length $\sqrt{g_{tt}}\Delta t$, due to the 'angle' between local space and Schwarzschild space.

### 8.3.2 The effect of the curvature of space

When the movement of light is in a radial direction relative to the gravity generating source, then the slope of curved space also plays a role. There is then another factor $\sqrt{g_{tt}}$ involved, as pictured in figure 8.1. Curved space means that the light, already 'slowed down' by the slower running clock, also moves at an 'angle' to Schwarzschild space, so that the movement in the radial direction is further retarded by a factor $\sqrt{g_{tt}}$. The local *radial speed* of light transforms to the distant reference frame by a factor $g_{tt}$, just like any other radial velocity.

### 8.3.3 The time delay of light

One can say that as observed by a distant observer, the *transverse speed* of light is $\sqrt{g_{tt}} = \sqrt{1 - 2GM/r}\, c$ and the *radial speed* is $g_{tt} = (1-2GM/r)c$, where $r$ is the radial Schwarzschild coordinate distance from the centre of mass $M$. Here the terms transverse and radial mean relative to the mass.

It must be said that scientists do not like this sort of talk, i.e. that the speed of light 'slows down' in a gravitational field. As touched on before, the reason for this is that if a *locally* stationary inertial observer performs a localized measurement of the speed of light in the gravitational field, the result is always $c$. It is only when a stationary distant inertial observer measures the time that light takes to pass close to a gravitational source, that there is a real (not apparent) delay.

For a light beam that grazes the surface of the Sun, say between Earth and Mars at superior conjunction,* the photon path is almost purely radial relative to the Sun. The 'speed' of light can be taken as approximately $g_{tt} = 1 - 2\bar{M}/r$ for both legs of the trip (as a first approximation).

*Superior conjunction occurs when a planet is on the opposite side of the Sun, as viewed from Earth.

The time delay is defined as the round trip difference between 'Newtonian time', where $c = 1$ and 'relativistic time', where $c = 1 - 2\bar{M}/r$. Since $r$ changes continuously during the trip, the time delay must be obtained by integration of time over the distance. For a photon passing relatively close to the Sun, the following approximation holds [Will(c)]:

$$\Delta t_d \approx 240 - 20\ln(\frac{d^2}{r_p})\ \mu\text{s}, \tag{8.5}$$

where $d$ is the distance of closest approach to the Sun, expressed in solar radii, $r_p$ the distance of the planet from the Sun in astronomical units (AU) and In is the natural logarithm. The delay for the 42 minute round trip of a Sun-grazing photon to Mars at superior conjunction and back works out to about 250 $\mu$s.

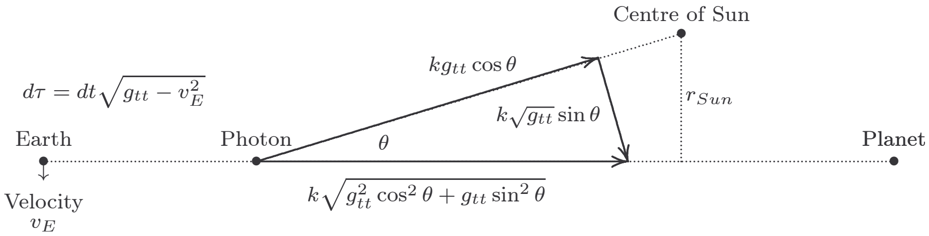

Centre of Sun
$kg_{tt}\cos\theta$
$d\tau = dt\sqrt{g_{tt} - v_E^2}$
$r_{Sun}$
$k\sqrt{g_{tt}}\sin\theta$
Earth
Photon
$\theta$
Planet
Velocity
$v_E$
$k\sqrt{g_{tt}^2\cos^2\theta + g_{tt}\sin^2\theta}$

**Figure 8.2:** The arrowed triangle represents the radial and transverse components and the resultant vector for a Sun-grazing photon's velocity (for one position of the photon). As used before, $g_{tt} = 1 - 2\bar{M}_{Sun}/r$, where $r$ is the radial distance between the photon and the centre of the Sun. The constant $k$ is just an arbitrary scale factor. Earth time is represented by $d\tau$ and Schwarzschild coordinate time by $dt$.

When corrected for the fact that there is a transverse component involved in the path of the photon and also for the fact that an observer on the surface of the Earth is not quite a 'stationary, distant inertial observer', the predicted time delay for a round trip to Mars is around 200 $\mu$s. Figure 8.2 illustrates the effects. The 'effective speed of light' in Schwarzschild coordinates can be closely approximated by

$$c_S \cong c\sqrt{g_{tt}^2 \cos^2\theta + g_{tt}\sin^2\theta}, \tag{8.6}$$

in whatever units are chosen for $c$. Here $\theta$ is the coordinate angle between the photon path and the radial from the mass $\bar{M}$. The speed of a photon is slightly faster in transverse directions than in radial directions and therefore decreases the predicted time delay. Further, the total time dilation at Earth's average orbital distance from the Sun ($r_E$) and average orbital speed ($v_E$), is

$$\frac{d\tau}{dt} \cong \sqrt{1 - \frac{2\bar{M}_{Sun}}{r_E} - v_E^2}\ , \tag{8.7}$$

where $d\tau$ represents the rate of clocks on Earth and $dt$ the rate of the distant* reference clock. The slight ellipticity of Earth's orbit is ignored.

*Here a distant clock is one at rest relative to the Sun and far enough away to be considered as measuring Schwarzschild time.

The fact that clocks on Earth run slightly slower than distant clocks causes the measured time delay to be smaller than what it would have been if Earth clocks ran at 'distant (Newtonian) rates'. The Earth's mass and rotation and the mass of the 'target' planet also play a role, but the effects are much smaller than the experimental errors and are usually ignored. Likewise, the fact that the photon's path will be deflected slightly by the Sun, contributes negligibly to the time delay.

The first predictions for the delay during superior conjunction and the actual measurements thereof were performed by Irvin Shapiro et. al. in the early 1960s. They bounced radar signals off the planets Venus and Mercury when they were close to superior conjunction. This type of time delay has since then been called the **Shapiro time delay**.

During the late 1960s and early 1970s, Shapiro bounced radar signals off Mars near superior conjunction. He later also utilized one of the Viking Mars landers, fitted with appropriate transponders. This eliminated the uncertainty of where the radar signal bounced of the Martian surface and allowed Shapiro and his team to reduce the observational errors to smaller than 0.1% of the effect they were after.

The results were in almost perfect agreement with the predictions of general relativity. An excellent description of this delicate experiment can be found in [Will(b)]. Will also offers an up to date technical web site, [Will(c)], which is well worth a visit.*

*You might want to read chapter 11 before visiting it, due to the rather abbreviated, yet technical nature of the site.

## 8.4 Gravity Probe B

There are two other effects of general relativity that are potentially testable in the solar system: the dragging of inertial frames and the geodetic effect. The purpose of Gravity Probe B was to test both in the vicinity of Earth. But first, what are they?

### 8.4.1 Geodetic Effect

The geodetic effect is caused directly by curved spacetime. It is somewhat similar to parallel movement* on curved three-dimensional surfaces.

> *Relativists call it parallel transport, meaning never rotating relative to the local surface.

Consider an arrow pointing due north on Earth's equator at zero longitude. Now parallel transport it to the north pole along the zero longitude meridian. From there, parallel transport it down the 90 degrees meridian to the equator.

The arrow will now point due east, while it never rotated relative to the (two dimensional) surface. This is similar to the geodetic effect in curved four dimensional spacetime.

Gravity Probe B does a similar thing by parallel transporting a gyroscope through the curved spacetime surrounding Earth. During each orbit of the spacecraft, the spin axis of a very stable on-board gyro is predicted to tilt by a tiny amount. Do that for a year or more and the tilt becomes measurable.

### 8.4.2 Dragging of Inertial Frames

The dragging of inertial frames happens around a spinning mass. General relativity predicts that a spinning mass drags spacetime around itself, almost like a spinning ball in a fluid will drag the fluid around with it. Particles submerged in the fluid will then be dragged with the fluid.

Compact, spinning bodies like neutron stars and black holes are predicted to drag spacetime around by a significant amount. The Earth's rotation is predicted to drag the spacetime around by a very small amount, but over a long period it might become measurable.

This is what Gravity Probe B attempts to achieve. For a spacecraft in orbit over the poles, the frame dragging will tilt a gyro in a different direction than what the geodetic effect will do. The two effects can thus be measured with the same gyroscope. For redundancy, the science package included four identical, very precise gyros.

At the time of writing, Gravity Probe B has been up there for more than a year, of which just about one full year was a "science run". The data analysis will take another year and the results are expected early in 2007.

## 8.5 The spacecraft

At a 650km high orbit, the satellite completed more than 5000 orbits during the year-long science run. The expected results are: a geodetic shift of 6. 6 arc-sec per annum and a frame dragging shift 0.042 arc-sec per annum in the spin axis of the gyros.

It is clear from the expected results that the four gyros had to be of extreme accuracy - a precision of 0.0005 arcseconds was the target. How did they do it? This quote from one of NASA's websites [GP-B] says a lot:

"**April 26, 2004:** Engineers don't often indulge in poetic flourish when discussing the things they build. So when words like 'beautiful' and 'elegant' and 'artful' frequently cross the lips of scientists and engineers as they talk about the design of Gravity Probe B (GP-B), one might suspect that this spacecraft is truly something special.

"Telescopes. Gyroscopes. Superconducting lead bags and SQUIDs. These are odd materials for art. Among engineers and physicists, though, there's no doubt: Gravity Probe B is a masterpiece." Read more about this "technological *tour de force*" on NASA's website [GP-B].

It will be very interesting to see how close the actual results come to the predictions of general relativity.

# Chapter 9

# Tidal Gravity

that tends to
stretch and
squeeze

Tidal gravity is most commonly known as the cause of the tides of the oceans on Earth. The Sun and the Moon both tugs on the Earth and cause a bulge on the surface of the Earth in some places and a flattening at other places. How the tides work will be discussed later, but first, what is tidal gravity?

Loosely speaking, tidal gravity manifests itself as unequal gravitational accelerations on different parts of a free-falling body, caused by the different distances and angles of the body parts from the centre of the source of gravity. We will first look at tidal gravity from the perspective of Newton's theory and then discuss the general relativistic interpretation of tidal gravity.

## 9.1 Newtonian tidal gravity

Consider what happens if an object with a diameter $d$ free-falls straight towards a large spherical mass $M$. Let the centre of the falling object be a radial distance $r$ from the centre of $M$.

The object's side nearest to the mass experiences a gravitational attraction that is larger than the gravitational attraction experienced by the centre of the object. Likewise the centre of the object experiences a gravitational attraction that is larger than the gravitational attraction experienced by the far side of the object, causing a stretching effect in the radial direction.

The stretching acceleration of the near side of the falling object relative to it's centre is (in geometric units)

$$\frac{-GM}{(r-d/2)^2} - \frac{-GM}{r^2} \approx \frac{-2GMd}{r^3}, \text{ if } r \gg d. \tag{9.1}$$

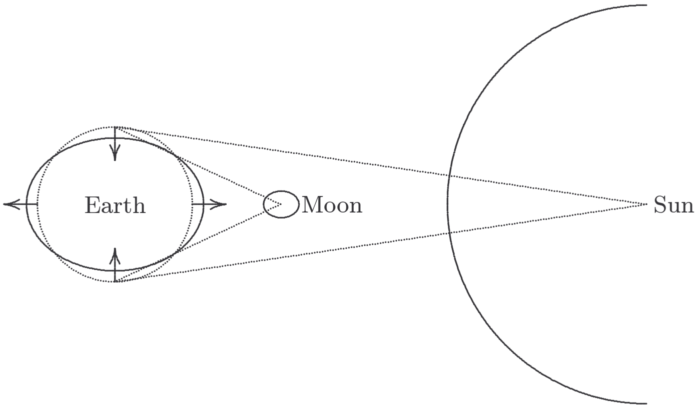


**Figure 9.1:** Tidal forces from the Sun and the Moon tend to stretch and squeeze the free falling Earth. The Moon's crust suffers the same effects. The magnitudes of the stretching and squeezing are relatively small and contrary to popular believe, they are not the direct cause of the tides in the oceans of the Earth. See text.

Likewise, the stretching acceleration of the far side relative to the centre is

$$\frac{GM}{(r+d/2)^2} - \frac{-GM}{r^2} \approx \frac{2GMd}{r^3}, \text{ if } r \gg d. \tag{9.2}$$

Because the gravitational acceleration works towards the centre of the massive object, there is also an active flattening effect in the transverse direction, as viewed from mass M (figure 9.1). The magnitude of the 'squeeze' effect is very closely the same as the stretching effect.

It is clear that because of the cubed distance denominator, tidal gravity falls off very rapidly with distance. This is significant in the case of the tides in Earth's oceans. Although the Sun has a much larger gravitational attraction on Earth than the Moon has on the Earth, the Moon produces the larger tidal effect on Earth, simply because it is so much closer to us.

The tidal stretching produced on Earth by the Moon and the Sun can be viewed as four acceleration vectors, i.e. a near side and a far side vector for the Moon and likewise, two for the Sun. The four vectors rotate around Earth (relative to it's surface) at different rates and with changing orientations and magnitudes.

To simplify things, we will analyse the tides with the Sun and the Moon roughly lined up (as at the new moon) and use the direction of their combined gravitational force as a reference, which we will refer to as the *tidal vector*.

The fact that the Earth is in orbit around the Sun and not free falling directly towards it, does not change the situation in any significant way. Any orbit is just a free-fall with some transverse velocity added.

This is all very easy as far as the Sun is concerned, but we know that the Moon is in fact the major contributor to tidal gravity on Earth. It is not so intuitively clear that the Earth is also free-falling relative to the Moon.

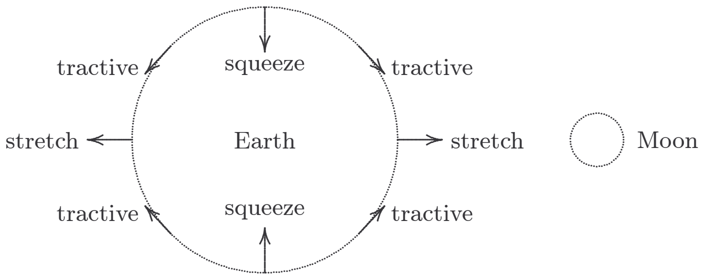


**Figure 9.2:** At points between the forces of pure stretch and pure squeeze, both stretching and squeezing are at work. This results in "tractive" force components working along the surface of the Earth, capable of dragging water horizontally.

Look at it like this: park the Earth and the Moon at an arbitrary distance from each other in some inertial reference frame and then set them free. Each will start to accelerate towards the other, as measured in the inertial reference frame.

While the Earth is orbiting the Sun, it is continously falling towards the Moon as well. More correctly stated, the Earth and the Moon orbit each other around their common centre of gravity (the barycentre), while the common centre of gravity orbits the Sun. Also see the box on page 138 for an interesting aside on the moon's orbit. Now back to tidal gravity and specifically the effect on the oceans of the Earth.

The tidal stretching and compression of the Earth discussed above, is often put forward as the direct cause of the tides in our oceans, but it is quite misleading, as will soon become clear. When the Sun and the Moon are lined up perfectly, the magnitude of the tidal acceleration is only about 175 nano-g on each side of the Earth.*

*175 nano-g is 175 thousand-millionth's of the normal Earth gravity, or $1.7 \times 10^{-6}\ m/s^2$.

This does stretch the Earth's radius by some tens of centimeters, but it does not cause a rise in the sea level relative to the land—the sea level and the adjacent land directly in line with the tidal vector 'rises' by approximately the same amount. However, at points on the earth's surface where both the compression and stretching effects are operating, there are components working along the surface of the Earth, as shown in figure 9.2. See, e.g., reference [Tides].

The maximum horizontal component is about 87 nano-g and occurs at points ±45 degrees (and ±135 degrees) from the tidal vector. If the Earth was not rotating relative to the tidal vector, this 'tractive' acceleration would eventually cause some water to flow towards points directly in line with the

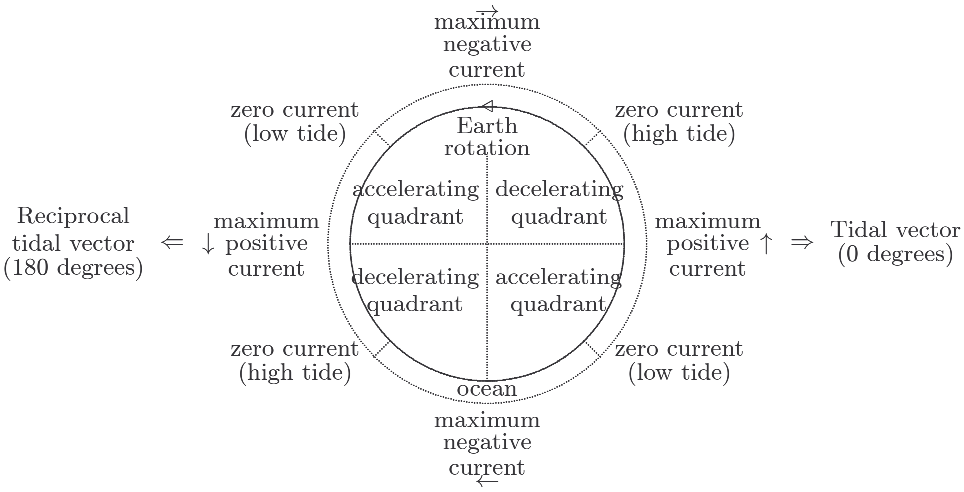


**Figure 9.3:** The formation of tidal currents along the equator of a hypothetical uniform, continent free ocean, as viewed from above the north pole. Tractive tidal forces alternately accelerate and decelerate water in the four quadrants, as shown. Positive currents are taken as water rotating faster than Earth's crust, flowing from west to east. The contiguous equatorial currents are ignored.

tidal vector.

The water would have heaped up a bit and at the same time caused a withdrawal of water at points 90 degrees from the tidal vector. The normal 1g gravity towards the centre of the Earth will tend to level the water and a stable equilibrium state would have been reached, with no movement of the water.

The Earth is however rotating at about 14.5 degrees per hour relative to the tidal vector and drags the water of the oceans with it. One can say that the water at the equator moves at an average speed of over 1600 km/h relative to the tidal vector.

This movement prevents a stable, equilibrium condition to be reached. As shown in figure 9.3, for the quadrants directly to the west of the tidal vector line, the tractive force causes a small acceleration of the water relative to the tidal vector, and in the eastwards quadrants, an equal deceleration.

As the Earth rotates relative to the tidal vector, each molecule of water in the open ocean along the equator will experience about 6 hours of this tiny acceleration of 87 nano-g maximum or 56 nano-g average, followed by 6 hours of equal deceleration.

The maximum positive flow, relative to the crust will occur near the two positions in line with the tidal vector. Likewise, the maximum negative flow will occur near the two points 90 degrees to the east and west of the tidal vector.

In the open oceans near the equator, it creates what is called a reversing tidal current. At other latitudes, the Coriolis force will tend to convert a reversing current into a rotating tidal current. The relative phases of the

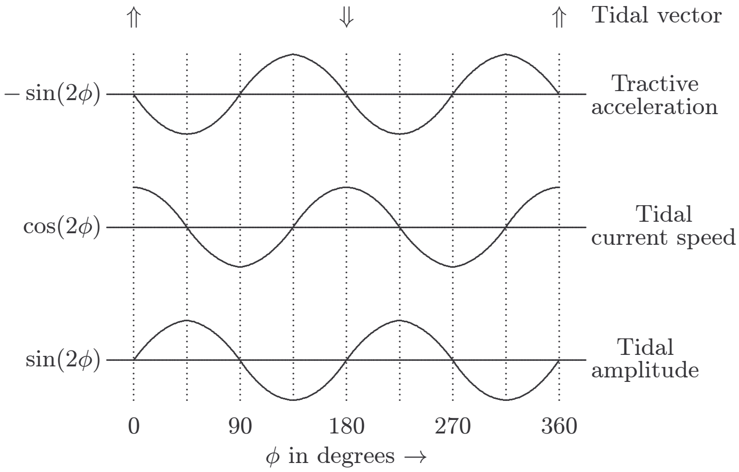


**Figure 9.4:** The phase relationships between the tractive acceleration, tidal current speed and tidal amplitude. Note that the peak tidal amplitude lags the tidal vector by 45 degrees and is precisely out of phase with the tractive acceleration. The amplitudes of the 'waves' are not to scale and serve only to show the phase relationships.

tidal currents are shown in figure 9.4.

To understand the phase relationships, consider water flowing through a uniform duct. If the in-flowing speed of water into any section of the duct is larger than the out-flowing speed from that section, the water level in the section must rise. This situation arises when the water is being decelerated over the length of the section.

The approximate speed of the tidal flow in open ocean is obtained by integrating the tractive acceleration, which has the general form $a_{tr} = -k\sin(2\phi)$, where $\phi$ is the angle measured from the tidal vector and $k$ some positive constant. The velocity of the tidal current is then

$$v_{tc} = -k\int \sin(2\phi)\; d\phi = \cos(2\phi) + \text{ constant.} \tag{9.3}$$

As stated above, the amount of heaping of the water (the tidal amplitude) is proportional to the velocity differential between inflow and outflow of a section. The differential of $\cos(2\phi)$ is $\sin(2\phi)$, which is proportional to $-a_{tr}$, so the tidal amplitude is precisely out of phase with the tractive acceleration.

In the open (and deep) ocean, the reversing or rotating tidal currents will cause only minor high and low tides, calculated to be in the order of a few tens of centimeters. The speed of the tidal currents is extremely small.

An acceleration of 56 nano-g applied to a free moving object for 6 hours will result in a speed change of only 42 metres per hour. The reversing tidal currents in the open ocean will therefore achieve maximum speeds of less than 20 metres per hour in each direction. *So how on earth do these slow moving currents cause the coastal tides?*

The secret lurks (partially) in the considerable depth of the oceans, averaging at more than 3 km. Since the tractive tidal acceleration works about equally at all depths, the volume of slow moving water is very large and so will be the kinetic energy locked up in the movement.

If the tidal current would hit a large continent head-on, then most of that kinetic energy will be converted into potential energy by a temporary rise in the level of the coastal waters. Depending upon the topography of the coast, the heaping of water will in general tend to increase for as long as there is a nett inflow of water towards the coast. The topography may also cause the in-flowing current to be considerably larger than the mean current in the deep sea.

When the main tidal current starts to reverse under the influence of the tractive force, the elevated coastal waters will flow out to the deep sea again, where the water level is lower. This outflow is mainly driven by normal Earth gravity that tends to level the water again, assisted just a little bit by the tractive tidal force, which is near maximum at the time of high water. The water flowing away from the coastline will have some momentum and a small overshoot is possible, which added to the tractive force, causes the low tide.

When the normal Earth gravity and the tractive force are moving water in the same direction, we have the possibility of a resonating system. Whenever the natural resonance* period of the mass of water between the continents (or shallower sections) happens to be close to the period of the periodically reversing (or rotating) tidal current, we can have a sustained oscillation that causes higher than average tidal currents and with that, tidal amplitudes of a few meters above mean sea level along the coastline.

The main period of the tidal currents will average at 12.42 hours, the so called semi-diurnal period, meaning roughly twice a day. This period is of the same order of magnitude as the resonant periods possible in the many parts of the oceans of the world. We can do a rough check by using an ideal square basin as a guide. Taking a water depth of roughly the average depth of the oceans, some 3500 metres, we get a resonant period of 12 hours for a basin of 4000 km long.

The resonating areas are not limited to only the basins formed by the continents, but may also be the result of basins formed by very deep water broken up by oceanic ridges. In essence, a whole system of resonant currents can build up in the open oceans. If the natural resonant period of a basin is not near one of the main tidal periods, there will still be times when the tractive currents tend to coincide with the resonant currents, causing larger than average currents.

There will then also be times when the two currents oppose each other and smaller than average currents will result. The real basins are obviously very irregular and the resonant current systems will also overlap in places,

*The resonance period of an ideal rectangular water basin is $T = 2L/\sqrt{gd}$, where $L$ is the length of the basin, $g$ the local gravitational acceleration and $d$ the depth of the water.

making a complete analytical treatment almost impossible.

However, for any specific locality, the complex tidal current system will be periodic in nature and the commonly used practical approach is a quasi-empirical one, known as *harmonic analysis and prediction*. Up to 37 harmonic constituents are added together, each with it's own frequency, phase and amplitude.

The frequency, phase and amplitude values for each constituent at a specific location are obtained by analyzing observed values for current speeds and tidal heights over extended periods of time. The 37 constituents include all astronomical effects, like non-circular orbits of the Moon and the Earth, short and long term changes in declination and also the so called *shallow water effects* for the specific place on Earth.

One problem in obtaining the harmonic constituents is that the full astronomical cycle of the Moon relative to the Earth lasts for 18.6 years. This is the time that elapses before the Moon crosses the equator again at precisely the same angle and place. Tidal specialists like to have a sample of at least that long.

Once they have extracted all the constituents for a specific place, remarkably accurate predictions of current strengths, tidal heights and times can be made for many years into the future. There are however non-cyclic effects caused by high winds, low and high pressure systems and other effects of the global weather that have to be taken into account on a sporadic basis.

The simplistic theory of an 'Earth circulating wave' of water in the oceans cannot be used to directly predict the coastal tides at any location. In principle, it is not even a correct interpretation of the cause of the coastal tides.

It may be argued that the small heaping effect that lags the tidal vector by some 45 degrees does contribute to the coastal tides. This heaping can however only occur to any significant degree in the relatively deep water of the open oceans. At the coast, it is the tidal currents that contributes significantly.

Another simplistic theory which is suspect, is that the Moon's drifting away from Earth at around 3.8 cm per year, is due to the friction caused by the tides in the ocean. As we have seen, the tidal currents are reversing or rotating and friction effects will largely cancel out—there are roughly an equal amount of tidal currents hitting continents in easterly and westerly directions.

The more reasonable theory for the Moon's orbital distance to increase is related to the tidal stretch and squeeze of the Earth as a whole, causing the crust to bulge out and be squeezed inwards. This effect is circulating continuously from east to west as the Earth rotates, causing some loss of angular momentum of Earth's rotation. This loss is transferred to the Moon in the form of additional orbital momentum, because the total angular momentum of the Earth-Moon system must be maintained.

The additional orbital momentum causes the Moon to continuously take up an ever so slightly larger orbit. The physical mechanism of transfer of angular momentum from the Earth to the Moon's orbit stems from the fact that there is a slight angular lag in the bulges on Earth's surface relative to the rotating tidal vector (as viewed from Earth). This is because it takes some time for the bulges to form and to subside again.

From the Moon's perspective however, the bulges have a 'lead angle', causing the gravitational force to point slightly ahead of the common centre of gravity of the Earth-Moon system. This offset creates a small force component in the direction of the Moon's orbit, causing it to continuously pick up a tiny amount of additional orbital velocity.

The Moon also suffers tidal forces due to Earth's gravity, and they are in the order of 22 times* larger than the Moon's tidal effect on the Earth. This caused large frictions in the originally faster spinning Moon and is the reason why the Moon now always shows more or less the same face to us. The Moon's rotation period has been decreased by the friction until it equaled the average orbital period of the Moon around the Earth.

Earth's rotational speed has also been decreased by tidal gravity. It causes in the order of 1.7ms in period shift per century. Stated differently, if atomic clocks and mean solar time were running precisely in step 100 years ago, then today our atomic clocks would run fast by 1.7ms every day.

This tiny amount has accumulated over time and today our atomic clocks tend to get ahead of mean solar time by about a second over an eighteen month period, on average. This is the main reason for the frequent "leap-seconds" that we experience, where we effectively set our clocks back by one second.

There are other random (unpredictable) factors caused by crustal and inner movements in the structure of Earth, that also affect Earth's rotation period. The 'scary' part of these factors is that we cannot, even in principle, predict future solar time precisely!

## 9.2 Relativistic tidal gravity

Relativity views tidal gravity as the result of curved spacetime. If spacetime is curved between two points, geodesics in spacetime either converges or diverges, depending on the direction of curvature.

If we take four free particles as representing a free falling object (figure 9.5), then the particle nearest to the mass has a time arrow that curves more sharply towards the mass than does the time arrow of the particle furthest from the mass, so the two particles diverge. In fact they will accelerate away from each other. The two particles perpendicular to the radial have time

*The Earth is about 81 times more massive than the Moon, but then the Moon's diameter is about 27% of Earth's and the tidal force on the Moon is proportional to $M_{Eearth} \times d_{Moon}$.

arrows with the same curvature, but because they are both curving towards the centre of the spherical mass, they converge, accelerating towards each other.

If we would place the particles at the points at 45 degrees to the radial, then the furthest pair will converge, and so will the closer pair. But the furthest pair would diverge from the closer pair. These relative geodesic movements correspond to the Newtonian tractive force along the surface of a spherical body.

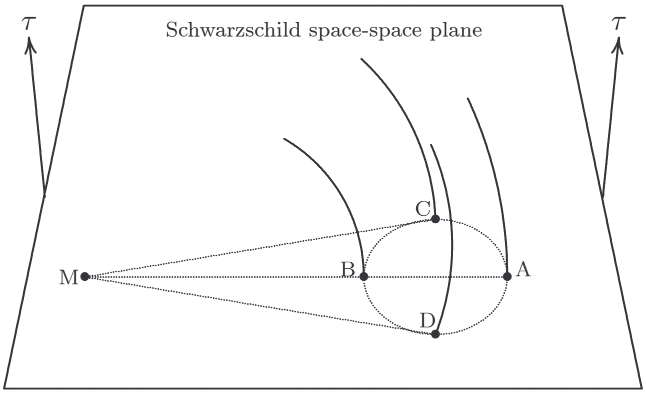


**Figure 9.5:** Spacetime geodesics as the mechanism of tidal gravity effects. The four geodesics rise above the space-space plane, into the space-propertime ($\mathbf{x}\mathcal{T}$) domain, taking particles A and B further from each other, while C and D move closer to each other in space.

The surface of a solid object does not follow the geodesic paths of the free particles, because it is prevented from doing so by the molecular binding forces holding the object together. And when the object is self-gravitating, it's gravitational acceleration also holds the object together.

In general relativity, tidal gravity is not seen as a 'real' force trying to squeeze or stretch the solid, free falling body, but rather that the natural geodesic movement of parts of the body is restrained by the molecular or self-gravitation forces.

It is similar to the relativistic view that when you stand on the surface of our Earth, the only force you experience is the surface pushing upwards at your feet and thus preventing you from following a geodesic through spacetime.

In the relatively small curvature environment inside the solar system, or in fact in most places in the universe, the results obtained for tidal gravity by means of general relativity will be virtually indistinguishable from that of Newton mechanics.

However, near objects like neutron stars and black holes, the results will be very different, because in such localities general relativity predicts gravitational accelerations much stronger than what Newton's theory predicts. The same holds for tidal gravity. However, one must be pretty close to such objects for the difference to become significant.

To illustrate this, let us replace our Moon abruptly by a black hole with the 'modest' mass of one and a half times that of our Sun. Apart from the fact that the Earth will rather quickly be 'swallowed' by the black hole, there is an even more serious effect.

The stretching components of the tidal gravity caused by such a huge mass so close to us will be about 5 times the normal surface gravity of Earth, but in the opposite direction. It means that on the sides nearest and furthest from the black hole, there will be an upward acceleration of 4g instead of the normal downward acceleration of 1g. Everything, including the Earth's crust, will start to fly away from the centre of the Earth.

On the sides 90 degrees from these points, there will be a downward tidal acceleration of $5+1 = 6$g; a really big squeeze! Earth will essentially be converted into a stream of rubble that will be swallowed by the black hole in due course. The transverse velocity of Earth relative to the hole would be too small to allow the fragments to enter a stable orbit around it.

Frightening and 'relativistic' as this fanciful scenario may sound, surprisingly, it is not a very relativistic case! Normal old Newtonian gravity does the demolition job of Earth equally well. Relativistic tidal effects only become very significant when the distance from the black hole is within a hundred times the radius of the hole's event horizon.

Our one and a half solar mass black hole has an event horizon radius of about 4.4 km, while our mean distance to the Moon is about 384,000 km. In this case, the relativistic effect would increase the Newton tidal gravity by only 6 parts in a million.*

However, as the pieces of what previously was Earth get closer to the black hole, the relativistic effects of spacetime curvature will become much more significant. There is no simple approximation for the tidal acceleration at very close ranges to a black hole.

The easiest is to calculate the relativistic gravitational acceleration vector at some point on the free-falling object and then to find the difference between this vector and the equivalent vector of the centre of mass of the free-falling object. This difference vector gives the tidal acceleration vector. We will analyze such a case in the next section.

## 9.3 Tidal gravity near a black hole

Let us plan for one of our astronauts, Pam, to hover her spaceship just outside a black hole's event horizon. Let the centre of gravity of the spaceship be at a circumferential radius* $r = 2.1GM/c^2$. The local radial gravitational acceleration that has to be countered by the engines of the spaceship is

*The relativistic factor is $1/\sqrt{1-4.4/384000} \approx 1.000006$.

*This means at a distance where a measured circumference is 2.1 times the circumference of the event horizon of the hole.

given by

$$\ddot{r} = \sqrt{g_{rr}}\frac{-GM}{r^2} \tag{9.4}$$

Let us revert to the geometric mass parameter $\bar{M} = GM/c^2$ again for simplicity and calculate the tidal gravity real close to the black hole, say at $r = 2.1\bar{M}$. At this distance $\sqrt{g_{rr}} = 1/\sqrt{1 - 2/2.1} = 4.80$, and

$$\ddot{r} = 4.8\frac{-\bar{M}}{(2.1\bar{M})^2} \approx \frac{-1.039}{\bar{M}}.$$

This shows that, for a constant fractional distance outside the event horizon, the gravitational acceleration is *inversely proportional to the mass of the black hole.*

To study the tidal forces, let us put a number to the mass of the hole, say 1,000m geometric mass, so that the ship's centre of gravity is at a distance of $r = 2.1 \times 1000 = 2100$m circumferential radius from the hole. The acceleration at the ship's centre of gravity will be $\ddot{r} = -0.001039$ m$^{-1}$.

Let the ship have a radius of 10m, so that the 'bottom' of the ship is at $r =$ 2090m from the hole. The acceleration there will be near enough -0.001103 m$^{-1}$, giving a 'stretching' tidal acceleration of $-0.001103 - (-0.001039) = -6.4 \times 10^{-5}$ m$^{-1}$ (at the top side of the ship, a positive stretching acceleration of the same magnitude will be felt).

Now increase the hole's mass tenfold, to 10,000m geometric. The same calculations yield the tidal acceleration to be $0.0001045 - 0.0001039 = 6 \times 10^{-7}$ m$^{-1}$. The sums show that the acceleration has decreased to about one tenth, as expected, while the tidal acceleration has decreased to about one hundredth of what it was.

As a rule of thumb one can say that for a constant fractional distance outside of the event horizon, and for a constant test distance (10m in this case), tidal acceleration is roughly inversely proportional to *the square of the mass* of the black hole.

The above accelerations look innocent enough, but recall that to convert m$^{-1}$ to m/s$^2$, we need to multiply by $c^2$. Even at $\bar{M}$ =10,000m, this makes the actual acceleration about $10^{11}$g and the tidal acceleration about $5 \times 10^9$g. Ouch! Neither spaceship, nor Pam can possibly survive that.

In order to find a survivable environment of say less than 10g (at $r = 2.1\bar{M}$), the black hole's geometric mass must be around $10^{15}$m, which is about a trillion ($10^{12}$) solar masses. With a $10^{15}$m black hole, the ship will be positioned $2.1 \times 10^{15}$m from the centre, so the tidal acceleration will be utterly negligible over the 10m size of the spaceship. This is due to the 'inverse square law' for tidal acceleration against mass for a constant fractional distance outside the event horizon.

In order to study tidal gravity in a 'friendlier' environment, Pam should put her spaceship into orbit around a 10,000m mass black hole. Since an orbit means free fall, the direct gravitational acceleration is not detectable, just the tidal gravity is. Now orbits at $r = 2.1\bar{M}$ are not possible, of course, and if they were, tidal gravity would still have pulled the spaceship (and Pam) apart.

Let us estimate the circumferential radius of the orbit that will produce, say, 5g of tidal acceleration over a 10 metre radial distance (the radius of the spaceship). Since we will presumably move out of the 'super gravity' environment, one can suspect that Newton gravity would give a reasonably good answer.

We use the approximation:

$$\Delta\ddot{r} = 2\bar{M}\ d/r^3,$$

which, for 5g (or $5.46 \times 10^{-16}$m$^{-1}$) of tidal acceleration over 10m radial distance, gives an orbital radius $r_o = 714\bar{M}$. A spreadsheet check with the full relativistic equations, confirms after a few iterations that the correct answer is closer to $r_o = 716\bar{M}$ (not bad for a Newtonian approximation).

For our 10,000m (or 10km) mass black hole, it means that an orbit of 7,160km circumferential radius is required. So we ask Pam to put her spaceship in this safe orbit around the black hole.

Conditions on board Pam's orbiting spacecraft will be very interesting, if not weird. If Pam positions herself exactly at the centre of gravity of the spaceship, she can relatively easily stay there, apart from the fact that she will be slightly stretched in a radial direction and slightly squeezed in transverse directions. We can say 'slightly', because over Pam's length of under 2m, the tidal forces are under 1g.

If Pam moves a little bit off the centre of the ship (radially towards the black hole), she will have to fight a tidal force that tends to pull her to the 'floor' of the ship. If she allows the force to move her to the floor, she can stand there, but she will experience a rather uncomfortable gravity of near 5g.

If she moves to the 'ceiling' of the ship, she can also stand there, 'upside down', with the same gravity of near 5g, pushing her against the 'ceiling'. If Pam succeeds to move to any of the 'side walls' of the ship, at the same orbital radius as the centre, she will find it very difficult to stay there—she will tend to fall towards the centre of the ship with a starting acceleration of 5g.

In order to stay near that side, she will have to hang on and support nearly five time her own weight in the process, which is very unlikely. If she cannot hang on, Pam will initially start to oscillate between the two 'side walls' of the ship, but small perturbations will cause her to end up either at the 'floor' or the 'ceiling' of the ship.

This is all caused by the tidal forces that tend to stretch the ship (and Pam) radially and squeeze them transversally. Relativistically speaking, these weird effects come from the fact that the whole ship follows the spacetime geodesic of the ship's centre of gravity. Points off the centre of gravity 'attempt' to follow different spacetime geodesics, but they are prevented from doing so by the rigidity of the ship.

The ship is thus stretched and squeezed, as far as the elasticity (and strength) of the structure allows. Pam, being more or less independent of the structure, follows her own geodesic, unless she hangs on, or are supported by the floor or the ceiling. Hence the accelerations (relative to the ship) that she experiences when she ventures off the ship's centre of gravity.

**Micro-gravity** The 7,160km orbital radius of Pam around the black hole is not much different from the radii of many Earth satellites—it is the radius of an Earth satellite at about 780km above the surface of the Earth. So let us put our other astronaut, Jim, in a 7,160km radius orbit around Earth, and find out what he would experience.

The effects will be qualitatively the same as what Pam experienced, but just many orders of magnitude smaller. For Jim, the approximation:

$$\Delta\ddot{r} = 2\bar{M}\ d/r^3,$$

will obviously be a good one. $\bar{M}_{Earth} = 4.43 \times 10^{-3}$m, so that the tidal acceleration over 10m is

$$\Delta\ddot{r} = 2 \times 4.43 \times 10^{-3} \times 10/(7,150,000)^3 \approx 2.4 \times 10^{-22}\ \text{m}^{-1}.$$

Multiply this by $c^2/9.81$, giving $2.2 \times 10^{-6}$g, or $2.2\mu$g. All equipment that is bolted down inside Jim's spaceship, will experience a gravitational 'force' of $0.22d\ \mu$g, in some direction or other, where $d$ is the distance from the centre of gravity of the spaceship in metres. This is called a 'micro-gravity environment', where many exiting experiments are being carried out presently.

It is not precisely zero gravity, as one were lead to believe by the popular press, except perhaps for one tiny point at the centre of gravity of the ship. Jim will also tend to very slowly float to the floor or the ceiling of his ship. In practice, because of the tiny tidal acceleration, he is more likely to be influenced by the airflow of the ventilation system of his ship.

Tidal gravity, rather than the lack of gravity, can also be used for some practical purposes. It has been used to stabilize satellites, in the sense that the satellite always shows the same face to Earth, without having to use thrusters at regular intervals. After initially stabilizing the satellite with thrusters, a long, pendulum type construction, 'hung' from the satellite, tends to keep the attachment point of the pendulum pointing at Earth, due to the tidal forces.

### Is the Moon a Planet or what?

Our Moon is the only natural satellite in the solar system that never falls away from the Sun. Because of it's relatively large distance from Earth, the Moon follows an orbit around the Sun just like the Earth, while the Earth and the Moon are orbiting around each other, somewhat like a double planet system.

The Moon never falls away from the Sun because when the Moon is on the Sun's side of Earth (during the new Moon), it is further from Earth than the point where the gravity of the Earth and the Sun cancel each other out. The nett gravitational acceleration acting on the Moon is towards the Sun, meaning the Moon's orbit must be curving towards the Sun.

From an Earth perspective, it appears that the Moon is curving towards Earth at that point simply because the Earth is then curving towards the Sun more tightly than the Moon does, as in double planet systems. So the Moon probably deserves the title *sister planet of Earth*, rather than natural satellite of Earth.

In his delightful little book 'Asimov on Astronomy', in a chapter named 'Just Mooning Around', Isaac Asimov speculated about the reasons why the Moon displays this planet-like behavior, while none of the other planets' moons do. He reckons that Earth might be close to the distance from the Sun where double planets could form.

Asimov then speculates about how it would have been if the Moon had a size comparable to Earth—a real sister planet—with oceans, an atmosphere and possibly life. Presumably because he knows human nature and our history of conflict, Asimov closed the chapter with:

*"What a shame if we have missed that...*
*Or, maybe (who knows), what luck!"*

### Where has the water gone to?

One morning at about 10 o'clock local time, sitting on the beach near the southern tip of Africa during the low cycle of a new moon spring tide, watching the bare rocks where water has been some hours ago, someone asked: where has all the water gone to? My 'shooting from the hip' answer was: it must have gone to where it is now high tide.

That evening I suddenly realized how ridiculous my answer was. When it is low tide in South Africa, it must be high tide near Australia and also near South America - some 9000 km either way. Since high and low tides are separated by just over 6 hours in time, it is totally impossible for the water to have gone to any of those places!

The more reasonable answer is that the water has gone out to sea, because during the high tide, the water level was higher at the coast than in the deep sea and normal gravity will tend to level the coastal waters with the deep sea. The water flowing towards the deep sea has some inertia and it overshoots the level water situation somewhat, causing the coastal low tide.

The very water molecules that previously covered those rocks have probably not gone very far. Most of them were likely to be somewhere just beyond the bare rocks, taking the place of other molecules that have also moved just a little bit further offshore, and so on.

In general, during the period preceding the low tide, there is some withdrawal of water from the South African south coast, in a generally westerly direction, caused by gravity and the main tidal current. At the time of the low of a spring tide, the main tidal current is just about reversing and will start flowing from the south-west again soon.

This will bring the high tide first to Cape Town and the south-western coast and a little later to the south-eastern coast. The tides reach Durban, which is 12.5 degrees east of Cape Town, on average half an hour after they reach Cape Town.

This fact is quite surprising at first, because one tends to expect the tidal highs and lows to come from the east - it is after all where the Moon "comes" from. Goes to show that the tides have more to do with tidal currents than with the bulge that circles the Earth in step with the Moon.

# Chapter 10

# Gravitational Waves

The 'hum, chirp and squeak'
of
our Cosmos

Gravitational waves are part of the "hot stuff" of relativity and cosmology today. It was also 'hot' for Einstein himself, way back in 1916, when he published a paper showing that his general theory of relativity predicts the existence of gravitational waves. For decades, many people considered them as of no practical value because they are so weak - just how weak we shall see later in this chapter.

There was even disagreement as to whether they were not perhaps just a quirk in the mathematics, with no real physical meaning. It was only in 1960, with the so called 'relativity revival', that other theorists provided rigorous proof (given that general relativity is right) that gravitational radiation must in fact be a physically observable phenomenon.

It is perhaps the fact that by the start of the twenty first century, gravitational waves were not yet directly observed, that makes them the hot topic that they are today.

## 10.1 Tidal gravity and gravitational waves

Gravitational waves are sometimes portrayed as simply the effect of oscillating tidal gravity that is felt by an object. This is a quite misleading picture. In a sense, gravitational waves are caused by tidal gravity that changes, but this is just the mechanism by which gravitational wave energy is transferred from the source to the 'fabric of spacetime'.

After being transferred to spacetime as ripples in curvature, gravitational waves propagate through space in a way analogous to electromagnetic waves—in other words as a transversal traveling wave that spreads out at the speed of light.

As we shall discuss below, the observable effects of a gravitational wave is quite different from a change in tidal gravity. If we use a fixed point on Earth as a reference, then the apparent rotations of the Sun and Moon around Earth cause periodic changes in the tidal gravity that is experienced, but this is not a gravitational wave.

It may however be argued that the rotation of the Earth and the Moon around their common centre of gravity does transfer gravitational wave energy to the vacuum. These waves may be observable at some distance from the Earth-Moon system, but they will be extremely weak.

**Binary black holes as a source of gravitational waves** The simplest source of gravitational waves to analyze is the case of two identical black holes in circular orbit around each other. If measured from a nearby point on the axis of rotation of the orbit (figure 10.1), a test object will experience periodically changing tidal gravity in the transverse directions.

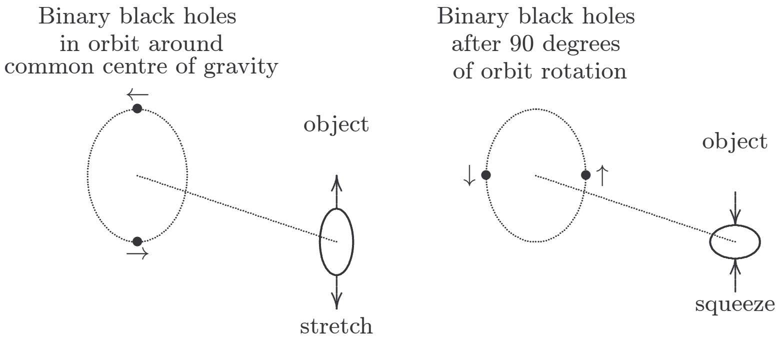


**Figure 10.1:** Orbiting binary black holes causing a periodic tidal stretch and squeeze on a nearby object, located on the axis of the orbit. Note that the period of the stretch and squeeze is half that of the orbital period of the black holes.

There will be a tidal stretch in the transverse directions corresponding to the orientation of the holes and a squeeze in the transverse directions 90 degrees from the stretch. The stretch and squeeze in the transverse directions will rotate with the orbit of the holes, so at each point in the reference frame, the stretch will become a squeeze and then a stretch again, with a period of half the orbital period of the two holes.

There will also be a detectable stretch in the longitudinal (or radial) direction, but that stretch will remain constant as the holes orbit around each other.

So are we measuring gravitational waves? Not quite. We essentially measure straightforward tidal gravity caused by the orbiting black holes. In pure Newtonian gravitational theory, tidal gravity would be all that there is.

As we have seen before, tidal gravity effects diminish rather rapidly with distance, because the strength is inversely proportional to the cube of the distance.* So tidal gravity is essentially a short range effect.

*Tidal acceleration is proportional to $\bar{M}d/r^3$, where $d$ is the diameter of an object and $r$ the distance from the source with mass $\bar{M}$.

In General Relativity, tidal gravity is explained as curvature in spacetime. So oscillating tidal gravity must be oscillations in the curvature of spacetime. The periodic squeezing and stretching is caused by curvature that alternates from being positive to being negative in the transverse directions relative to the source.

Einstein's relativity theory demands that such oscillations in spacetime curvature will progress through space at the speed of light, as a transversal wave (similar to electromagnetic waves). Objects through which this wave passes, will alternately experience a squeeze and a stretch in directions perpendicular (or transverse) to the direction of the wave's movement.

One can also say that gravitational waves are spacetime curvature that alternates from being 'concave' to being 'convex', as shown in figure 10.2.

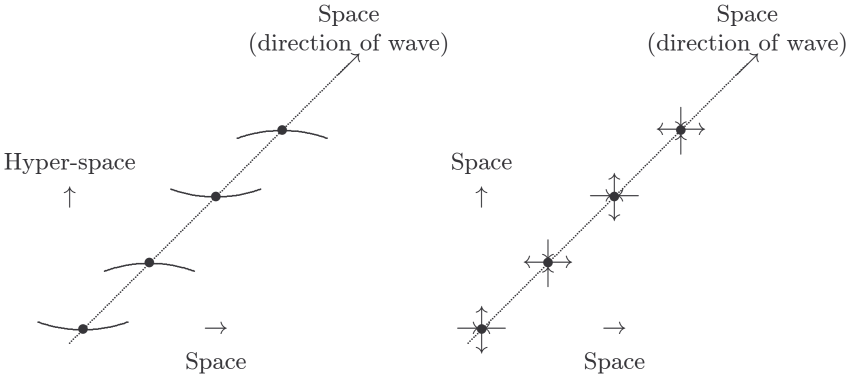


**Figure 10.2:** Gravitational wave principles. On the left is a space-propertime diagram, showing alternating directions of curvature of spacetime for one transverse space direction (the curvature is in the hyper-space direction). There are similar oscillations in other transverse space directions, which are not shown here. On the right, the stretch- and squeeze effects of the curvature oscillations on normal three dimensional space are shown.

One can say (loosely) that the orbiting black holes act as a transmitter of gravitational waves, while the space in their immediate vicinity works like an antenna. Relativists call the region inside one 'reduced wavelength' ($\lambda/2\pi$) from the centre of the transmitting system, the 'near zone',* which effectively acts as an antenna.

*E.g., [MTW, section 36.10, figure 36.3]

This transmitter/antenna combination generates two distinct waves, 90 degrees out of phase and polarized at 45 degrees relative to each other. The reason for the 45 degree relative polarization comes from the fact that the spacetime oscillations have a frequency of double that of the orbital frequency of the black holes, as illustrated in figure 10.3.

The two polarizations are referred to as the 'plus' waves and the 'cross' waves and indicated by $+$ and $\times$ respectively.

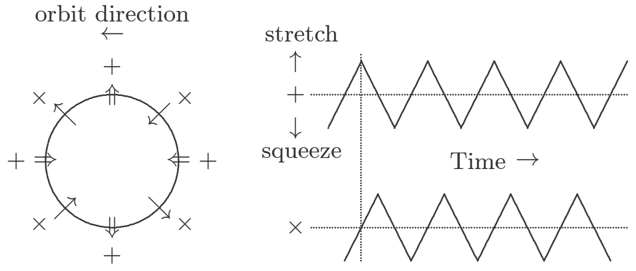


**Figure 10.3:** The two gravitational wave polarizations ($+$ and $\times$), showing their relative spatial orientation of 45 degrees (left) and relative phase shift of 90 degrees (right). The 45 degrees polarization results from the fact that the waves repeat themselves after half an orbit of the holes. Gravitational waves will not generally have a triangular shape, as shown here for simplicity.

The energy density of gravitational waves is proportional to the inverse of the square of the distance (i.e. $1/\mathbf{x}^2$) traveled, just like electromagnetic waves of light, radio transmissions, etc. Very importantly though, the amplitude of such waves (gravitational and electromagnetic), is proportional to the inverse of the distance traveled ($1/\mathbf{x}$), as depicted in figure 10.4.

As far as amplitude is concerned, we can say that in comparison to tidal gravity, a gravitational wave has a long range effect—it falls off relatively slowly with distance. Oscillating tidal gravity transfers orbital energy to spacetime as curvature vibrations at close range and gravitational waves then propagate this energy to large distances, where direct tidal gravity changes are utterly negligible.

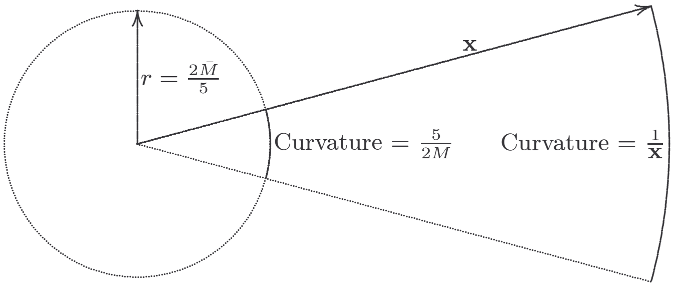

$r = \frac{2\bar{M}}{5}$
$\mathbf{x}$
Curvature $= \frac{5}{2M}$
Curvature $= \frac{1}{\mathbf{x}}$

**Figure 10.4:** A simplistic view of how the strength of a gravitational wave is diminished by distance. The radius of the circle comes from Einstein's field equations for two coalescing black holes with total mass $M$. If the geometric curvature $(\frac{5}{2M})$ of the circle represents a gravitational wave strength of 1 unit, then the geometric curvature $(\frac{1}{\mathbf{x}})$ of the circle segment at distance $\mathbf{x}$ represents a strength of: $(\frac{1}{\mathbf{x}})/(\frac{5}{2M}) = \frac{2\bar{M}}{5\mathbf{x}}$ units.

The emission of gravitational waves by orbiting bodies causes a loss of orbital energy, making the two black holes to spiral inwards towards their common centre of gravity. They will eventually spiral into each other to form a single black hole.

The unified black hole will initially be very non-symmetrical - it will have two bulges protruding from it's "equator"—the plane perpendicular to it's rotation. The bulges will be moving at close to the local speed of light and will emit large amounts of energy in the form of gravitational waves.

The loss of energy will, within a short time, cause the bulges to dissappear and the black hole will become circularly symmetrical around it's equator. By that time, more than 5% of the original mass-energy of the two black holes would have been lost due to gravitational wave emissions.

The whole process is called the coalescence of two black holes and it is thought to happen often during the early life of galaxies. Because the process is relatively well understood, this type of occurrence is a prime candidate in the quest to measure gravitational waves directly.

## 10.2 Detection of gravitational waves

The active search for gravitational waves started in the early 1960s, when Joseph Weber started to work with his gravitational wave detectors, the so-called Weber bars. We will return to this later, but first it is worthwhile to briefly examine the indirect evidence for the existence of gravitational waves.

In 1974, Hulse and Taylor discovered the first known binary pulsar, believed to be two neutron stars orbiting each other at fairly close range. They are estimated to be about 16,000 lightyears from Earth and weigh in at about 1.4 solar masses each. The orbital period of the binary system is about 27,000 seconds, from which the average separation between the two neutron stars can be computed to be about the same as the diameter of our Sun ($\approx 6.4$ lightseconds).*

This separation is not close enough to emit gravitational waves of significant amplitude and to make matters even worse, the period of the waves will be about 13,500 seconds, giving a frequency of 1/13,500 Hz, or $74\mu$Hz. So direct detection of the gravitational waves is highly unlikely.

If however, the binary pair is spiraling in towards each other due to the orbital energy lost to gravitational waves, this should be detectable as a decrease in the orbital period. The theoretical decrease is also very tiny, just $75\mu$s per year, and thus difficult to measure.

In 1983, Taylor and colleagues refined the measurements enough to report a period decrease of $76\pm2\mu$s per year. There is no other plausible explanation for this decrease in orbital period—so the existence of gravitational waves

*The separation for circular orbits is $(\frac{\sqrt{\bar{M}}\times Period}{2\pi})^{\frac{2}{3}}$, where $\bar{M}$ is the combined mass.

has been verified experimentally.

### 10.2.1 Weber bars

In the late 1960s, Weber had designed and built a detection system that could measure gravitational wave stretching and squeezing of a solid bar of aluminum, 1 to 2 metres long and half a metre in diameter, to an accuracy of $10^{-16}$ metre (that is one-tenth the diameter of the nucleus of an atom).

The reason for the size of the bar is that it would have a natural resonance frequency around 1000 Hz, a frequency that should theoretically be present in some of the strongest gravitational waves. Once the bar is disturbed by the tiny gravitational wave energy at a frequency near it's natural frequency, the bar will amplify the signal slightly and Weber picks up this vibration by piezoelectric crystals glued around the middle of the bar.

By stringing a large number of crystals in series, the voltage output would be detectable if the bar is exited by a gravitational wave with an amplitude of at least $10^{-16}$ metre of displacement. Although Weber claimed success in the late 1960s and early 1970s, other experimenters failed to confirm the results experimentally.

It is today accepted as unlikely that Weber did pick up gravitational waves, because more in-depth theoretical research has shown that the likely strength of gravitational waves reaching Earth would be a factor hundred thousand times smaller than the sensitivity of Weber's design.

The reason for the incredibly small amplitude is the large distance at which significant gravitational waves are likely to be created. The standard argument goes something like this: supernova explosions are strong sources, but the occurrence is about one per 100 to 300 years in a typical galaxy. We are unlikely to see one near us soon.

Coalescing black holes are strong sources, but they occur mostly at the core of young galaxies. All the galaxies near us are old and the closest promising candidates are thought to be at 'cosmological distances', more than 1 billion ($10^9$) lightyears away.

Very near two coalescing black holes, the gravitational waves will stretch and squeeze any object by about the same amount as it's size. Relativists call this a strength of 1 unit—it amounts to 100% stretch and squeeze. At large distances the strength is approximated by two-fifths of the total mass of the holes, divided by the distance (all in geometric units), as per figure 10.4.

For coalescing black holes with a combined mass of ten of our suns ($\approx 1.5 \times 10^4$m), at a distance of $10^9$ lightyears ($\approx 10^{25}$m), the strength is $\frac{2}{5} \times \frac{1.5\times10^4}{10^{25}} \approx 10^{-21}$. The strength of a specific gravitational wave is then multiplied by the dimensions of the object to determine the amplitude of the stretch or squeeze.

For Weber's original 1 to 2 metre long bars, the stretch and squeeze caused by the gravitational wave are $10^{-21}$ metre, or one millionth of the diameter of an atomic nucleus. The best bar detectors of today have design sensitivities in the order of $10^{-17}$ strength, still a factor ten thousand from the theoretical requirement. So how do the experimental physicists plan to bridge the gap?

### 10.2.2 Laser interferometry

The plan is to build a *laser interferometer* as a gravitational wave detector, using, instead of bars, three mirrors suspended by wires from overhead supports. The three mirrors will be arranged in a horizontal "L" shape, with each of the arms of the L several kilometres long.

By means of laser interferometry, experimentalists hope to detect the tiny changes in the length of the two arms, caused by the fact that one arm will be stretched at the same time as the other arm is shortened by a passing gravitational wave.

The USA government is investing funds into the building of a National Science Foundation facility called *LIGO*, for *Laser Interferometric Gravitational-wave Observatory* Figure 10.5 shows a schematic of the type of interferometer used in the LIGO system..*

*For a summary of the mechanics, optics and politics of LIGO, see chapter 10 of [Thorne]. Alternatively, see [Sigg] on the Internet.

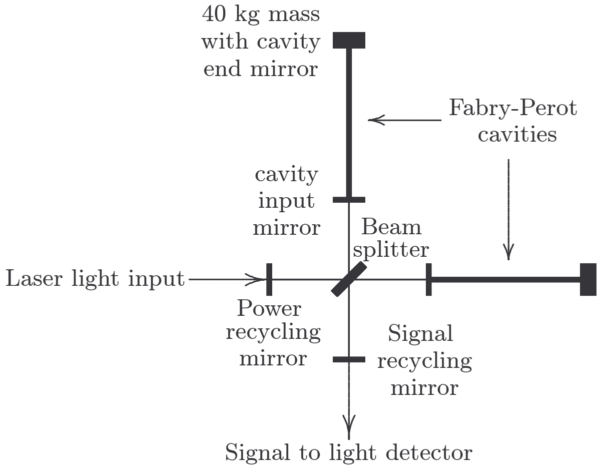


**Figure 10.5:** The LIGO optical paths in simplified schematic form. Properly tuned in length, the two Fabry-Perot cavities recycle and resonate the light input inside them to achieve hundreds of times the light energy of the input. The same thing happens, to a lesser degree, between the cavity input mirrors and the recycling mirrors. If the arm lengths are the same, then no light arrives at the detector, since the interference is destructive. See text for more details.

All interferometric gravitational wave detectors use variants of the Michelson interferometer, similar to the device that Michelson and Morley used in 1887 in their attempt to detect the movement of the Earth through the aether.

The basic principle is that of light passing through a beam splitter that sends half the light down one arm and half the light down the other arm. At the end of each arm, the light is reflected back to the beam splitter that recombines the light, half towards the input source and half in a direction 90 degrees from the beam splitter, where a light detector is situated.

With proper tuning of the length of the two arms, the light can be made to interfere destructively in the direction of the detector, so that no light is measured there. If however, a gravitational wave shortens the one arm and lengthens the other, (in other words, it disturbs the length tuning of the arms), the interference will not remain destructive and there will be a measurable light output going to the detector.

The LIGO apparatus uses a dual recycled Michelson interferometer with Fabry-Perot cavities, as shown schematically in figure 10.5. The Fabry-Perot cavities are tuned in length to match the wavelength of the input laser's light and the cavity input mirrors reflect most of the light returning from the cavity rear mirrors back to the rear of the cavities.

This light adds to the input light and sets up a resonating system that builds up the light intensity in the cavities to hundreds of times the input light. Some of the light does however escape through the cavity input mirrors and would have been lost if it was not for the power- and signal recycling mirrors, which reflect most of that light back into the system.

The design goal is to get a total light amplification in the system of a few thousand times. In essence, this means that the laser light is recycled thousands of times through each arm, multiplying the phase shift caused by stretching and squeezing of the arms by the same amount. This phase shift causes some light to "leak out" towards the photo-detector because the destructive interference is no longer complete.

The output light will be amplitude modulated by the gravitational wave. Detection is done much like in a standard superheterodyne radio receiver, where the carrier frequency is converted to an intermediate frequency that can be amplified and then demodulated to extract the gravitational wave signal.

A prototype laser interferometric detector with arms of 40 metres has been built at Caltech in the late 1990's to demonstrate the engineering feasibility. One detector with 2 km arms and two detectors with arms of 4 km are under construction - two at Caltech and one at MIT. One needs at least two detectors at different locations in order to eliminate local noise sources.

In late 2001, the LIGO team was busy with engineering runs in order to debug and calibrate the horrendously complex optics, control systems and noise suppression mechanisms. If all goes well, they will be able to reach the elusive sensitivity of $10^{-21}$ strength between 10 Hz and 10 kHz. In fact, over this frequency range, the design sensitivity of the 4 km detectors is ten times better than the minimum theoretical requirement of $10^{-21}$ strength.

In the centre range of 50 to 500 Hz, the hope is to achieve a sensitivity hundred times better than the minimum theoretical requirement, i.e. $10^{-23}$

strength. This is the frequency range where coalescing neutron stars should be transmitting strongly.

The most important effect of a 100-fold increase in sensitivity is that the expected rate of detection of strength $10^{-21}$ sources becomes a million times better. The distance at which sources become detectable increases only 100-fold, but the volume of space where detectable sources may lie increases by $100^3 = 10^6$.

The planned schedule of LIGO is to start scientific runs in 2003 with the 2 km detector and soon after that with the other two. There are also laser interferometric detectors being constructed in Germany, Italy and Japan. Once all the planned detectors are in operation, it gives a relatively long baseline system for pinpointing the direction in space from which the waves are coming with reasonable accuracy.

### 10.2.3 Cosmic music

The bandwidth of LIGO (10 Hz to 10 kHz) is just about the whole audio frequency spectrum, meaning that if you plug the output from the detector into audio equipment, you can listen to the gravitational waves directly - a sort of cosmic music. On the LIGO web pages, one can find simulations of gravitational waves that you can listen to.

For our previous example of two coalescing black holes with a combined mass of ten Suns, the audio will sound like a "chirp", building up in amplitude and frequency to reach a crescendo near 1 kHz. The chirp effect is caused by the inspiral phase, when the period of the orbit decreases and the orbital velocity increases - so both the amplitude and the frequency of the waves increase.

Just before the two holes merge, the orbit velocity approaches the speed of light, causing the crescendo at maximum amplitude. After the crescendo it 'rings down' in amplitude at a constant frequency of about 2 kHz.*

*The crescendo frequency is about 10 kHz divided by the number of solar masses represented by the combined mass of the two black holes.

The merged, but still deformed hole will usually spin at close to the maximum possible rate and produces waves at a constant frequency of roughly twice the crescendo frequency.* As the hole smooths itself out, the gravitational wave amplitude rapidly diminishes to zero.

*The maximum possible spin rate is that of the *extreme Kerr black hole*, as discussed in chapter 6.

### 10.2.4 What lies ahead

Further into the future, there are plans to put a laser interferometer into space. It is called the LISA project, for Laser Interferometer Space Antenna, with the intention to put 3 satellites in solar orbit forming a large equilateral triangle.

The main objective of the LISA mission is to observe very low frequency ($10^{-4}$ Hz to $10^{-1}$ Hz) gravitational waves from galactic and extra-galactic binary systems and gravitational waves generated in the vicinity of the very massive black holes believed to occupy the centres of many galaxies. These monster black holes can "eat" whole stars and the movement of such large masses generates strong gravitational waves.

The frequency is however very low because the event horizons of such black holes are so large that the orbital period of the in-falling stars, even approaching the speed of light, is relatively long.

For a non-rotating black hole with mass $M$, light will take $6\pi M$ seconds to orbit the hole, giving for a hole weighing in at a million Suns, a minimum period of $6\pi \times 5 \times 10^{-6} \times 10^{6} \approx 100$ seconds, equivalent to a maximum gravitational wave frequency of $10^{-2}$ Hz. Only a space based system can possibly to detect the occurences of super-massive black holes swallowing stars.

## 10.3 The purpose of it all

All the effort and expense for the projects mentioned above, would be utterly senseless if the aim was just to detect gravitational waves for the sake of detecting it. Gravitational waves promise to open a new branch of astronomy - it is after all why LIGO is called an observatory!

Despite the immense difficulties involved in the detection and recording of gravitational waves, if successful, they will provide a new and different view of astrophysical processes largely hidden from electromagnetic astronomy, such as super-massive black holes in the centres of galaxies gobbling up matter and the inner dynamics of supernova and neutron star cores.

It may eventually tell us just how right Einstein was - or when and where his general theory of relativity starts to break down. Further, by measuring low frequency background gravitational wave signals from the very early universe, it may help to discriminate between various cosmological models.

# Chapter 11

# Parametrized Post-Newtonian formalism

good enough for
many
practical purposes

In the relatively slow motion, weak gravitational fields of the solar system, general relativity can be approximated to something between Newton dynamics and 'Einstein dynamics'. Einstein's field equations are prohibitively hard to solve for multi-body systems like the solar system, even approximately.*

*In fact, rigorous solutions for multi-body problems do not exist in general.

Full solutions are not quite necessary in the weak field, low velocity limit. This is where scientists adopted the engineering-like approach of 'good enough for most (or many) practical purposes'. Eddington, Robertson and Schiff started the movement towards 'post-Newtonian' approximations of general relativity.

Note that, despite the name, 'post-Newtonian' does not mean modified Newton gravity, but rather simplified Einstein gravity. Nordtvedt later expanded on the scheme and it became known as 'Parametrized Post-Newtonian' (PPN) formalism.

It was realized that most post-Newtonian theories of gravity predicted (in the slow moving, weak field limit) a spacetime metric that has a similar structure. Most gravitational theories (including general relativity) could, in this limit, be expressed as an expansion of the Minkowski metric, but with different parameters as coefficients.

The modern, generalized and unified version of PPN formalism is due to Will and Nordtvedt. This chapter is included because if engineers get involved in observational relativity, the PPN formalism is most probably the mathematics that will confront them. The aim of this chapter is merely to introduce the topic, so that further reading can, hopefully, be easier.

## 11.1 Post-Newtonian approximation

To understand the reasoning behind the post-Newtonian approximation, it is good to revisit the exact Schwarzschild line element for the gravitational field outside a stationary, spherically symmetrical and non-rotating mass $\bar{M}$:

$$ds^2 = -\left(1 - 2\frac{\bar{M}}{r}\right) dt^2 + \left(1 - 2\frac{\bar{M}}{r}\right)^{-1} dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2, \tag{11.1}$$

where $\theta$ and $\phi$ are spherical coordinate angles in two orthogonal directions. The first term represents a time interval and the second term a radial space interval ($dr$ suggests a change in the radial distance $r$). The last two terms represent transverse space intervals, relative to mass $\bar{M}$.

As stated in previous chapters, movement in the radial and transverse directions encounter different amounts of space curvature. The $dr^2$ term has a coefficient that is a function of the inverse of the gravitational time dilation, while the transverse terms do not have such a coefficient.

It is this difference between radial and transverse movement that complicates matters significantly when real problems must be solved in Schwarzschild coordinates. To overcome this difficulty, relativists perform a transformation of coordinates that yields a 'conformally flat space' (not flat spacetime) coordinate system, called *isotropic coordinates*, which loosely means 'the same in all space directions'.

The isotropic radial space distance is obtained by the transformation:

$$r = \bar{r}\left(1 + \frac{\bar{M}}{2\bar{r}}\right)^2, \tag{11.2}$$

where $r$ is the usual Schwarzschild radial distance and $\bar{r}$ is the isotropic radial distance. This transforms the Schwarzschild metric to (e.g., [MTW] exercise 31.7)

$$ds^2 = -\left(\frac{1 - \bar{M}/2\bar{r}}{1 + \bar{M}/2\bar{r}}\right)^2 dt^2 + \left(1 + \frac{\bar{M}}{2\bar{r}}\right)^4 \left(d\bar{r}^2 + \bar{r}^2 d\theta^2 + \bar{r}^2 \sin^2\theta d\phi^2\right). \tag{11.3}$$

When comparing this line element to the Schwarzschild line element above, note the following:

$$\left(\frac{1 - \bar{M}/2\bar{r}}{1 + \bar{M}/2\bar{r}}\right)^2 = 1 - 2\bar{M}/r \quad \text{(identically)}, \tag{11.4}$$

although it may not look like it. They operate on the same time parameter $dt$.

Also note the crucial difference between the isotropic and the Schwarzschild coordinates: the radial space term $(d\bar{r}^2)$ and the transverse space terms $(\bar{r}^2 d\theta^2 + \bar{r}^2 \sin^2\theta d\phi^2)$ are now multiplied by the same coefficient $(1+\bar{M}/2\bar{r})^4$.

This means that *space* is 'conformally flat' in this metric, meaning space curvature is identical in the radial and transverse directions, simplifying calculations for complex cases significantly. *Spacetime* is however not conformally flat, because the coefficients for time and space are different.

Relativistic gravity in the solar system is usually studied in isotropic coordinates and the relativistic ephemeris for the solar system, drawn up by the Caltech Jet Propulsion Laboratory, uses this coordinate system. So does the PPN formalism, which will be elaborated upon below. Taken all together, the isotropic coordinate system is an important scheme.

As an aside, it also has the added attraction that the metric $ds$ does not diverge to infinity at $r = 2\bar{M}$, or $\bar{r} = \bar{M}/2$. As $\bar{r}$ gets smaller than $\bar{M}/2$, the corresponding Schwarzschild radial distance $r$ increases again and tends to $+\infty$ as $\bar{r} \to 0$, so that isotropic coordinates reach the event horizon, but never enter the black hole. (See figure 11.1.) If $\bar{r} \gg \bar{M}$, then $\bar{r} \approx r - \bar{M}/2$, which is very close to $r$.

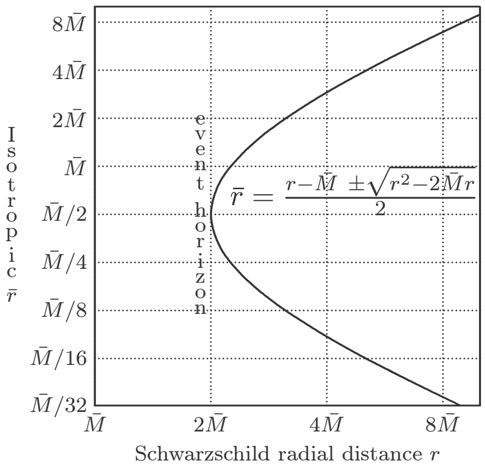


**Figure 11.1:** A log-log plot of isotropic radial distance $\bar{r}$ as a function of Schwarzschild radial distance $r$.

What will the gravitational redshift factor look like in isotropic coordinates? It has the value

$$g_{tt} = \sqrt{1 - 2\bar{M}/r} = \frac{|1 - \bar{M}/2\bar{r}|}{1 + \bar{M}/2\bar{r}}. \tag{11.5}$$

Figure 11.2 shows a plot of $g_{tt}$ agianst $\bar{r}$. Each value of redshift corresponds to two values of $\bar{r}$, but they both represent the same Schwarzschild radial distance $r$.

Despite the fact that isotropic coordinates offers conformally flat space, the full isotropic line element is still not all that easy to work with. For the type

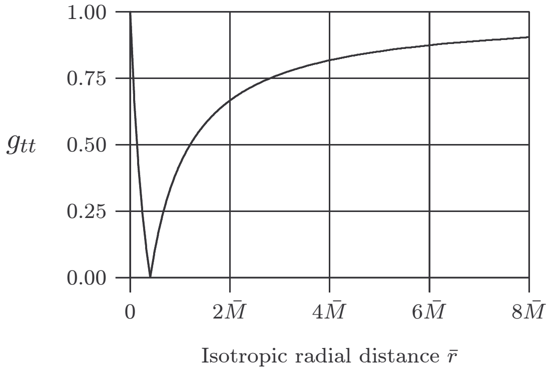


**Figure 11.2:** A linear plot of the gravitational redshift factor $g_{tt}$ as a function of isotropic radial distance $\bar{r}$, where $g_{(r)} = \frac{|1-\bar{M}/2\bar{r}|}{1+\bar{M}/2\bar{r}}$. For practical situations one can ignore values of $\bar{r} < 0.5$.

of gravitational fields and velocities found in the solar system, higher order terms of $\bar{M}/\bar{r}$ are negligible in the spatial part of the metric (the right-most term) and the isotropic line element is usually approximated to

$$ds^2 \cong -\left[1 - 2\frac{\bar{M}}{\bar{r}} + 2\left(\frac{\bar{M}}{\bar{r}}\right)^2\right] dt^2 + \left(1 + 2\frac{\bar{M}}{\bar{r}}\right)\left(d\bar{r}^2 + \bar{r}^2 d\psi^2\right), \tag{11.6}$$

called the *first post-Newtonian* (or 1PN) approximation.

The value of $(\bar{M}/\bar{r})^2$ is also extremely small,* but as we will see later, it plays an important role in distinguishing rival theories of gravity from general relativity.

Since it is no longer necessary to treat radial and transverse spatial intervals differently, the polar space coordinates $[d\bar{r}^2 + \bar{r}^2 d\psi^2]$ can be replaced by the simpler Cartesian values $[dx^2 + dy^2 + dz^2]$, giving

$$ds^2 \cong -\left[1 - 2\frac{\bar{M}}{\bar{r}} + 2\left(\frac{\bar{M}}{\bar{r}}\right)^2\right] dt^2 + \left(1 + 2\frac{\bar{M}}{\bar{r}}\right)\left(dx^2 + dy^2 + dz^2\right), \tag{11.7}$$

the most common form encountered.

In many books and articles, the bars are dropped from the radial distance $\bar{r}$ and it must then not be confused with the Schwarzschild radial distance $r$—they are different, yet very close to each other, as we have seen above.

The notations $r'$ or $r_i$ are also sometimes used for isotropic radial distance. This can all be very confusing to laypeople, but relativists are so familiar with their equations, that they read them correctly by simply noting their form.

*For planet Mercury at perihelion, $(\bar{M}/\bar{r})^2 \approx 10^{-15}$, as compared to $\bar{M}/\bar{r} \approx 3\times10^{-8}$.

## 11.2 Parametrized Post-Newtonian approximation

The post-Newtonian approximation is of a form that allows easy incorporation of other theories of gravity, by means of parameters included in the coefficients of the line element $ds$. Such schemes are called Parametrized Post-Newtonian (PPN) approximations.

The two most important PPN parameters are $\beta$ and $\gamma$ and they enter the above post-Newtonian equation as follows:

$$ds^2 \cong -\left[1 - 2\frac{\bar{M}}{\bar{r}} + 2\beta\left(\frac{\bar{M}}{\bar{r}}\right)^2\right]dt^2 + \left(1 + 2\gamma\frac{\bar{M}}{\bar{r}}\right)\left(dx^2 + dy^2 + dz^2\right), \tag{11.8}$$

where for general relativity, $\beta = \gamma = 1$.

Loosely speaking, the meaning of $\beta$ is the amount of non-linearity in the superposition law for gravity, where $\beta = 1$ means no non-linearity. It can also be thought of as the 'equivalence principle parameter'. Theories of gravity that conform to the Einstein equivalence principle all have $\beta = 1$, while those that do not conform to it have $\beta \neq 1$.

The parameter $\gamma$ represents the amount of space curvature produced by one unit of rest mass, where general relativity is taken to produce one unit of curvature per unit rest mass. There are quite a few theories of gravity that gives a slightly different curvature per unit mass, i.e. $\gamma \neq 1$.

The PPN approximation is not normally stated in terms of the line element $ds^2$, but the same information is expressed in terms of the metric tensor $g_{\mu\nu}$, where $\mu, \nu =$ 0,1,2,3, i.e., it is a 4x4 matrix, as shown below (a bit over elaborated, as a 'bordered-matrix', to make the relationship with the line element as clear as possible).

$$(g_{\mu\nu}) = \begin{array}{c c} & \begin{array}{cccc} dt & dx & dy & dz \end{array} \\ \begin{array}{c} dt \\ dx \\ dy \\ dz \end{array} & \begin{pmatrix} g_{00} & 0 & 0 & 0 \\ 0 & g_{11} & 0 & 0 \\ 0 & 0 & g_{22} & 0 \\ 0 & 0 & 0 & g_{33} \end{pmatrix} \end{array},$$

where

$$\begin{aligned} g_{00} &= -\left[1 - 2\frac{\bar{M}}{\bar{r}} + 2\beta\left(\frac{\bar{M}}{\bar{r}}\right)^2\right], \\ g_{ij} &= \left(1 + 2\gamma\frac{\bar{M}}{\bar{r}}\right)\delta_{ij}. \end{aligned}$$

Indices $i, j = 1, 2, 3$, i.e., this is for space, not spacetime. Only the space coefficients $g_{11}$, $g_{22}$ and $g_{33}$ are non-zero, as is clear from the matrix.

The factor $\delta_{ij}$ is called the 'Kronecker delta', which has the following meaning: if $i = j$, then $\delta_{ij} = 1$; if $i \neq j$, then $\delta_{ij} = 0$.

The reason why only the diagonal elements of the matrix are non-zero is due to the simplistic case considered—the gravitational field outside of an isolated, spherically symmetrical, non-rotating mass, permanently at rest at the origin of the coordinate system.

The other elements are there for describing the metric when some of these conditions are not true. We will touch on some of them later in this chapter.

For the simplistic case considered so far, the value $-\bar{M}/\bar{r}$ is equivalent to the Newtonian potential ($\Phi$). The modern version of the PPN formalism uses the generalized Newtonian potential

$$U = -\Phi = \int \frac{\rho'}{|\mathbf{x} - \mathbf{x}'|} d^3x'. \tag{11.9}$$

It is an integration of densities ($\rho'$) multiplied by the volume element ($d^3x'$), so it corresponds to mass ($\bar{M}$) divided by the appropriate space distance elements ($|\mathbf{x} - \mathbf{x}'|$), which in turn corresponds to the distnace $\bar{r}$ in the simplistic case.

The boldface $\mathbf{x}$ refers to the measurement point and $\mathbf{x}'$ to a point inside the matter distribution. This generalizes the potential to any number of stationary, spherically symmetrical masses. With this generalization, the line element is written as

$$ds^2 \cong -(1 - 2U + 2\beta U^2)dt^2 + (1 + 2\gamma U)(dx^2 + dy^2 + dz^2). \tag{11.10}$$

If the individual masses move in the coordinate system, more non-zero elements are needed in the metric tensor. The velocity of the gravitational sources introduces linear momentum and angular momentum to the metric elements $g_{0j}$ as follows:

$$g_{0j} = -\frac{1}{2}(4\gamma + 3)V_j - \frac{1}{2}W_j, \tag{11.11}$$

where

$$V_j = \int \frac{\rho' v'_j}{|\mathbf{x} - \mathbf{x}'|} d^3x', \tag{11.12}$$

and

$$W_j = \int \frac{\rho'[\mathbf{v}' \cdot (\mathbf{x} - \mathbf{x}')](x - x')_j}{|\mathbf{x} - \mathbf{x}'|^3} d^3x'. \tag{11.13}$$

$V_j$ is can be thought of as a 'linear momentum potential' and $W_j$ as an 'angular momentum potential', although they are not named as such in the literature (they are usually just called 'functionals' of the metric). The metric is now written in shorthand form as:

$$\begin{aligned}
g_{00} &= -1 + 2U - 2\beta U^2, \\
g_{0j} &= -\frac{1}{2}(4\gamma + 3)V_i - \frac{1}{2}W_j, \\
g_{ij} &= (1 + 2\gamma U)\delta_{ij}.
\end{aligned}$$

Apart from $\beta$ and $\gamma$, there are another eight PPN parameters, grouped into three generic groups.

The first of these groups (the third row in table 11.1) has to do with whether the theory has 'preferred-location' effects, e.g., galaxy induced anisotropy in the local Newtonian gravitational constant $G$. It is indicated by the parameter $\xi$, which in general relativity equals zero, because general relativity takes $G$ as a constant everywhere.

The next group is 'preferred-frame' effects, i.e., does the theory give preference to a universal rest frame (a frame that is stationary relative to the matter of the universe at large). It is indicated by a set of three parameters, $\alpha_1$, $\alpha_2$ and $\alpha_3$. General relativity does not have a preferred frame of reference and all three parameters are zero.

The last group is concerned with 'violation of conservation of total momentum' effects. It is indicated by five different parameters: $\alpha_3$, $\zeta_1$, $\zeta_2$, $\zeta_3$ and $\zeta_4$, where $\alpha_3$ contributes to two effects.

General relativity is a 'fully conservative' theory, so all five of the last group of parameters are zero. Table 11.1 summarizes* all ten 'modern' PPN parameters.

| **Parametrized post-Newtonian parameters** | | |
|---|---|---|
| **Parameter** | **Meaning** | **Value (GR)** |
| $\beta$ | Non-linearity in the superposition law for gravity | 1 |
| $\gamma$ | Space curvature produced by unit rest mass | 1 |
| $\xi$ | Preferred location effects | 0 |
| $\alpha_1$ | Preferred frame of reference effects | 0 |
| $\alpha_2$ | | 0 |
| $\alpha_3$ | | 0 |
| $\alpha_3$ | Voilation of conservation of total momentum | 0 |
| $\zeta_1$ | | 0 |
| $\zeta_2$ | | 0 |
| $\zeta_3$ | | 0 |
| $\zeta_4$ | | 0 |

**Table 11.1:** The list of ten PPN parameters with their general meanings and their values in general relativity (GR). Note that $\alpha_3$ appears in two categories, because it is a measure of both effects.

Most post-Newtonian theories that rival general relativity have at least some of the ten parameters that deviate from the values given in the table. Experimental relativists are always looking for measurements that can determine the limits on the values of the PPN parameters.

We will examine some of the tests in terms of the PPN parameters later, but first we must conclude the discussion of the PPN formalism.

*Based on a more complete table from [Will(c)], section 3.2. (available on Internet).

In the full PPN metric, there are, apart form the potentials $U$, $V_j$ and $W_j$ explained above, another seven gravitational potentials, some of them not applicable to general relativity. Since this section of the book deals with general relativity, only those potentials that are applicable to general relativity will be briefly discussed.

In the modern version of PPN, there are four additional potentials that qualify, all labeled by the uppercase Greek 'phi':

$$\begin{aligned}\Phi_1 &\equiv \text{kinetic energy potential}\\ \Phi_2 &\equiv \text{potential energy 'potential'}\\ \Phi_3 &\equiv \text{internal energy potential}\\ \Phi_4 &\equiv \text{pressure potential}\end{aligned}$$

All of them make a contribution to the principle gravitational metric element $g_{00}$, i.e., they increase the strength of the gravitational field.

Kinetic energy potential ($\Phi_1$) comes from the fact that a mass that is moving in the coordinate system appears to become more massive and thus increases it's gravitational effect.

Potential energy 'potential' ($\Phi_2$) has to do with the strength of the gravitational field that the gravitational source under consideration finds itself in.

Internal energy potential ($\Phi_3$) means that an object with internal energy, e.g., temperature and compression, creates a larger gravitational field than what it's rest mass alone would create.

Pressure potential ($\Phi_4$) says that the pressure inside a mass adds to the gravitational field that it creates.

With these potentials factored in, the full PPN metric element $g_{00}$ for general relativity is

$$g_{00} = -1+2U-2\beta U^2+(2\gamma+2)\Phi_1+2(3\gamma-2\beta+1)\Phi_2+2\Phi_3+6\gamma\Phi_4+O(\epsilon^3). \tag{11.14}$$

The last term, $O(\epsilon^3)$, indicates the 'order of smallness' of the terms that was neglected in the approximation. In this case it essentially means that terms of order $U^3$ and smaller have been omitted.

The parameter $\epsilon$ is defined by it's similarity to other variables, e.g. $\epsilon \sim U \sim \bar{M}/\bar{r} \sim v^2$, where $U$ is the PPN potential and $v$ the circular orbital velocity of a planet at radius $\bar{r}$ from mass $\bar{M}$. In the PPN limit, both $U$ and $v^2$ (and therefore $\epsilon$) are very much smaller than unity in geometric units. As a further example of the use of $O(\epsilon^n)$, the other two PPN metric elements will now read

$$g_{0i} = -\frac{1}{2}(4\gamma+3)V_i + O(\epsilon^{5/2}), \tag{11.15}$$

$$g_{ij} = (1+2\gamma U+O(\epsilon^2))\delta_{ij}. \tag{11.16}$$

The reason for the $O(\epsilon^{5/2})$ in $g_{0i}$ is that the momentum potential $V_i$ is of smallness order $\epsilon^{3/2}$, because $V_i \sim \frac{\bar{M}}{\bar{r}}v \sim \frac{\bar{M}}{\bar{r}}\sqrt{\bar{M}/\bar{r}} \sim \epsilon^{3/2}$. The above metric approximation is known as the '*first post-Newtonian approximation*', or the *1PN approximation* for short.

Further, the reader is advised to note that earlier PPN schemes used $\epsilon^2 \sim U$, whereas the modern (Will and Nordtvedt) formalism uses $\epsilon \sim U$. Many books and papers will still be found with a PPN metric similar to the above, but with the order of smallness expressed as $O(\epsilon^6)$, $O(\epsilon^5)$ and $O(\epsilon^4)$ respectively, i.e., with the smallness orders squared when compared to the above metric.

The metric given so far is all that is required for most of the tests of gravitational theory in the solar system, where possible deviations from unity for the values $\beta$ and $\gamma$ are determined. To be completely general, in the sense that it must cater for all reasonable theories of gravity, the metric must be expanded to include all ten PPN parameters.

Further, generalization to also include the effects of preferred frames of reference, preferred location effects and violation of the conservation of total energy and momentum, demands ten potentials, rather than the seven ($U$, $V_j$, $W_j$ and the four $\Phi$'s) used so far.

At the risk of frustrating some readers (and perhaps to the delight of others), we will not dicuss further generalization in this book, but rather refer the reader to the excellent, yet quite technical summary given by Clifford M. Will in [Will(c)]. For a complete treatment, see Clifford Will's book [Will(a)]. With the discussions given so far, even a novice 'amateur relativist' should be able to follow most of what Will has to say on this topic, at least in [Will(c)].

## 11.3 Some solar system tests

In this section, a brief overview will be given of the most common measurements and values of the parameters $\beta$ and $\gamma$.

**The deflection of light** Perhaps the most well known test is the deflection of light, first performed by Eddington just after the end of World War I. For a light ray grazing the Sun, the deflection predicted by general relativity is $\delta\theta = 1.75$ arcseconds.

In PPN formalism, the value is changed by $\gamma$ as follows:

$$\delta\theta \approx \frac{1}{2}(1+\gamma)1.75 \text{ arcseconds.}$$

Eddington's result were not very convincing, because the error was $\pm 30\%$ and follow-on optical tests were not much better. The development of very-long-baseline radio interferometry (VLBI) improved the accuracy greatly—they can today measure angles down to about 100 $\mu$arcseconds.

This deflection measurement technique relies on the passing of pairs of strong quasi stellar radio sources very close to the Sun (as we view them from Earth, due to Earth's orbit around the Sun). According to [Will(c)], the PPN value obtained by such deflection measurements is $0.99956 < \gamma < 1.00012$. It is usually given in the literature as $\frac{1}{2}(1+\gamma) = 0.99992 \pm 0.00014$, presumably because this relationship features in other tests as well.

**The perihelion shift of planet Mercury** This was the other early test of Einstein's theory, which predicts a relativistic shift of $\dot{\omega} = 42.98$ arcseconds per century. This value is modified by both $\beta$ and $\gamma$ as follows:

$$\dot{\omega} = 42.98 \left( \frac{1}{3}(2 + 2\gamma - \beta) \right) \text{ arcseconds per century.}$$

After a long period of radar observations of Mercury, the perihelion shift is today known to an accuracy of better than $0.1\%$. From this, the limiting value of the $\beta$, $\gamma$ combination is usually given in the form $|2\gamma - \beta - 1| < 3 \times 10^{-3}$. From the limits $0.99956 < \gamma < 1.00012$ determined by the VLBI experiments, one can extract the limits $0.99612 < \beta < 1.00324$.

**The time delay of light** This measurement was tentatively discussed in chapter 5. It was originally the most accurate test for the value of $\gamma$, until it was surpassed in accuracy by the VLBI light deflection, as quoted above.

The two-way PPN time delay for a ray which passes close to the Sun on it's way to a planet at superior conjuction and back to Earth, is

$$\delta t \approx \frac{1}{2}(1+\gamma)[240 - 20\ln(d^2/r)] \ \mu\text{s},$$

where $d$ is the ray's closest approach to the Sun in solar radii and $r$ the distance of the target planet from the Sun in astronomical units.

The most accurate time delay measurements were made by Irvin Shapiro and his team, using the Viking Mars landers as transponders for a radar signal transmitted from Earth. The predicted two-way time delay is only in the order of 250 $\mu$s, equivalent to some 75 kilometres of light travel distance.

In order to achieve the 0.1% accuracy quoted for the time delay, the 'Newtonian distance' between Earth and Mars, at the time of superior conjunction, had to be known to within 38 metres. This is an extremely tough requirement, considering that the Earth and Mars are then some 278 million km apart and we have no way of measuring the 'Newtonian distance' directly at the time of superior conjunction. So how did Shapiro do it?

He radar ranged Mars over many weeks, starting while it was still far enough away from superior conjunction so that the time delay effect is negligible. Because the orbit of the Earth around the Sun is pretty accurately known, a fairly accurately measured segment of the orbit of Mars around the Sun was obtained.

With this knowledge, the PPN coordinate positions for Mars and Earth near superior conjunction were predicted using standard astronomical methods, giving a 'Newtonian round trip time'. This was then compared to the measured round trip time near superior conjunction and the difference gave the time delay.

Finally a least-squares fit of the two sets of data (measurements and predictions) was performed to estimate the parameter $\gamma$ as $1 \pm 0.002$. The VLBI light deflection tests that were treated above, later improved this accuracy to $\gamma = 1 \pm 0.0003$.

## 11.4 The Brans-Dicke theory of gravity in brief

This theory had a line of development by Jordan, Fierce, Brans and Dicke, but is generally known as the Brans-Dicke theory. It is a special case of a general class of scalar-tensor modifications to general relativity. The Brans-Dicky theory enters the PPN formalism only as a modification to parameter $\gamma$, via the positive 'Dicke coupling constant' $\omega$:

$$\gamma = \frac{1+\omega}{2+\omega}, \tag{11.17}$$

meaning $\gamma$ goes from 0.5 to 1 as $\omega$ goes from 0 to $\infty$.

Recall that $\gamma$ represents the amount of space curvature produced by unit rest mass and it is clear that the Brans-Dicke theory predicts less space curvature than general relativity does. When the theory was first promulgated, Brans and Dicke favored a value of $\omega = 5$, (or $\gamma = 0.857$) because it was the smallest value of $\gamma$ that was reasonably compatible with the experimental evidence of the time (circa 1973).

As we have seen above, modern observations give the limit as $\gamma = 1 \pm 0.0003$. This forces $\omega > 3000$, making the Brans-Dicke theory virtually indistinguishable from general relativity.

## 11.5 So was Einstein really right?

The best we can say today, is that his theory must be very close to how things actually work. In almost a century of observations, not a single one disagreed with general relativity. It does however not say that, as observational errors decrease, we will not eventually find one that disagrees.

There is a saying amongst relativists that 'every new test is a potentially deadly one'. The general feeling today is that the breakdown point of general relativity may be inside black holes, very close to the central singularity, where spacetime curvature diverges to infinity.

Another place where the theory might break down can be the presumed cosmological singularity at the beginning of time. Neither of these places are accessible for observation, however.

It may also be that some accessible place can show up a discrepancy in the theory. This will not make general relativity 'wrong', just like general relativity did not make Newton's theory 'wrong'. There may just be a limit to it's applicability, demanding a theory with wider applicability.

The search is already on, and the thing searched for is called *quantum gravity*. Maybe, in a few decades (or years?) from now, someone will be able to write "An Engineer's Perspective on Quantum Gravity".

# Chapter 12

# Introduction to Cosmology

how
engineers might view
the cosmos

Cosmology is the study of universe at large. It is an attempt to make physical sense of the material cosmos, where it came from and where it is heading.

This text is written for the engineers and the likes of them. One could ask the question: why would engineers be interested in cosmology? For one thing, the cosmos can be viewed as one extremely large machine that follows the laws of physics in it's operation. Machines of all sorts, and especially the design of such machines, are the domain of engineers.

Further, the 'cosmic machine' fulfils the very practical purpose that it provides a home for mankind to live in. So it is natural that we would like to know how the machine works. The problem is that it is somewhat as if we are elementary particle-sized beings, living on one of the spinning ball bearings in this gigantic machine.

With the little that we can observe from where we live, we try to reverse engineer the machine—at least the cosmologists do, if they will forgive me for the engineering approach. In this approach, the objective is to reconstruct the engineering specifications, algorithms, parts lists and raw material requirements of this complex machine.

## 12.1 Cosmic design

Taking their cues from the cosmologists, engineers may possibly view the cosmological 'machine' (from inside out, or 'bottom-up') as follows. Our 'ball', the Earth, is an elementary part of a basic component (the bearing), which we call the solar system. Many such basic components make up a 'sub-assembly' called the Milky Way galaxy.

We know that there are other similar galaxies that are gravitationally coupled to our Galaxy, amongst them the Andromeda galaxy and the two Magellanic clouds. So the 'sub-assemblies' are integrated into a larger 'assembly' of galaxies that is called the Local Group.

There are other groups of galaxies, some much larger than our Local Group, which are called clusters due to their size. The Local Group can be viewed as a 'mini cluster'. The groups and clusters are integrated into a 'sub-system' called a supercluster.

Our Local Group is part of the Virgo supercluster, so called because the Virgo cluster is an extremely large collection of galaxies that we observe in the direction of the Virgo constellation of stars. There are many such superclusters that we can observe and they are all integrated into a larger 'system'—the observable universe.

The integration seems to be via connecting elements made of filaments or sheets of galaxies, leaving huge, bubble-like volumes of empty space between them. The empty 'bubbles' are called cosmic voids.

At present it does not seem as if there are larger scale substructures in the cosmic 'machine', so for the purposes of this book (and for the benefit of the system engineers), this is our 'system'.

This is a gross oversimplification of the structure of the real cosmos, but it serves to illustrate the idea. Figure 12.1 depicts the simplified hierarchy and some characteristic sizes of the system.

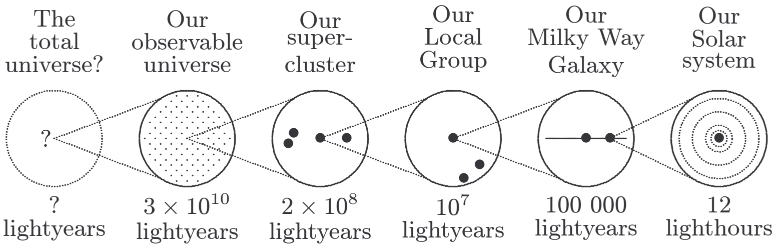


**Figure 12.1:** A simple hierarchical view of our place in the cosmos and some characteristic sizes. The leftmost circle represents the entire universe, of unknown size. The other circles represent small portions of the circle on it's left. Only major or important components are shown in each circle.

A supercluster is in diameter some $10^{11}$ times as large as our solar system. Further, a typical supercluster has a diameter of almost 1% of the observable universe. This gives some idea of how large a supercluster actually is.

Everything up to groups and clusters of galaxies seems to be gravitationally bound in the sense that their components seem to be orbiting the centre of gravity of the structure.

Superclusters do not appear to be gravitationally bound in this sense, due to the large distances between the clusters and the fact that gravitational influence diminishes with the square of distance. For example, the distance between us and the Virgo cluster is presently some 60 million lightyears. The Virgo cluster, being so massive, does gravitationally influence other clusters in the supercluster, but not in the sense of creating bound orbits.

The approximate diameters and distances that were quoted above are all in units of light travel time, as is implied by the units light years. We will later see that light travel time does not necessarily mean the physical distance between objects.

## 12.2 The expanding universe

The expansion of the universe is normally visualized by means of a balloon that is being inflated. A better and perhaps, by modern knowledge, a more correct visualization is that of an infinitely large lattice of rods, normally referred to as *Escher's infinite lattice* [Gribbin, page 43]. Figure 12.2 shows just one (cubic) element of such a lattice, consisting of nodes (the black dots) and connecting rods.

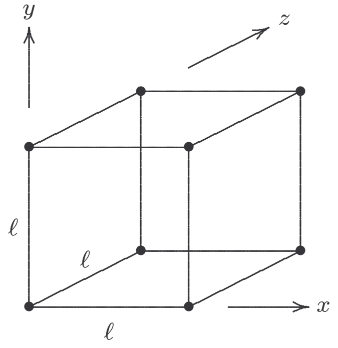


**Figure 12.2:** One element of Escher's lattice, with rod length $\ell$. Repeat this element indefinitely in all directions and we have **Escher's infinite lattice**.

Imagine this element being duplicated indefinitely in all directions and we have Escher's infinite lattice. Let the rods represent space and the nodes superclusters, with the length of all rods being $\ell$ units at the present time. If in every time increment, say $\Delta t$, all rods were to lengthen by some constant amount, say $\Delta\ell$ units, the recession speed between any two adjacent nodes will be

$$v_r = \Delta\ell/\Delta t$$

units (see figure 12.3). The recession speed between any node and the second node from it will be

$$v_r = 2\Delta\ell/\Delta t$$

units, because between them there are two rods that each lengthens by $\Delta\ell$ units in time $\Delta t$. In general we can say that the recession speed between any two nodes is

$$v_r = n\Delta\ell/\Delta t = nk,$$

where $n$ is the number of rods between the nodes and $k = \Delta\ell/\Delta t$, which is constant, as defined above. This is essentially *Hubble's law* for the recession velocity of galactic clusters, discovered by Edwin Hubble in the early 1920's.

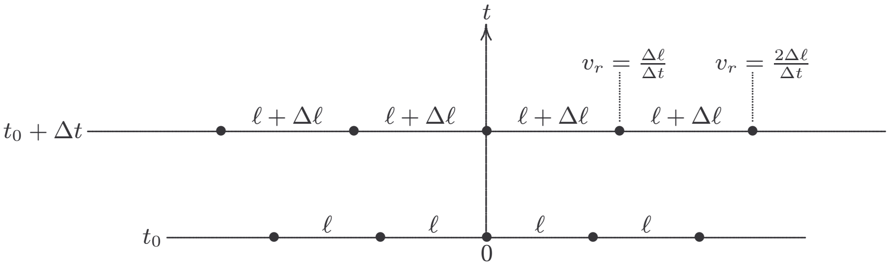


**Figure 12.3:** Two views of a one dimensional portion of Escher's infinite lattice, one for time $t_0$ and one for time $t_0 + \Delta t$. In the time $\Delta t$, distances between nodes stretch from $\ell$ to $\ell + \Delta\ell$. The recession speed of the first node form the origin is $v_r = \frac{\Delta\ell}{\Delta t} = k$ and the speed of the second node from the origin is $v_r = \frac{2\Delta\ell}{\Delta t} = 2k$.

The law states that the apparent recession velocity of a cluster, as measured by it's cosmological redshift is directly proportional to it's distance from us. The constant of proportionality has been aptly named the *Hubble constant*, $H$, i.e.,

$$v_r = sH, \tag{12.1}$$

where $s$ is the distance to the cluster as measured by light travel time. Since it is possible that $H$ may change over time, the present value of the Hubble constant is usually denoted $H_0$ and is expressed in km/s per Megaparsec (Mpc).

One Mpc is about 3.3 million lightyears. By the turn of the century, the measured value of $H_0$ has converged to around 64 km/s/Mpc. A value of about 73 km/s/Mpc has lately appeared to be the 'best buy', but it may still change in time to come.

So we will stick, for the time being, to the older value. This means that a galaxy at a distance of 1 Mpc (or 3.3 million lightyears) should appear to recede from us at a radial speed of 64 km/s.

Now this is one enormous speed if reckoned by terrestial standards. Unfortunately for astronomers, this is a rather low velocity by astronomical standards. Earth is moving at about 30 km/s in it's orbit around the Sun. The Sun itself is moving at over 200 km/s around the centre of the Milky Way.

Now add in the Milky Way's orbit around the centre of gravity of the Local Group and the fact that the whole Local Group seems to be moving at around 600 km/s. This movement is in a direction towards a presumed conglomeration of superclusters called the 'Great Attractor'.

You can imagine the problem. To extract the *pure Hubble flow* (the 64 km/s in the example) from the so called *peculiar motion* of galaxies of up to 600 km/s, is a difficult task. It is only for galaxies or clusters at distances of a few hundred million lightyears from us that the Hubble flow dominates the observed velocity.

For example, a cluster at 330 million lightyears will have a Hubble flow of about 6400 km/s, which is some ten times that of the peculiar velocities. How distances of hundreds of millions of lightyears are measured will be discussed in a later chapter.

Our constant $k$ is essentially the same thing as $H$, although we use a 'quantized' distance in the form of the number of 'Escher rods' between us and the cluster. Make the rods short enough and $k$ is indistinguishable from $H$, at least for the present state of expansion of the universe.

If we arbitrarily assume that at present the Escher rods are one lightyear long and we express velocity as a fraction of the speed of light (so that velocity is a dimensionless quantity and the speed of light is unity), $k$ has the units lightyear$^{-1}$.

The conversion from $H$ to $k$ involves changing km/s to a fraction of light speed and converting Mpc to lightyears. The result is

$$k \approx H \times 10^{-12} \text{ lightyear}^{-1}.$$

A very interesting (but quite naive) question to ask at this point is: if the universe works like Escher's infinite lattice, how far can an object be from us before it's recession velocity reaches the speed of light? It is just the inverse of $k$, i.e., $10^{12}/64 \approx 1.6 \times 10^{10}$ lightyears.*

*Recall that $v_r = nk$, so that $n = v_r/k$. Here $v_r = 1$ and $n$ is the number of rods of length 1 lightyear each.

This roughly correlates with the radius of the observable universe—we can presumably not observe anything that, due to expansion of the universe, is receding from us at a speed higher than the speed of light. We will later see that this statement is a bit misleading in an expanding universe.

If Escher's infinite lattice always followed the above law in the past, then there must have been a time when the length of all the rods between the nodes must have been zero. How long ago would that have been? We simply calculate how long it would take a rod, starting at zero length, to reach it's present length (which is one lightyear according to our units).*

*Recall that $k = \Delta\ell/\Delta t$, where here $\Delta\ell = 1$ lightyear.

The answer is again

$$t = 1/k = 10^{12}/64 \approx 1.6 \times 10^{10} \text{ years},$$

the same numerical value as the one that we obtained above for the radius of the observable universe. It makes some sense that one can observe things only as far as light has had time to travel since the 'beginning'.

With rod lengths close to zero, the 'beginning' must have been a place approaching infinite density, but possibly still of infinite size—an infinite number of rods of infinitesimal length still add up to an infinite size!

On the other hand, our observable universe is of finite size and at the beginning this 'patch' of the universe must have been very close to a single point. Do we know where that single point is (or was)? Yes, it is right here where we are, because our present place must have been at that point—just like every other place in our observable universe must have been at that point. The 'single point' has become the observable universe.

The simplistic 'Escher model' does not represent the present thinking about the dynamics of the universe. As we shall see in chapter 16, it comes very close (at least for the present epoch of expansion) to one of the 'latest and greatest' theoretical schemes that fits observational evidence pretty well.

## 12.3 The cosmologist's approach

In this introduction, only some of the principles and symbology that cosmologists use in their technical treatment of cosmic models will be given. A more detailed treatment will follow in later chapters.

Firstly, cosmologists prefers to work with what they call *comoving coordinates*, a fancy name for a rather simple concept. They are effectively just counting the rods in the 'Escher model'.* The number of rods between two any two nodes obviously remain constant, no matter how much the rods expand, or shrink for that matter.

This is equivalent to the comoving distance between clusters and is denoted by the symbol $r$,* which is obviously not the same as the *proper distance* between the two clusters. Proper distance is measured by a non-stretching ruler, which can be done by measuring the time an electromagnetic signal takes to propagate the distance.

The proper distance increases in an expanding universe and cosmologists obtain it by multiplying the comoving distance ($r$) by a time varying *expansion factor*, denoted by $a(t)$, i.e., $a$ as a function of time. The proper distance is thus

$$\ell = a(t)r, \tag{12.2}$$

where the expansion factor $a(t)$ is chosen so that at present, $a(t_0) = 1$.

When the rods of Escher was half their present size, then $a(t) = 0.5$. As an example, a galaxy that today has a comoving distance of $r = 100$ million lightyears will have a proper distance $a(t_0)r = 100$ million lightyears, because $a(t_0) = 1$. A long time ago, when the expansion factor was $a(t) = 0.5$, the comoving distance of the galaxy was still 100 million lightyears, but it's proper distance was $a(t)r = 50$ million lightyears.

*There is no official 'Escher model' of the universe—it is just a convenient reference.

*Comoving distance is also denoted by the symbol $\chi$. See table 12.1 below.

Cosmologists sometimes drop the $t$ in $a(t)$, and just refer to the time varying expansion parameter as $a$. The expansion factor $a(t)$ is of great importance in cosmology, because it tells us how the expansion has changed over time.

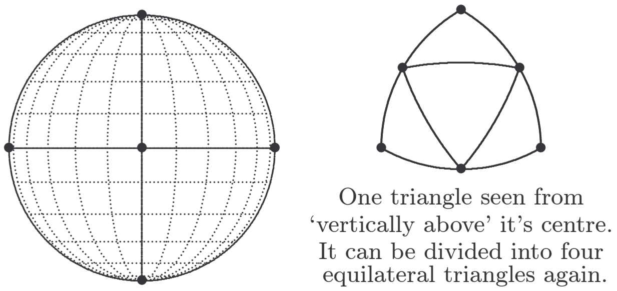
The 'Escher model' is essentially a flat, Euclidean space model. Space may not be completely flat, but may have curvature. Curvature was treated thoroughly in the first part of this book, on relativity. We will 'borrow' from general relativity a hypothetical three dimensional 'hyperspace' and then mathematically embed normal three dimensional space is into this hyperspace domain.

Now normal space can curve into the extra (hyperspace) dimensions, just like the two-dimensional surface of the Earth curves into the third space dimension. With this postulate, one can construct an Escher lattice, so that it has curvature, yet with all rods having the same length. This is impossible in normal three dimensional space.

It is also impossible to visualize such a lattice, let alone sketch one. We are forced to drop one space dimension and consider only two dimensions of space as all there is. The third dimension can now be 'viewed' as one dimension of hyperspace.

'Viewed' is actually a bad word choice because the third dimension is now hypothetical since all observers must be considered two dimensional beings. Such observers cannot directly view the third dimension. We will later see that they can, in principle, measure the curvature of the two dimensional surface, however.

Assume for now that hyperspace is a perfect sphere and that our normal space is the surface of this hypersphere. On the surface of this sphere, we now try to build a two dimensional 'Escher lattice'. Using curved squares will not work in this limited form of hyperspace, but we can fit eight curved equilateral triangles perfectly onto the surface of the sphere, in such a way that they cover the complete surface with no overlaps or gaps. Figure 12.4 shows the four triangles on half of the sphere.



One triangle seen from 'vertically above' it's centre. It can be divided into four equilateral triangles again.

**Figure 12.4:** The solid lines (left) represent four equilateral triangles fitted onto half of a sphere. All inside angles of each triangle span $\pi/2$ radians, so that inside angles of each triangle add up to $3\pi/2$ radians, instead of the $\pi$ radians of a flat triangle. Each triangle can be subdivided as shown on the right.

Each of these triangles can be subdivided into four equilateral triangles again, as shown on the right of the figure. This subdivision can be repeated as many times as one wishes.

Every resulting 'connecting rod' will curve towards the centre of the sphere with the same curvature as the original (quarter circumference) rods. In the limit, when the number of subdivisions approach infinity, the infinitesimal little triangles will be indistinguishable from normal flat, Euclidean triangles of the same size.

Their inside angles will add up to $\pi$, as near as can be. However, their sides will still have the same curvature, with a radius of curvature equal to the radius of the sphere. The curvature may just not be detectable, due to the tiny size of the triangles in comparison to the radius of the sphere.

The triangles do not have to be tiny—one can just make the radius of the sphere extremely large and the same effect will appear with large triangles.

The reason for laboring this rather trivial point, is that the cosmos may perhaps be somewhat like this, with our observable neighborhood equivalent to one of the triangles, apparently perfectly flat, due to a very large radius of curvature.

If the curvature is positive, the universe is called 'closed', since it is finite, yet unbounded, like the surface of the sphere. In an expanding universe, positive curvature (traditionally) means a universe with an expansion rate that is insufficient to sustain the expansion against the mutual pull of gravity. It is expected to reverse the expansion at some time in the future and collapse towards an infinitely dense state again.

Curvature can also be negative, which can be thought of as the 'inside' of a hyperbola, as will become clearer later. Such a universe is called 'open' and is taken to be unbounded, because it is infinite in size.

Traditionally, negative curvature means an expansion rate that is large enough to sustain the expansion, against the mutual pull of gravity, forever. A universe with zero curvature is traditionally called 'flat' and is also unbounded and infinite.

Zero curvature means that the expansion rate is just large enough so that the expansion rate will decrease asymptotically to zero, so that it will just not collapse. We will later see that these traditional definitions have become a bit blurred lately, due to the discovery that the universe may be 'flat', yet it may actually be expanding at an increasing rate!

The symbology that cosmologists use to describe the curved, expanding universe is illustrated in figure 12.5, where just one of the curved 'Escher rods' is shown—the circular arc between an observer and a distant object.

The time varying *radius of curvature*, denoted by $R(t)$ is equivalent to the radius of the expanding hypersphere. The present radius of curvature is usually written as $R$, as shown in the figure. The expansion parameter $a$ is now $a(t) = R(t)/R$, meaning the radius of curvature at time $t$ as a fraction of the present radius of curvature.

The comoving distance angle $\chi$ has the same meaning as the linear comoving distance $r$ used in the linear 'model', because for any given object, the angle

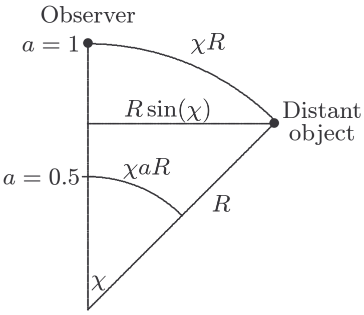


**Figure 12.5:** The symbology of curved space for a closed cosmos with positive curvature. $R$ is the present radius of curvature, $\chi$ the comoving distance and $a$ the expansion factor. For a flat cosmos, $R$ tends to infinity and $\chi$ tends to zero, but $\chi R$ remains finite. $R\sin(\chi)$ is shown because it is used in cosmological equations for the closed cosmos. For the open, negative curvature cosmos, $\sin(\chi)$ is replaced by $\sinh(\chi)$, where $\chi \to i\chi$ and $R \to -iR$.

$\chi$ remains constant as $R(t)$ grows with the expansion of the universe. The comoving distance $\chi R$ is equivalent to the proper length of a stretched Escher rod.

The reader may have found the last paragraph a little confusing and the author can sympathize with that—cosmology can be very confusing, especially since the specialists seem unable to agree on the terminology and symbology for presenting the subject. As an example, two modern textbooks that was consulted, [Peebles] and [Peacock], use different terminology and symbology, as listed in table 12.1.

The differences in symbology are striking and there are some difference in terms as well. It must be said that Prof. Peacock uses ‘scale factor’ and ‘radius of curvature’ somewhat interchangeably, while Prof. Peebles does not use the term ‘scale factor’ at all. In their defense, one must say that contemporary cosmologists suffer under a huge legacy of well established terminology and symbology—well established, but not very consistent.

This text uses the terms and symbology from Prof. Peebles because they are in some sense more intuitively ‘accessible’, e.g. ‘expansion factor’ rings a bell that ‘normalized scale factor’ does not. One might however argue that, judging by the publication dates, Peacock’s textbook may be the ‘more modern’ one.

## 12.4 An engineering approach

Engineers are meant to take the science that scientists develop and turn it into items of practical use. In cosmology, the engineer’s role is usually limited to designing equipment and machinery that can be used to measure the universe at large, so that cosmologists can do their science.

| Comparison between some Peacock and Peebles parameters | | | | |
|---|---|---|---|---|
| | **Parameter name** | **Peacock** | **Peebles** | **Units** |
| 1 | Comoving distance (angle) | $r$ | $\chi$ | rad. |
| 2 | Present scale factor | $R_0$ | - | Mpc |
| 3 | Present radius of curvature | - | $R$ | Mpc |
| 4 | Present comoving distance | $R_0 r$ | $R\,\chi$ | Mpc |
| 5 | Present normalized scale factor | $a$ | - | - |
| 6 | Present expansion factor | - | $a$ | - |
| 7 | Time varying normalized scale factor | $a(t)$ | - | - |
| 8 | Time varying expansion factor | - | $a(t)$ | - |
| 9 | Time varying scale factor | $R$ or $a(t)R_0$ | - | Mpc |
| 10 | Time varying radius of curvature | - | $R(t)$ or $a(t)R$ | Mpc |

**Table 12.1:** A comparison between some of the parameter names and symbols used by Peacock and Peebles respectively. Inside the double rows: 2/3, 5/6, 7/8 and 9/10, the two names describe the same parameter. This book uses the Peebles parameter names and symbols.

When working in this type of environment, engineers need to understand what cosmologists are speaking about and probably need to understand the physics and mathematics that cosmologists are using, at least to some degree. Engineers in other environments have little reason to understand the cosmologists; that is unless they have an interest in the cosmological 'machine'.

For their benefit the physics can be presented in a simplified form. The spherical model of fig. 12.5 can be replaced by a flat* model without loosing much of the principles involved.

Later chapters will show that this model, using rather simple mathematics, yields very closely the same results than the more complex standard cosmological mathematics.

In the next chapter, we will turn to those parameters, but before we go there, a few introductory words about the units of measurement used in this text.

*Here 'flat' means using non-spherical mathematics and not 'flat' in terms of large scale curvature.

## 12.5 Cosmological units

Th SI convention of units is useful for 'ordinary' mass, time and distance, but becomes a bit cumbersome when working with cosmological mass, distance and time. Astronomers like to use the unit 'solar mass' for expressing mass in the cosmos and parsec (pc) for distance. A solar mass is self explanatory and a parsec is the distance at which an object would have an annual parallax of one arc second.

Annual parallax means the peak (not peak to peak) shift in angular position of an object against the very distant stars, observed over a period of one year. So the baseline is the radius of Earth's orbit ($\approx$ 150 million kilometres) and one parsec $\approx \frac{150 \text{ million km}}{\tan(\text{one arc second})} \approx 3.09 \times 10^{13}$ kilometres or 3.2616 light years.

Because the parsec and it's common multiples, e.g. the kiloparsec (kpc) and the Megaparsec (Mpc) are not intuitively understandable by novices, popular books normally express distances in lightyears or orders of magnitude thereof, e.g., million lightyears or billion lightyears, where billion usually means a thousand million.

Since there are traditionally two meanings of 'billion' (USA $10^9$ and UK $10^{12}$),* this text will adopt the 'more engineering-like' convention of *Giga-lightyear* (abbr. Gly $= 10^9$ lightyears) for cosmological distance and *Gigayear* (abbr. Gy $= 10^9$ years) for cosmological time. Cosmology books sometimes use Ga (*Giga-annum)* for the latter.

*Oxford dictionary, 1990: "**billion ....** **2** (now less often, esp. *Brit.*) a million million ...".

# Chapter 13

# The Einstein-de Sitter Universe

The
simplest
model

Way back in 1932, Einstein and de Sitter presented the ‘standard’ model of the cosmos to the world.* It is the simplest possible model and has been the favorite amongst cosmologists until the 1980s. In a way it was the ‘de facto industry standard’ for over 50 years.

*This must not be confused with the ‘De Sitter model’, which was an earlier effort of De Sitter alone [Peebles, chapter 5].

In short, this model started in an extremely dense state, much like the elementary ‘Escher model’ discussed in the introduction. The expansion rate must have been extreme in the beginning, but fine tuned so that the gravitational pull of the matter in the universe was precisely balanced by the kinetic energy of expansion. In other words, potential energy and kinetic energy of expansion had to balance out, with a nett energy of zero.

This fine balance had to be maintained until the present and, according to this model, will be maintained forever. This chapter will examine the major properties of the ‘Einstein-de Sitter model’. More modern variations of this model will be examined in later chapters.

## 13.1 Einstein-de Sitter spacetime

What must be stressed is that particles, atoms, molecules and later congregations of matter are not expanding into pre-existing space. It is space itself that is expanding.

When we draw a diagram of Einstein-de Sitter spacetime, as shown in figure 13.1, it appears as if the expansion were driving matter outwards at speeds exceeding that of light. But light itself was being driven outwards with the expansion, so that relative to the (stretching) fabric of spacetime, light was still moving at its normal speed and matter were always moving at less than the speed of light.

The curve in figure 13.1 shows the "edge" of the observable universe as it would have looked in one space and one time dimension over the history of the universe. Actually, 'looked' is not a good word choice, since no observer could have 'seen' the 'edge' of the universe like that.

Since we are limited to observe the universe by means of light and other electromagnetic radiation, which do not have infinite propagations speed, we see a completely different picture. We can however use observational data to draw such a graph, but it will always be somewhat model dependant—in this case the Einstein-de Sitter model.

The scales of time and space shown is further dependant upon the value of the Hubble constant, $H_0$, for which 50 km/s/Mpc was used here, simply because it was the favorite value for most of the time of Einstein and de Sitter.

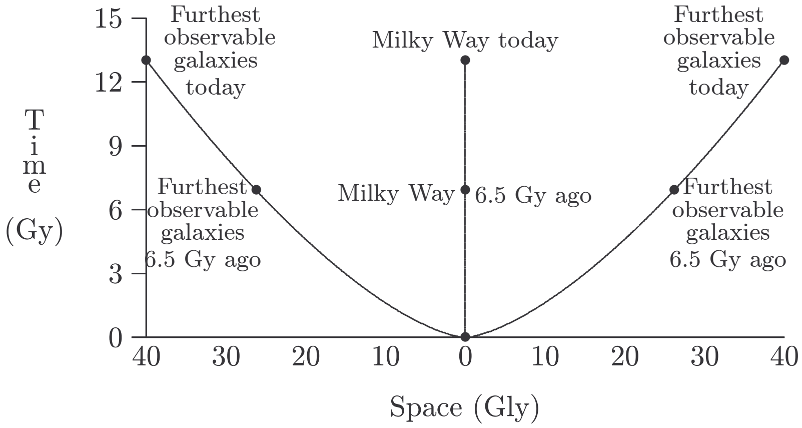


**Figure 13.1:** A spacetime diagram for the observable universe according to the Einstein-de Sitter model, where the expansion curve is parabolic. The shown positions of the remote galaxies are not where they are observed, but their actual (presently unobservable) positions.

We will now briefly look at what 'observable' space in an expanding universe means.

## 13.2 Observable space

Our main observational method of the universe is through electromagnetic radiation (photons) of various wavelengths. Unlike material objects, which can be stationary in space, photons cannot be stationary. They always have to move at the speed of light in some or other direction in space.

If we trace the path of a photon that was tranmitted in our direction from near the edge of the (expanding) observable universe at a very early time, it will follow one side of the teardrop-shaped curve shown in figure 13.2. The reason for the shape is that when the expansion rate was very high, the photon would effectively have been dragged away from us (the central world line).

Since the photon always moves at precisely the speed of light through local space, it will move away from the 'edge' and eventually find itself in a region where the rate of expansion is slow enough for it to start approaching the central world line. Eventually it will reach us and can be detected.

Photons from the very edge of the observable universe will take the full age of the universe to reach us. Areas further than the current edge will reach us some time in the future. From areas inside the current edge we will observe a continuous stream of photons, which will be elaborated on in the next paragraph.

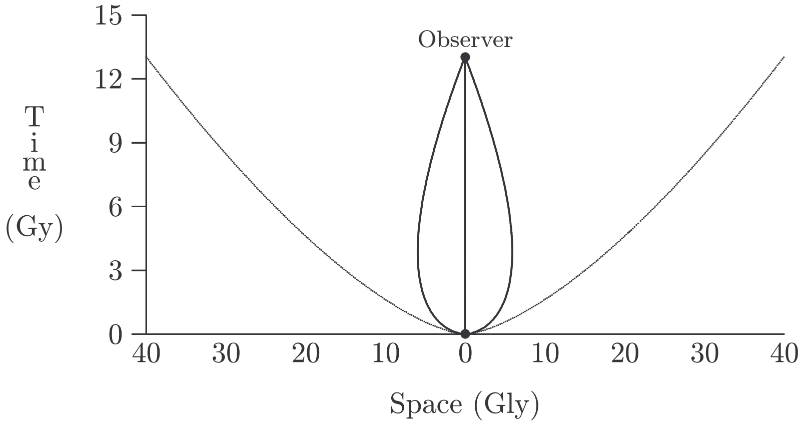


**Figure 13.2:** The teardrop-shaped curve in the centre represents the paths of two photons, transmitted in opposite directions from the edge of the observable universe when the universe was very, very young. They are presently being observed in the Milky way for the first time. For every instant in time, there is a slightly different 'teardrop', representing the path of another pair of photons.

Essentially, their are an infinite number of 'teardrops', one for every instant of observation. However, at one instant, we can in principle observe an infinite number of photons, coming from different distances, all following the same 'teardrop' path.

In just one space dimension, observing multiple photons may be difficult, if not impossible. In more than one dimension, the problem largely dissappear. We can then resolve different photons arriving simultaneously from slightly different directions.

In figure 13.3, the parabolic expansion curves for the spacetime of two galaxies at intermediate distances are drawn, one at one third and one at two thirds of the distance to the end of observable space.

We observe them as they were when the universe was 0.46 Gy and 3.8 Gy old respectively, as shown in the figure. Light took $\approx 13 - 0.46 \approx 12.5$ Gy and $\approx 13 - 3.8 \approx 9$ Gy respectively to reach us from those galaxies. This is also their respective distances in *light travel time*.

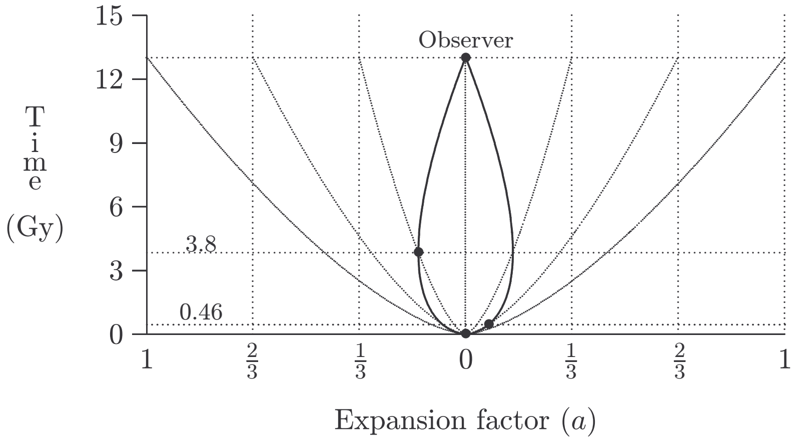


**Figure 13.3:** The spacetime expansion curves of two pairs of galaxies, presently at 10 and 20 Gly from us respectively. Where the curves intersect the 'teardrop' is where the galaxies were when we observe them today—in principle at least—they may be too far to be observed in practice.

## 13.3 Standard expansion model concepts

In order to comprehend cosmological models, we must first firmly establish some basic concepts around the Hubble constant and it's units. The Hubble constant $H_0$ (pronounced 'H naught' or 'H zero') is defined as the apparent speed of recession of a distance object per unit distance.

In the SI convention, the units of $H_0$ should really be (metres/second)/meter, giving second$^{-1}$. This would give an extremely small value for $H_0$, so Edwin Hubble decided to use the units km/s/Mpc, giving a 'friendly' range of values, between 50 and 100 km/s/Mpc.*

Cosmologists further define a dimensionless *Hubble parameter h*. It has

*One Mega-parsec (Mpc) is about 3.3 million lightyears.

the value $h \equiv H_0/(100$ km/s/Mpc), with an original 'best fit value' around 0.5, meaning $H_0 \approx 50$ km/s/Mpc.

The parameter $h$ is often used in conjunction with other dimensionless cosmological parameters, to make them valid for any value of the Hubble constant $H_0$, especially when such parameters are extracted from observational data. More about that later.

It must be noted that the Hubble constant is not necessarily constant, because it must have been much higher in the past. So $H_0$ is referred to as the present Hubble constant or also the local Hubble constant. At other times, the value is denoted by just $H$ or by $H(t)$, meaning the 'time varying Hubble constant'.

In the Einstein-de Sitter expansion model, the mutual gravity of all the matter in the universe must be balanced by the expansion rate of the entire universe. this must be done in such a way that the expansion rate is just high enough to prevent an eventual re-collapse of the universe.

This requirement demands a very specific expansion law. In imitation of the escape velocity: $dr/dt = \sqrt{2GM/r}$, the rate of change of the expansion factor equals

$$\frac{da}{dt} = \sqrt{\frac{H_0^2}{a}}, \tag{13.1}$$

if appropriate units are chosen. $H_0$ is the Hubble constant and $a$ the dimensionless (time varying) expansion parameter* at time $t$. Here $H_0^2$ is equivalent to $2GM$, a constant energy, because $H_0$ represents the 'velocity' of expansion and energy is proportional to velocity squared.

*Recall that $a$ is defined to be unity at the present time and 0.5 when the observable universe was half it's present size.

This is the *expansion law for the Einstein-de Sitter universe*. It can also be written as

$$dt = \frac{a^{\frac{1}{2}}}{H_0} da$$

and integrated against $a$, to find

$$t = \frac{1}{H_0} \int a^{\frac{1}{2}} da = \tfrac{2}{3} \frac{a^{\frac{3}{2}}}{H_0} + \text{ constant.}$$

It is assumed that $t = 0$ for $a = 0$, making the constant of integration zero. In standard textbooks, e.g., [Peebles, Peacock], this relationship is usually written as

$$H_0 t = \tfrac{2}{3} a^{\frac{3}{2}}, \tag{13.2}$$

which was used to plot the expansion curves of the preceeding figures of this chapter.

Since $a$ is a dimensionless parameter, the above implies that $H_0$ has the units (time)$^{-1}$. If time is measured in Gy, then $H_0$ has the units Gy$^{-1}$.

The conversion from the standard units for $H_0$ (km/s/Mpc) to Gy$^{-1}$ is a factor $\frac{1}{978}$ in appropriate units. Cosmologists seem to happily live with this 'dual use' of the symbol $H_0$, but engineers normally don't.

The author prefers to show the difference more clearly by defining a normalized Hubble constant

$$\bar{H}_0 \equiv \frac{H_0}{978} \text{ Gy}^{-1},$$

as will be used in equations in the rest of this book. Because at present, $a = a_0 = 1$, equation 13.2 immediately gives us the present age of the universe, as predicted by the Einstein-de Sitter model:

$$t_0 = \tfrac{2}{3}\frac{1}{\bar{H}_0} \text{ Gy}. \tag{13.3}$$

In the early days, when $H_0$ was taken as around 50 km/s/Mpc (or $\bar{H}_0 = 0.051$ Gy$^{-1}$), this gave an age for the 'standard' universe of about 13 Gy. We will see in later chapters that a new 'standard' model has been developed, giving about the same age, but with a Hubble constant around 0.07 Gy$^{-1}$.

In the Einstein-de Sitter model, this Hubble value gives an age below 10 Gy, which is incompatible with other observational data. From equation 13.2, we also have the following useful relationship

$$\frac{a}{a_0} = a = \left(\frac{t}{t_0}\right)^{\frac{2}{3}}, \tag{13.4}$$

telling us that at as we look back in time, say to when the universe was an eighth of it's present age, the expansion factor was a quarter of what it is today, because

$$a = \left(\frac{1}{8}\right)^{\frac{2}{3}} = \frac{1}{4}.$$

An expansion factor $a = 1/4$ means that the present visible universe was then one quarter of it's present size.

## 13.4 Redshift

Engineers are mostly familiar with the Doppler shift caused by objects moving through the air or through space (like radio waves). Redshift is essentially the same thing, but in cosmology it may have two origins.

The first is like Doppler shift, where a source is moving through space relative to us. It becomes a blueshift if the source is moving towards us. Then there is cosmological redshift caused by the expansion of space.

Since photons move through space as electromagnetic waves, it is reasonable to accept that their wavelengths stretch with space. If a photon was emitted into space when the expansion factor was $a = 0.5$, then today, with expansion factor $a = 1$, all distances have stretched by a factor $\frac{1}{0.5} = 2$ and so has the wavelength of the photon.

If the photon had a wavelength $\lambda$ when emitted, it will now have a wavelength $\frac{\lambda}{a} = 2\lambda$. The increase in wavelength will be $\Delta\lambda = \frac{\lambda}{a} - \lambda = \lambda$ in this case, with $a = 0.5$. Expressed as a fraction of the original wavelength:

$$z = \frac{\Delta\lambda}{\lambda} = \frac{1}{a} - 1 \tag{13.5}$$

which will be unity in this case. The parameter $z$ is called the cosmological redshift We can also express the expansion factor $a$ as a function of redshift $z$:

$$a = \frac{1}{z+1} \; . \tag{13.6}$$

Equations 13.5 and 13.6 are very important relations in cosmology. Since the present value of $a$ is unity, light emitted very recently will have a redshift approaching zero.

If the expansion factor was very small when the photon was emitted, the photon's redshift would be very large, as is clear from figure 13.4. Here the redshifts for the galaxies that we worked with before (fig. 13.3) are shown against the expansion factor.

It is interesting to note that turn-of-the-millennium technology only allowed observation of galaxies up to just over $z = 6$. The only observations at significantly larger redshifts were the cosmic microwave background (CMB) radiation, weighing in at just over $z = 1000$. More about the CMB later in this chapter.

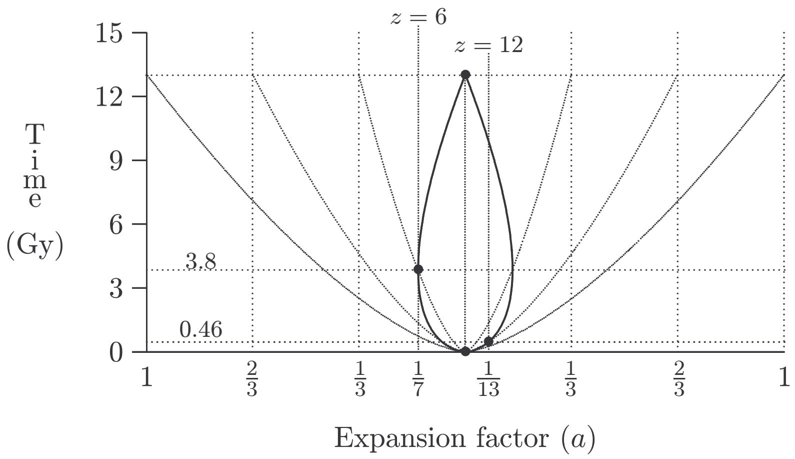


**Figure 13.4:** The expansion curves against expansion factor, allowing the redshift of galaxies to be calculated where the teardrop and the curves intersect.

## 13.5 The other 'Hubble quantities

*Hubble time* is the inverse of the Hubble constant, in appropriate units of course. Cosmologists sometimes confuse people outside their trade by stating the Hubble time as

$$t_H = \frac{1}{H_0} = \frac{9.78}{h} \text{ Gy},$$

implying, but not saying, that $H_0$ is expressed in Gy$^{-1}$ and $h = 100$ km/s/Mpc. One should rather avoid this potential confusion and use the normalized Hubble constant, simply stating

$$t_H = \frac{1}{\bar{H}_0} = \frac{978}{H_0} \text{ Gy}.$$

It is essentially the time it would have taken the present observable universe to expand from near zero size to it's present size, given that the expansion rate was always the same as today.

If we take the 'old' value of $\bar{H}_0 \sim 0.05$, it means that $t_H \sim 20$ Gy, which is a 'characteristic' timescale, but not the age of the universe. We have seen above that in the Einstein-de Sitter model, the age of the universe is two-thirds of the Hubble time.

There are more modern models that utilize the same Hubble time, but yields a different fraction than two-thirds, caused by a different expansion law. These will be dealt with later in this chapter.

*Hubble distance* is simply the distance that light can travel in the Hubble time. Because the speed of light is 1 lightyear per year, the value of the Hubble distance is the same as Hubble time if expressed in the units of Gly.

The Hubble distance, also called the Hubble radius ($r_H$), is a characteristic scale for the universe. Like in the case of the age of the universe, the radius of the observable universe is not equal to the Hubble radius.

For the Einstein-de Sitter model, it is again two-thirds of the Hubble radius, based on light travel time. Based on co-moving coordinates, the radius of the observable universe is about 40 Gly, assuming $\bar{H}_0 \sim 0.05$ Gy$^{-1}$ (or 50 km/s per Mpc) for now.

## 13.6 Look-back time

When we observe a distant celestial object, the radiation that reaches us traveled along the surface of the "cosmic teardrop". The space distance that the radiation traveled is however not the distance as measured along the surface of the teardrop, because that distance includes a component of time.

The space distance that the light traveled is called the look-back distance, which is the same as the look-back time, in appropriate units, of course.

The look-back time is $t_0 - t$, where $t_0$ is the present time and $t$ the time when the radiation left the object.

For the standard Einstein-de Sitter universe, we have seen that $a = (t/t_o)^{\frac{2}{3}}$, or $t = t_0 a^{\frac{3}{2}}$, so that

$$t_0 - t = t_0(1 - a^{1.5}).$$

Since $a = \frac{1}{1+z}$ (eq. 13.6) and $t_0 = \frac{2}{3\bar{H}_0}$ (eq. 13.3), we can express the look-back time $t_0 - t$ in terms of the redshift $z$ as

$$t_0 - t = \frac{2}{3\bar{H}_0}\left(1 - (\frac{1}{1+z})^{1.5}\right) \ , \tag{13.7}$$

perhaps one of the most used equations in 'standard' cosmology—at least in the pre-1980 era. It gives the (light travel) distance to a remote cosmic object in terms of two measurable quantities, the redshift and the Hubble constant.

Let us use equation 13.7 to calculate the look-back time to a galaxy at redshift $z = 6$, taking $\bar{H}_0 = .05$ Gly$^{-1}$.

$$t_0 - t = \frac{2}{3 \times 0.05}\left(1 - (\frac{1}{7})^{3/2}\right) \approx 12.6 \text{ Gly}.$$

## 13.7 Cosmic microwave background radiation

If the universe started out in an almost infinitely dense state (the 'big bang'), the temperature must have been extreme. As determined by modern theory, the temperature must have been in the order of $10^{15}$ degrees Kelvin, e.g., [Smoot], by the time the normal particles that makes up matter emerged.

However, during the first 400,000 years or so, radiation was so strongly coupled to the elementary particles, that the universe was not transparent. In effect, the photons were scattered by the charged elementary particles. The universe was opaque.

When the temperature dropped enough so that electrons could bind with nuclei, neutral atoms were formed, allowing radiation to become free to move through space. The universe became transparent. It is called the time of 'last scattering', with a temperature of around 3,000 degrees Kelvin.

Today, astronomers observe this radiation as the cosmic microwave background (CMB), at about 2.7 degrees Kelvin (equivalent black body temperature). For a detailed account of how the CMB was accurately charted, refer to [Smoot].

So how much has the universe expanded since last scattering? The relationship between temperature and the expansion factor happens to be linear, so that the expansion factor must have increased by about a factor $3,000/2.7 \sim 1,100$.

This gives an expansion factor at last scattering of $a_{ls} \sim 9.1 \times 10^{-4}$ and a redshift of $z_{ls} = (1-a)/a \sim 1,100$. Astronomers and cosmologists usually round this to $a_{ls} \sim 10^{-3}$ and $z_{ls} \sim 1,000$.

An interesting question: why can we continously observe the CMB, or in other words, why have those photons from the last scattering not whisked past us, never to be seen again? The answer is that there are apparently much more universe than what we can observe today.

The big-bang happened everywhere simultaneously. As time goes on, we observe further and further regions of space as they were at last scattering. If the universe is infinite in size, the CMB will never become unobservable, but it will be observed at lower and lower temperatures (larger redshift) as time goes on.

If the universe happens to be finite but not closed on itself (i.e. it is larger than today's observable universe, but it has boundaries or an edge), then there may come a time when there will be no CMB as we know it today—when the last CMB photons have whisked past us, never to be seen by us again.

## 13.8 Summary

We now have a feeling for the Einstein-de Sitter model in terms of the expansion law and Hubble's constant. We have seen how the redshift relates to the expansion factor, the age of the universe, look-back time and the CMB.

All was done in terms of the standard 'flat' model. It is now time to look at various other expansion models. As we have seen previously, the simple flat model takes as an assumption that the kinetic energy of expansion exactly balances out the potential energy caused by the mutual gravitational pull of the matter of the universe. What if they do not balance out?

# Chapter 14

# The Friedmann Equation

When the energy equation
seems
not to balance

## 14.1 A 'Newton cannon ball'

We will start this discussion by looking at an analogy with the familiar case of a cannon ball, shot up straight from the surface of the Earth.

If we ignore the drag from the atmosphere and influences from the Sun, the Moon and other planets, it is easy to write down the Newtonian equation of motion for the cannon ball. Radial velocity $\dot{r}$ changes against distance $r$ as

$$\dot{r} = \pm\sqrt{\epsilon + 2GM/r}, \tag{14.1}$$

where $\epsilon$ is a constant proportional to the total (Newtonian) energy per unit mass of the cannon ball. Total energy is made up of positive kinetic energy and negative potential energy, so $\epsilon$ can be positive, zero or negative.

It is easy to see that $\epsilon = 0$ will give the standard escape speed $\dot{r}_e = \sqrt{2GM/r}$. A positive $\epsilon$ will produce a speed greater than the escape speed and visa versa for negative $\epsilon$. So a cannon ball with $\epsilon \geq 0$ will escape from Earth and if $\epsilon < 0$, it will eventually fall back to Earth.

Now this looks suspiciously much like the case of the open universe, expanding forever, and the closed universe that will one day collapse back to it's origin. But does this simple Newtonian energy balance also hold for the expanding universe? The answer is: almost.

## 14.2 The Friedmann equation

During the 1920's, Friedmann showed that Einstein's field equations have a solution for an expanding universe that is extremely close to the 'Newton cannon ball' equation discussed above. In it's simplest form, the Friedmann equation is virtually indistinguishable from the Newton energy balance. The Friedmann equivalent* of eq 14.1 is

*This is not the 'real' Friedmann equation, which will be given shortly

$$\dot{R} = \pm\sqrt{-k + \frac{2GM}{R}}\ , \tag{14.2}$$

where $M$ is the total mass of the universe and $R$ the present radius of the entire universe (not the observable part). $\dot{R} = dR/dt$ is the 'velocity' of expansion and $k$ is a 'curvature switch', selecting between open, flat and closed universes.

The $k = 0$ case is obviously the 'flat' universe. The $k = -1$ case is an open universe with negative curvature, but positive total energy. The $k = +1$ case is a closed universe with positive curvature, but negative total energy, meaning that the expansion will reverse sometime in the future.

Cosmologists do not work with the (unknown) mass of the universe in the Friedmann equation. They replace $M$ with a function of the density ($\rho$) of the universe, i.e., $M = \frac{4}{3}\pi\rho R^3$, meaning multiplying the volume of the sphere ($\frac{4}{3}\pi R^3$) by it's mass density ($\rho$). Substituting $M$ into eq. 14.2 gives the standard Friedmann equation

$$\dot{R} = \pm\sqrt{-k + \tfrac{8}{3}\pi G\rho R^2},$$

which can also be written as

$$\dot{a}/a = \pm\sqrt{-k/R^2 + \tfrac{8}{3}\pi G\rho}, \tag{14.3}$$

since $\dot{R}/R = \dot{a}/a$, where $a$ is the expansion factor. Here $-k/R^2$ represents 'curvature density', with a sign depending on the value of $k$, and with a magnitude depending on the radius $R$.

## 14.3 The density parameter $\Omega$

Cosmologists work with a dimensionless density parameter, defined as

$$\Omega = \frac{\rho}{\rho_c} = \frac{8\pi G}{3\hat{H}_0^2}\rho, \tag{14.4}$$

where $\rho_c$ is a critical density that will produce a flat Einstein-de Sitter universe with $\Omega = 1$. For an open universe, $\Omega < 1$, which can be called an "under-dense" state and visa versa for a closed universe, which can be called "over-dense".

The Friedman equation can be worked into this most useful form [Peebles] at least for a 'matter only' universe:

$$\dot{a} = \bar{H}_0 \sqrt{1 - \Omega + \frac{\Omega}{a}}, \tag{14.5}$$

i.e., the expansion rate in terms of the Hubble constant, the density parameter and the expansion factor, all measurable quantities, at least in principle. It is mostly cast into the form

$$\frac{\dot{a}}{a} = \bar{H}_0 \sqrt{\frac{1 - \Omega}{a^2} + \frac{\Omega}{a^3}}, \tag{14.6}$$

because it is easier to understand—e.g., if the amount of matter remains constant, the matter density $\Omega$ is inversely proportional to the cube of the expansion factor, i.e., to the volume of the universe.

This is the expansion law for a 'curved' universe, where only matter density is accounted for. Note that when $\Omega = 1$, we have the expansion law of the Einstein-de Sitter model.

The quantity $1-\Omega$ is sometimes called the 'curvature parameter', i.e. $\Omega_R = 1 - \Omega$. (See e.g., [Peebles]). When $\Omega_R = 0$ there is no curvature and $\Omega_R$ goes positive and negative corresponding to positive or negative curvature.

Equation 14.5 was used to plot the curves in figure 14.1, not directly, but by numerically integrating $dt$ with respect to $a$ from $a = 0$ to $a = 1$. Since $\dot{a} = da/dt$, the integral is

$$t = \frac{1}{\bar{H}_0} \int_0^1 \frac{da}{\sqrt{1 - \Omega + \Omega/a}}. \tag{14.7}$$

A "middle of the range" Hubble constant of 64 km/s/Mpc were used.*

The 'closed', 'flat' and 'open' curves tell us some important things. Firstly, the age of the universe as predicted by the standard cosmology can be read off the graphs for the three curves: about 8.6 Gy for the closed curve, 10 Gy for the flat curve and just over 12 Gy for the open curve, consistent with a Hubble constant of 64 km/s/Mps.

Secondly, all three curves cut the line $a(t) = 1$ with identical slopes. This must be so, because they were all calculated for the same $H_0$, which is the slope of the curves at $a(t) = 1$.

Thirdly, the rate of change of the slopes is the highest for the closed model, explaining why it is the one that will eventually return to $a(t) = 0$. The closed model starts expansion faster than the others, but being more dense, the expansion eventually stops and gravitational collapse follows.

*It seems that at the time of writing, the 'best fit' value of $H_0$ is 72 km/s/Mpc.

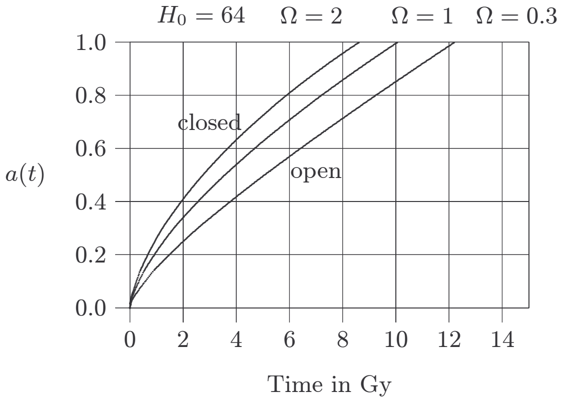


**Figure 14.1:** The expansion factor $a(t)$ against time for different values of $\Omega$. The 'closed' curve is a section of an ellipse and the slope of the curve will eventually become negative and return to $a(t) = 0$, i.e., gravitational collapse. The 'flat' curve is a section of a parabola and the slope will tend towards zero as time tends to infinity. The 'open' curve is a section of a hyperbola and will have a positive slope forever.

## 14.4 Density accounting

At this point it is appropriate to introduce the fact that energy density need not necessarily be made op by mass alone. Since mass is the same thing as energy, all forms of energy must be accounted for. Cosmologists use three forms of energy in their "accounting": mass energy, radiation energy and vacuum energy. The respective contributions to $\Omega$ are denoted by $\Omega_m$, $\Omega_r$ and $\Omega_v$.

For the present value of $\Omega$, the accountant's job is simple, because $\Omega = \Omega_m + \Omega_r + \Omega_v$. If however, we want to refer to the value of $\Omega$ at some other epoch, it is a bit more involved.

The density of matter decreases inversely proportional to the cube of the expansion factor (i.e., $a^3$). The radiation energy density decreases inversely proportional to $a^4$, being reduced by both the expanding volume and the redshift. Vacuum energy density remains constant because as more space is created by the expansion, so more vacuum energy can be created in a linear fashion.*

*That is if the vacuum energy is not exactly zero at all times.

This gives $\Omega$ as a function of the expansion factor

$$\Omega(a) = \frac{\Omega_m}{a^3} + \frac{\Omega_r}{a^4} + \Omega_v \ , \tag{14.8}$$

where $\Omega_m$, $\Omega_r$ and $\Omega_v$ is the values of the present time. This illustrates the epoch dependency of $\Omega$ on the various factors making it up.

When the universe was very young and small ($a \ll 1$), the radiation component $\Omega_r/a^4$ dominated the accounting. But since $\Omega_m$ is much larger than $\Omega_r$, as soon as $a$ grew larger, the mass component $\Omega_m/a^3$ started to dominate. And if the vacuum energy density $\Omega_v$ has even a very small value, then as time goes on and $a \gg 1$, $\Omega_v$ will start to dominate the accounting.

We can now rewrite the Friedmann equation (eq. 14.6 above) in terms of the full $\Omega$ as

$$\dot{a}/a = \bar{H}_0 \sqrt{\frac{1-\Omega}{a^2} + \frac{\Omega_m}{a^3} + \frac{\Omega_r}{a^4} + \Omega_v} \ , \tag{14.9}$$

where $\Omega = \Omega_m + \Omega_r + \Omega_v$. Remember that all the $\Omega$'s have their present values, with the time dependencies sorted out by the $a^n$ denominators.

Equation 14.9 is the general expansion law for most of the history of the universe.* Like before, the inverse of this equation can be numerically integrated to find the expansion curve for any makeup of $\Omega$.

All the above "accounting" for $\Omega$ seems to complicate matters considerably. And so it does, making it almost impossible to perceive how a "flat" $\Omega = 1$ can be maintained. Fortunately the present day contribution of radiation energy is extremely small. It was only in the first one ten-thousandths or so of the age of universe that radiation energy was important. After that, matter density dominated radiation density almost totally.

Vacuum energy is not as clear-cut as that, because as we have seen above, it does not get diluted by the expansion of the universe. The question is, does vacuum energy make any significant contribution to the total energy density, at least at the present time?

We will return to this question shortly, but first we will look at the question of the stability of the "flat" state, in other words, is $\Omega = 1$ a stable condition?

## 14.5 The stability of $\Omega = 1$

We have seen in the previous sections that for the universe to be flat, i.e. expanding just fast enough to prevent collapsing back into a singularity, the total of all contributions to the density parameter $\Omega$ must add up to unity. But to maintain $\Omega = 1$ needs a balancing act about as difficult as making a pencil stand on it's fine point on a flat table top. The smallest perturbation from unity will make $\Omega$ evolve away from that value as the universe expands.

Cosmologists are reasonably confident that the present value of $\Omega$ falls in the range 0.93 to 1.07. Then they calculate back to close to the big bang and find that $\Omega$ must have then been within a factor 1 in $10^{60}$ from unity then. If not, $\Omega$ would have evolved to lie outside the observational range today.

Because this is such an important aspect of cosmology, we will briefly examine why and how $\Omega$ evolves away from unity, once it deviates from that value.

A slightly imperfect analogy is the case of the cannon ball shot radially away from Earth where friction with the atmosphere is ignored. If the initial

*It does not account for 'inflation', which is the topic of the next chapter.

radial velocity is exactly equal to escape velocity $\sqrt{2GM/r}$, then the radial velocity will decrease, but it will always be equal to the escape velocity for the distance at which the cannon ball finds itself.

If the initial radial velocity was ever so slightly smaller than the escape velocity, the cannon ball will slow down too fast until it eventually starts to drop back to Earth again. In other words the velocity will eventually drastically diverge from the escape velocity.

For the universe at large, the same effect is at work. Slightly too small an expansion rate will cause the expansion rate to diverge away from the 'critical expansion rate' $H_c$, which can be expressed as follows:

$$H_c = \frac{\dot{a}_c}{a} = \sqrt{\frac{8\pi G}{3}\rho_c} \; . \tag{14.10}$$

An expansion rate slower than $H_c$ will cause the expansion to stop and contraction to take over. An expansion rate higher than $H_c$ will cause the density to drop too fast and the density parameter will diverge to zero.

Only if $H = H_c$ and $\Omega = 1$ will the critical balance be maintained indefinitely. It is however not possible to establish (by observation) whether $\Omega = 1$ precisely, due to inevitable uncertainties in the observations. Cosmologists come quite close, however.

## 14.6 The energy of the vacuum

The energy of the vacuum was first utilized by Einstein in his efforts to make his own equations of general relativity compatible with a static universe. Even standard Newton cosmology would not allow such a static universe, because the mutual attraction of all the matter in the universe would surely make it to contract. That is unless the universe as a whole is rotating relative to some or other absolute reference frame, which is apparently not the case.

The believe that the universe was static was however so strong in his time, that Einstein brought in a *cosmological constant* into his equations to act as a repulsive force, thus preventing his static universe from contracting.*

*The cosmological constant is proportional to the energy parameter of the vacuum: $\Lambda = 3\bar{H}_0^2\Omega_v$.

Later, when Hubble and others proved that the universe is actually expanding, Einstein repudiated the cosmological constant in his equations.

But Einstein's declaration that it was "the biggest blunder of his scientific career" did not make the cosmological constant go away completely. It was toyed with from time to time and today it is a hot topic in cosmology.

Why? Firstly, astronomers cannot find enough mass or other forms of energy to balance the expansion equations. Secondly, one of the best theories that explains how the universe got to be expanding in the first place, *inflation theory*, needs vacuum energy. Thirdly, it appears that the universal expansion is not slowing down as it "properly" should.

It may even be accelerating, if one takes recent observations at face value. We will return to inflation and possible acceleration in the expansion rate later, but first we need to get a feeling for what this *energy of the vacuum* is.

A "feeling" is about all that is within the scope of this book. Vacuum energy falls within the realms of quantum theory, which is a pretty complex subject.

The quantum vacuum is not an empty place. There are fluctuating quantum fields with all possible wavelengths that move in all possible directions. If averaged over time, these fields cancel out and we have a classical vacuum—resembling what we think of as empty space, with zero average energy.

If however the fields do not cancel out, we have, in the jargon of physics, a *false vacuum*. Over short periods of time, the quantum fields do not have to cancel out and if the resultant field is positive, then according to quantum theory, this positive field can act on matter as a repulsive cosmological force, something like "anti-gravity".

In an expanding universe, such a force extracts energy from the vacuum and converts it into additional kinetic energy of expansion. The vacuum must therefore end up with nett negative energy. This negative energy of the vacuum produces a contractive cosmological force, balancing the extra kinetic energy.

In this way, vacuum energy can work both ways (repulsive and contractive) at the same time. It is the negative component of the vacuum energy that is added into the 'accounting' equation for $\Omega$ (eq. 14.8 on page 186). In this way, the positive kinetic energy of expansion is precisely balanced by the negative contractive energy, maintaining a 'flat', $\Omega = 1$ universe.

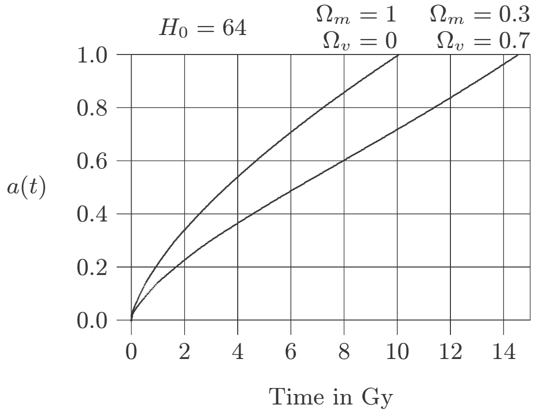


**Figure 14.2:** The expansion factor $a(t)$ against time for the $\Omega_m = 1$ and the $\Omega_m + \Omega_v = 1$ cases. The left curve has the same form as the 'flat' case in figure 14.1. Note how much older the universe with appreciable vacuum energy is—more than 14 Gy. ($H_0 = 64$ km/s/Mpc).

## 14.7 Summary of Friedmann model

In summary then, if the expansion energy accounting book does not balance, we have either an open or a closed universe. If the book balances, even if the rate of universal expansion is increasing due to vacuum energy, we have a 'flat', $\Omega = 1$ universe.

Recent observations seem to indicate that we live in such a flat universe, possibly dominated by vacuum energy already. Figure 14.2 shows a comparison between the expansion curves for a 'normal' flat universe and a flat universe with vacuum energy operating already.

It is reasonably assumed that $\Omega_r \approx 0$ today and that $\Omega = \Omega_m + \Omega_v = 1$. A Hubble constant of 64 km/s/Mpc was used, although the latest indications are that a value of 72 is the 'best fit'. This brings the age of the universe down to about 13.6 Gy.

In the next chapter we will see that vacuum energy could have been the mechanism that have started the whole process of universal expansion in the first place—the so called "inflationary big bang".

# Chapter 15

# Inflation

an engineer's view
of
inflationary expansion

If the Friedman model is taken to it's lower limit (when $a \to 0$), the expansion rate becomes extremely large. From what we have discussed in the previous chapter and specifically equation 14.9, it is clear that under this condition the radiation density ($\Omega_r$) dominates the expansion equation, i.e.

$$\dot{a}_{(r\to 0)} \to aH_0\sqrt{\Omega_r/r^4} \to \infty.$$

This very fact presents a puzzling problem to cosmologists; not so much the very rapid expansion rate, but rather the so called 'horizon problem' that it creates. In short, the horizon problem means that if space expanded at a tremendous rate right from time zero, light and other radiation that moves *through* space could not have kept up with the expansion—only parts of space that were in very close proximity to each other could have exchanged any sort of radiation information.

However, if we look at space today we find that the cosmic microwave background radiation has very, very close to the same temperature everywhere. The question is, how did parts of space that could not (ever) have exchanged radiation, ended up with this uniform temperature?

## 15.1 Guth's insight

In 1979 Alan Guth published a ground breaking paper that showed how the very, very early universe could have started with a very slow expansion rate and then went through a period of exponentially increasing expansion.* After some time the exponentially increasing expansion rate went over to

*A good summary is found in [Guth].

the conventional Friedman expansion, where the rate of expansion started to slow down. In short, he introduced a time variable cosmological 'constant' that acted like an anti-gravity force that could have 'inflated' the universe at a tremendous rate.

A phase transition then stopped the inflationary epoch and the normal 'flat' expansion dynamics took over. Later analysis showed up flaws in the argument,* but the main elements survived in essence. A physical treatment of the subject of inflation is outside the scope of this book, firstly because it is technically very intimidating and secondly because the experts do not seem to agree on the model.

*It had to do with the 'first order' phase transition that Guth proposed and was replaced by Linde (1982,1983) and also by Steinhardt (1982) with a 'second order' phase transition.

In the standard Friedman model there is general agreement on the model and just disagreement about some of the parameters in it. For the inflationary epoch, however, there is no agreement on the model's characteristics, let alone it's parameters. In this chapter, we will attempt to add some value for the reader by using an engineering approach—take the simplest model of the physicists at face value and look at it's implications.

## 15.2 The simplest model

In it's simplest form, inflation theory postulates that the time varying Hubble 'constant' $H(t)$ might actually *have been constant* [Peebles, page 407], in the very early universe, i.e.,

$$H(t) = \dot{a}/a = \text{constant}, \qquad (15.1)$$

meaning that the expansion rate $\dot{a}$ was directly proportional to the expansion factor $a$. Now this may look innocent enough, but a simple 'everyday' example shows that it is not so ordinary.

Make a spacecraft accelerate from rest in such a way that it's velocity $\dot{s}$ is always directly proportional to $s$, the distance traveled, i.e., $\dot{s} = ks$, where $k$ is some positive constant. For simplicity, let $k$ be unity, so that at one meter distance, the speed must be 1 m/s, after 10 meters, 10 m/s and so on.

From this we can suspect that we are dealing with an exponential situation. The faster the craft travels, the less time it takes to traverse 1 meter distance and the less time it has to make that next 1 m/s speed change. This relationship between distance and time is indeed exponential and can be expressed as

$$s = be^{\alpha t}, \qquad (15.2)$$

where for this example, $\alpha = 1$ and $b$ is an arbitrary constant (a scale factor), which cannot be determined from the information we have.* The curve is shown in figure 15.1, where $b = 10^{-6}$ was chosen, so that it gives a convenient scale.

Spacecraft distance vs time

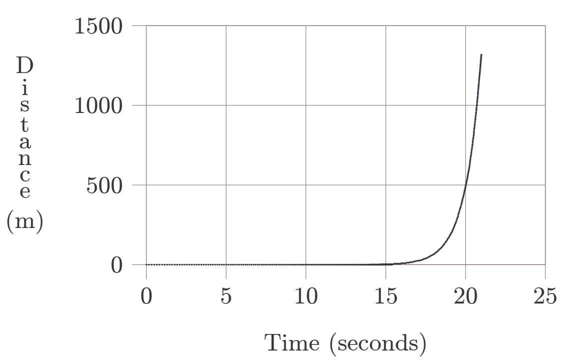


**Figure 15.1:** A plot of the distance that the spacecraft traveled against time if it's speed has to be directly proportional to the distance traveled. Note how close the distance (and by definition, also the speed) remains to zero, up to about 15 seconds and then quickly increases until it will approach infinity as times goes on.

The distance remains small for the first 15 seconds or so and then diverges relatively quickly to a large value, increasing ever faster. The 'time constant for divergence' (which we shall call $t_d$) is obtained by taking the natural log of the right hand side of the equation and setting it to zero, i.e. $\ln(be^{\alpha t_d}) = \ln(b) + \alpha t_d = 0$. Therefore

$$t_d = \frac{-\ln(b)}{\alpha} = \frac{-\ln(10^{-6})}{1} = 13.82 \text{ sec.}$$

In a way the spacecraft's speed control starts to go unstable at 13.82 seconds into the mission.

Obviously, a real spacecraft can never behave exactly like this, no matter how badly us engineers design the stability of the control loop. We just do not have the propulsion system to cope with this instability! But according to inflation theorists, the very early universe could have had the required energy of 'propulsion'—at least for a *really* short time.

Assuming that equation 15.1 above is valid, cosmological inflation can be viewed as analogous to the 'unstable' spacecraft—after some time of slow expansion there will be an exponential increase in the expansion rate. Let $r$ be the spatial radius of a spherical piece of space that represents our observable universe at a time near $t = 0$. By analogy to equation 15.2 above, the radius $r$ will change with time $t$ as

$$r = be^{\alpha t}, \qquad (15.3)$$

*Many different initial accelerations can satisfy the conditions set.

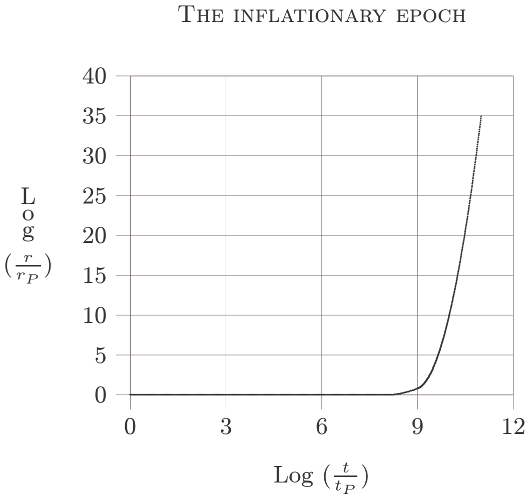


**Figure 15.2:** A log-log plot of $r$ against time for the inflationary epoch, where the expansion rate ($\dot{r}$) is proportional to the radius ($r$). Like in the case of the 'unstable' spacecraft, the expansion goes unstable rather rapidly. $r_P$ is the Planck length and $t_P$ the Planck time.

where $b$ is some constant (a scale factor) and $\alpha = \sqrt{\frac{8}{3}\pi G\rho_v}$, the energy density of the vacuum, which is assumed to be constant.

Figure 15.2 shows a plot for the first 12 time units or so. The Planck time ($t_P \approx 10^{-43}$ seconds) is used as unit time and the Planck length ($r_P \approx 10^{-43}$ lightseconds) as unit length.

We assume that at time $t_P$ the radius $r$ was equal to $r_P$, the smallest size that has any physical meaning. It looks much like the plot for the spacecraft, even though this plot is done on a log-log scale. See box on page 201 for more on the units and constants used.

We find that in the first 9 time units or so, the expansion rate was very slow. Then at about 9 time units the expansion rate 'exploded'. At about 11 time units, according to the inflation theorists, the extreme input of vacuum energy stopped because the false vacuum went through a phase transition back to the classical vacuum.

The expansion then takes on the classical shape, where the expansion rate slows down under the mutual gravitational attraction of the matter and radiation.

Since the expansion rate was very slow near time zero, it allowed all parts of the small sphere to be 'causally connected'. This means that any influence, propagating at the speed of light, could reach from one end of the sphere to the other end—in fact it could do so many times over. This made the sphere very homogeneous and having the same temperature everywhere (almost*

*Quantum fluctuations is thought to have caused small deviations from the mean temperature.).

Then, in the incredibly small time interval of about $10^{-32}$ seconds, the radius $r$ grew from less than $10^{-28}$ metres to in the order of 1 metre. The expansion rate $\dot{r}$ reached something in the order of $10^{24}$ times the speed of light.*

> *The expansion rate of space is not constrained by the speed of light—only movement through space is constrained.

The small quantum fluctuations in the mean temperature and density that existed at 9 time units were magnified immensely and due to the tremendous rate of expansion, no further causal connection between regions were possible. Light and other influences operating at the speed of light could no longer 'smooth out' the temperature of the sphere. This left just enough 'lumpiness' in the density to enable matter to congeal (under mutual gravitational attraction) into the sort of structures that we observe today.

If inflation continued for much longer at this rate, then $\dot{r}$ (and $r$ for that matter) would have diverged quickly to (near) infinity and there would have been no structures like galaxies, clusters etc. But apparently the universe was saved from that fate by the fact that it cooled down during this massive expansion.

Theory has it that the vacuum 'supercooled' and then smoothly, but rapidly 'froze' into the classical vacuum, somewhat like supercooled water freezing abruptly into ice, where a lot of latent energy is released. In the case of the universe, that energy went into the creation of the elementary particles of matter. There was then only radiation, electrons and quarks, the latter being the building blocks of protons and neutrons.

Broadly speaking, this (possibly) is the expansion history of the universe for the first $10^{-32}$ seconds of it's life. At that stage the present observable universe had a size measured in metres and was expanding at one tremendous rate.

It was filled with all the radiation energy and the mass energy of the building blocks of the matter that we experience today. In the absence' of additional vacuum energy, the expansion rate $\dot{r}$ started to slow down, due to the extreme gravitational field caused by the energy content.

The expansion rate was however precisely balanced to the energy content of the early universe. Let us take the full expansion law, equation 14.9 of the previous chapter, repeated her for convenience:

$$\dot{a}/a = \bar{H}_0\sqrt{\frac{1-\Omega}{a^2} + \frac{\Omega_m}{a^3} + \frac{\Omega_r}{a^4} + \Omega_v} \ .$$

Now simplify it fully for the conditions just after inflation, i.e., $\Omega = 1$, $\Omega_v = 0$. Also, since $a \ll 1$,* we can take $\Omega_r/a^4 \gg \Omega_m/a^3$, leaving

> *Just after inflation ended, the expansion factor was $a \approx 10^{-27}$.

$$\left(\frac{\dot{a}}{\bar{H}_0}\right)^2 \cong \frac{\Omega_r}{a^2}. \qquad (15.4)$$

The left side of the equation is proportional to expansion energy 'density', because expansion rate squared is proportional to energy (think about $\frac{1}{2}mv^2$) and $\bar{H}_0$ is the present expansion rate. This says that the early expansion energy density was proportional to the radiation energy density at that time.

But, how could radiation energy act as the 'braking' force, working against expansion, in a universe consisting mainly out of radiation? The answer is simply that any form of energy, be it mass energy, pressure energy, radiation energy, or whatever, will cause gravitational attraction.

Figure 15.3 shows the radius of our observable universe against time, in a log-log plot from the 'beginning' up to now. This type of plot magnifies the early times tremendously and compresses the present epoch. Only a small section on the right hand side of the plot represents the epoch from 'last scattering' (at ≈ 300,000 years) up to now (i.e., the visible universe). But it shows the inflationary epoch, with it's exponential expansion dramatically.

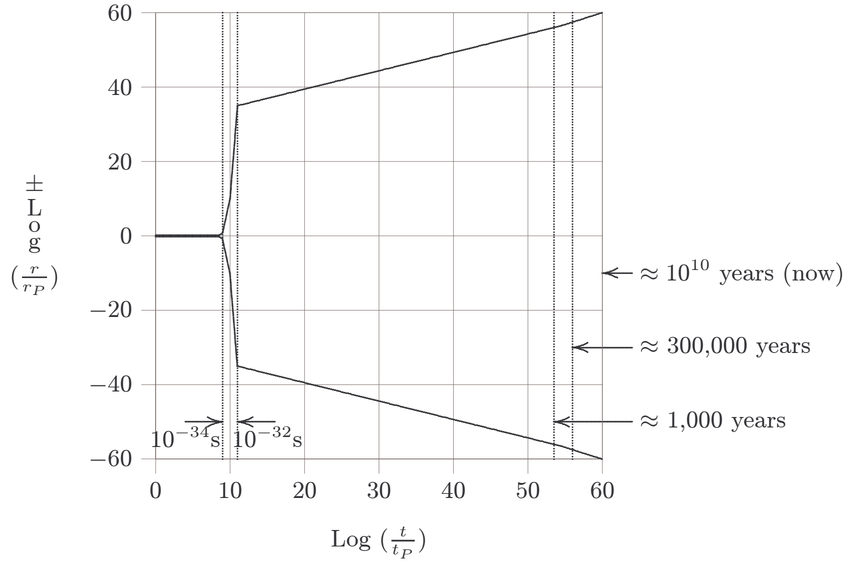


**Figure 15.3:** A log-log plot of radius of the observable universe against time from one Planck time unit until today. The rapid inflation lasted from about $10^{-34}$ to $10^{-32}$ seconds. After that the universe was radiation dominated up to about 1,000 years. Then the expansion curve changed to the matter dominated form that we presumably have today. See text for alternatives to that.

During the epoch from $10^{-32}$ seconds to around 1,000 years, radiation dominated in the density equation and the slope of the straight line is $\frac{1}{2}$, meaning $r_1/r_2 = (t_1/t_2)^{\frac{1}{2}}$. Thereafter, matter started to dominate and the slope of the line becomes $\frac{2}{3}$, meaning $r_1/r_2 = (t_1/t_2)^{\frac{2}{3}}$, the decreasing expansion rate of the standard model. This model ignores the possibility of vacuum energy becoming a factor in the density equation at a later stage.

It is important to note that in a log-log plot, a straight line does not represent linear expansion, but rather a power function, $r_1/r_2 = (t_1/t_2)^n$, where

$n$ is constant. If lines are not straight, like during the inflationary epoch, then $n$ grows with time and the situation is unstable.

When there is an abrupt change in the log-log slope, like at time $10^{-32}$ seconds, it does not mean that the expansion rate changes abruptly. It simply means accelerating expansion becomes decelerating expansion and the expansion rate $\dot{r}$ changes rapidly, yet smoothly at all times.

## 15.3 Is the expansion rate actually decreasing?

If the contribution of vacuum energy dropped to exactly zero after inflation ended (and remain zero), the log-log curve will today have a slope of 2/3 and will remain so forever, which means that $\dot{r}$ will approach zero as the age of the universe tends to infinity.*

*A constant slope of less than unity on this log-log curve represents a parabola on a linear plot.

This is however not what today's astronomers observe. There are strong indications that the expansion rate in the matter dominated phase is not dropping at the rate that the standard model requires and it may even be increasing. The only plausible explanation is that some form of additional expansion energy is at work and this could possibly come from the fact that the vacuum contribution did not drop to exactly zero after inflation. So a mild form of inflation may still be present today.

As we have seen, vacuum energy is not diluted by the expansion of the universe and (if present) will eventually dominate the expansion, causing the log-log curve to flare out as shown in figure 15.4, meaning the expansion rate increases.

An expansion rate that first decreases and then increases makes the observed higher values of the Hubble constant $H_0$ compatible with an older universe—14 Gy or more. This is old enough to accommodate the apparent age of the oldest stars.

Apart from solving the 'age problem', the accelerating expansion rate apparently has little influence on us (and our descendants) here on Earth due to the timescales involved. It looks dramatic on the log-log scale because of the compression of time on the right hand side of figure 15.4.

However, $10^{14}$ years means that the universe is then ten thousand times older than it is today, or some hundred trillion years of age. What is important though is that apparently the slope of the log-log curve has recently* started to increase and is exceeding the unity value today, meaning the cosmic expansion rate is increasing.

*Perhaps as "recently" as 5 Gy ago.

The line drawn at $10^{14}$ years is thought to be roughly the boundary between

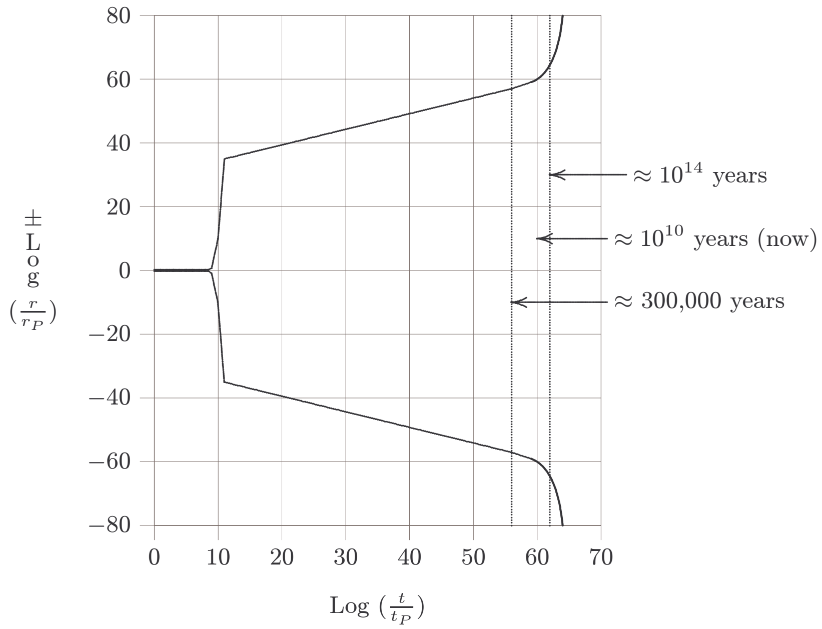


**Figure 15.4:** The radius against time if vacuum energy is still operating today. If present, the effect of vacuum energy will increase dramatically in the distant future, when the universe is thousands of times older than it is today. Theory has it that at $10^{14}$ years the universe will begin to degenerate into dead stars and black holes only.

the 'generative' and the 'degenerative' epochs. Most of the ordinary matter of the universe will become locked up in dead stars—white dwarfs, neutron stars or black holes.

Eventually, after a time greatly exceeding $10^{14}$ years, protons will decay, meaning white dwarfs and neutron stars will 'evaporate' into radiation and other elementary particles, so that only black holes will remain. An unimaginable long time after that, even black holes will become extinct due to Hawking radiation.*

*After Stephen Hawking, who proved this effect theoretically.

Because of the huge timescale, there will be no descendants of ours on Earth at $10^{14}$ years because our Sun is expected to 'flame out' in another 5 Gy or so, which is practically still on the "now" line of figure 15.4.

Should there be any observers left (somewhere in our galaxy) when the universe reaches the age of $10^{14}$ years, they will find it a much less exciting place for astronomy than it is today. All other superclusters would have drifted out of view. The increasing expansion rate would have surely taken them outside the observable universe for that time.

From the far-far future back to the present epoch. If we plot the expansion of the universe on a linear scale, the differences between a standard (no vacuum energy) expansion and a vacuum driven expansion is shown much more clearly for the present time. Figure 15.5 shows such a plot from the 'beginning' up to when the universe will be more than twice it's present age. Now the inflationary epoch and the other early epochs are lost due to the

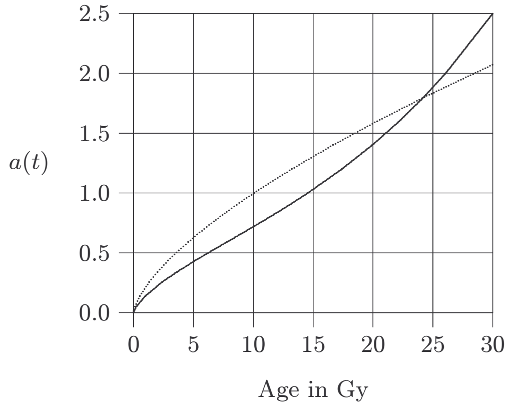


**Figure 15.5:** The solid curve plots the expansion factor $a(t)$ against time for a flat universe ($\Omega = 1$), with vacuum energy today contributing 70% of the expansion energy. For Ho = 64 km/sec/Mpc, the present age of the universe will then be near 15 Gy (read off at $a = 1$). The plot is extended to more than twice the present age to show the increasing expansion rate more clearly. For comparison, the dotted curve is for a flat universe with no vacuum energy present today. See text for more details.

resolution, but the present is shown in proportion.

The solid curve is for when today's vacuum energy component makes up 70% of the total energy needed to make the universe flat. The dotted curve (part of a parabola) is for a flat universe with no vacuum energy component operating today—i.e., there are enough ordinary matter and dark matter to make the universe flat. The two curves have both been drawn for a present Hubble constant $H_0 = 64$ km/sec/Mpc or $\bar{H}_0 = 0.065$ Gy$^{-1}$.

Of significance is the fact that at $a = 1$, the slopes of the two curves are equal, meaning that the time varying Hubble constant $H(t) = \dot{a}/a$ is the same for the two curves. So where curves cut $a = 1$, it gives the present age of the universe for the two curves respectively.

For the assumed value of the Hubble constant it gives roughly 10 Gy for the matter only universe and close to 15 Gy for the universe with 70% vacuum energy. This shows graphically why vacuum aided expansion makes a specific value of the Hubble constant compatible with an older universe.

Something that is not so easy to understand is why the vacuum aided expansion curve lies below the the matter-only (dotted) curve for much of the time. The answer lies in the fact that the inflationary epoch 'delivered' a universe with the expansion rate precisely balanced with the amount of matter (visible and dark) that caused the slowing down effect through mutual gravitational attraction. For the vacuum aided expansion curve, the amount of matter must today be less than what it would have been for a matter-only universe—only 30% for the case plotted.

So during early years, when there was negligible vacuum effect, the amount of matter to be balanced by expansion rate must have been less in the same proportion. This 'lighter' universe would initially have expanded slower than the 'heavier' one. It can be easily deduced from the expansion law equations, given earlier.

### More about the units and constants of the plots of the inflationary epoch

Working in units of seconds at these incredibly tiny time intervals is not a lot of fun. One solution is to convert time to units of Planck time ($t_P$), i.e. multiples of $10^{-43}$ seconds (the exact value is $t_P = 5.3906 \times 10^{-44}$ seconds). Time $10^{-43}$ seconds will then be 1 Planck time unit, time $10^{-42}$ seconds will be 10 Planck time units and so on. This leads to the simplest scheme of all - plot the radius $r$ against $\log(t/t_P)$. Then $10^{-43}$ seconds will be expressed as zero time units, $10^{-42}$ seconds as one time unit and so on. Since $r$ has magnitudes comparable to $t$, it is best to also express $r$ in the same type of units, i.e. as $\log(r/r_P)$, where $r_P$ is the radius expressed in Planck lengths, which is about $10^{-43}$ light-seconds. With these values it is convenient to work in log base 10 units and since we are interested in ratios, we can write the inflationary expansion equation as

$$\frac{r}{r_P} = \frac{b \times 10^{\alpha t}}{r_P}.$$

If we choose the constant $b = r_P$ (one Planck length), then $b$ and $r_P$ cancel out and we have the log-log curve for the inflationary expansion plot as

$$\log \frac{r}{r_P} = \alpha t.$$

The constant $\alpha$ depends on the inflationary model used. For the inflationary part of the plot, a value of $\alpha = 10^{34}$ per second was chosen so that time constant of inflation is $10^{-34}$ seconds (9 log-time units). This results in inflation happening between 9 and 11 log-time units, in the range given by most standard inflation models.

For the epochs after inflation ended (at 11 time units), the straight parts of the log-log curve is obtained by the slopes of the corresponding epoch: from 11 to about 53.5 time units (1,000 years), the slope is 1/2 and after that the slope is 2/3 (assuming that no vacuum energy is at work).

The approach followed here is really engineering-like! We accept what the scientists tell us without fully understanding the physics behind it and then apply it, albeit somewhat empirically! Although the analysis is not rigorous or precise, it gives us a reasonable idea of why the curves look like they do.

# Chapter 16

# Measuring the shape of expansion

a wrap-up
of
'engineering cosmology'

To cosmologists, the shape of the expansion curve is all important. If they know it accurately, they will know, amongst other things, the age of the universe. They would know the distance to objects with a given redshift pretty accurately. One can understand why cosmologists wants to know the geometry of the universe.

## 16.1 The observables

The distance to remote objects in the universe is not directly measurable. The only direct distance related observables are the redshift and the apparent luminosity of remote objects. The relationship between these two observables and distance is however not all that clear cut.

As far as the redshift is concerned, there is a possibility that part of the redshift of distant can be attributed to 'non-cosmological' effects. This means that the redshift may not be a simple function of how much the universe has expanded since the observed radiation left the distant object. The luminosity of very distant objects may be altered by partially obscuring dust between us and the object that makes it appear dimmer than it should be.

However, by looking at objects like galaxies and supernovae near and far,* astronomers are today pretty sure that they have the tools to make long range distance measurements with errors of less than 20%. In engineering

> *A supernova is essentially a massive star that ran out of fuel, causing it's core to collapse, creating a shock wave that blows the star apart.

terms, this is pretty coarse, but if that is the best we can do, we live with it.

In order to use the redshift to measure distance, one needs accurate values for the Hubble constant $H_0$ and the shaping parameters of the expansion curve, i.e., $\Omega_m$ and $\Omega_v$. The latest values (at the time of writing) for these parameters came from the Wilkinson Microwave Anisotropy Probe (MWAP). WMAP determined that $H_0 = 71 \pm 3.5$ km/s/Mpc, with 73% vacuum energy.

## 16.2 What we expect to find

Figure 16.1 shows the situation for a luminous object with an observed redshift of $z = 1$, meaning that the light we observe left the object when the Hubble radius $a(t)r_H$ was half of what it is today. The two black dots, where the 'half-size' dotted horizontal line crosses the two curves, are the solutions for where the object was for the two cases: (i) the standard flat, decelerating expansion curve and (ii) the (presently) accelerating expansion curve, with a 73% vacuum energy component.

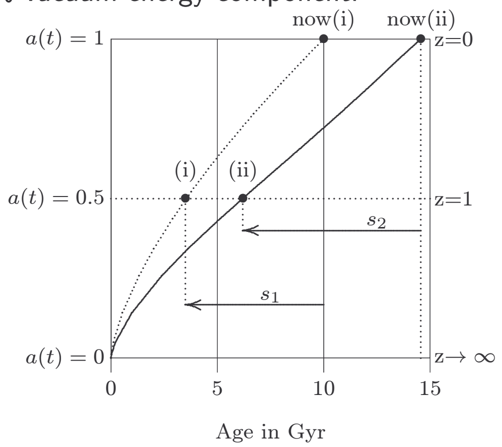


**Figure 16.1:** The two black dots are the positions for a celestial object at redshift $z = 1$. Position (i) is for a standard flat universe and position (ii) for a universe with accelerating expansion at the present time. The arrows ($s_1$) and ($s_2$) are the distances of the source from an observer living in the present time. The observer's position is at either 'now(i)' or 'now(ii)', depending on which curve (if either) is valid. This shows graphically that for a specific redshift, the distance to the object depends upon the shape of the expansion curve.

A Hubble constant $H_0$ of 71 km/s/Mpc was assumed. The two left-arrows indicate the look-back times (geometrically the same as the look-back distances) of the objects. It is immediately clear that in the case of the vacuum driven, accelerating expansion curve, the distance ($s_2$) to the object is larger than the distance ($s_1$) for the standard case. The distances can be found by numerical integration as $s_1 \approx 6$ and $s_2 = 7.7$ Gly respectively.

Now if this object had a known absolute luminosity, we could calculate the distance to the object by measuring the apparent luminosity that we observe. This would tell us which one of the two curves is the closest to the 'real thing'. If we could do that for a number of objects at different redshifts, we could in principle plot the 'real' curve and thus know the shape of the expansion.

In practice, absolute measurement of the distance to remote objects by means of apparent luminosity is fraud with uncertainties. A better approach is to measure relative distances by means of the luminosity ratio between a distant object and a nearby one—provided we know that the distant and nearby sources have the same absolute luminosities, or at least that we know the ratio between their respective luminosities.

In section 16.3, some of the measurement techniques of absolute distance (the so called 'distance ladder') is discussed in more detail. Here we will just touch upon two methods, the supernovae and the Ceiphed variable stars or*Cepheids* for short.

Cepheids are pulsating yellow giant stars radiating ten thousand times as much energy as the Sun, so they can be accurately observed at distances of more than 50 million lightyears. Further, they pulsate with a period between 3 and 50 days, depending on their luminosity—the longer the period, the more luminous.*

*Cepheid luminosity is proportional to the period raised to the power 1.3

Cepheids are also present at closer ranges, like in the Small Magellanic Cloud (SMC) and in our own Galaxy. The distances to these closer Cepheids are fairly well known from other measurements, so that an absolute distance to apparent luminosity function for Cepheids can be established to some degree of accuracy (about $\pm$ 10%).

Cepheids are not directly usable to measure the shape of the expansion curve, because they are not observable over large enough distances. They are however important in the sense that their observable range overlaps with the ranges where supernovae are reasonably abundant, like in the Virgo Cluster, some 53 million lightyears away. The Virgo cluster contains thousands of galaxies and produces enough supernovae and Cepheids that the Hubble Space Telescope can measure.

Since they are at roughly the same distance, some absolute distance to apparent maximum luminosity function can in principle be established for a specific type of supernova (the SN Ia, as will be discussed later). There is one problem though—at such distances, it is not possible to know whether the supernovae and the Cepheids are actually at the same distance from us. They appear to be part of the same cluster of galaxies, but there may be significant differences in distance.

If we combine these uncertainties with the 10% accuracy of the Cepheid function, it means using the apparent maximum brightness of SN Ia supernovae alone, we cannot presently know the absolute distance to very distant

SNe Ia to better than about 20%. We therefore have to rely mainly on relative luminosities and distances to determine the shape of the expansion curve.

The other observable parameter is the redshift of distant objects, which can be determined very accurately. However, to use redshift as a distance indicator, we need a reasonably accurate Hubble constant. Here the problem is that at the largest distance that we know with good accuracy—the Virgo cluster—we cannot measure the *cosmological* redshift all that accurately.

At the distance of the Virgo cluster, the possible relative velocities due to gravitational and other effects are of the same order of magnitude as the pure Hubble flow (the recession velocity due to the expansion alone). This means the redshift that we measure is not purely due to the expansion of the universe.

The way astronomers attempt to overcome this problem is to measure the redshift and the apparent luminosities of SNe Ia in the Coma cluster, which is some 6 times further away than the Virgo cluster. Cepheids are not detectable at such a large distance, but the larger distance makes the influence of non-Hubble velocities less than 10% of the pure Hubble flow.

By comparing SNe Ia luminosities in Virgo and Coma, cosmologists reduce the measured redshift of Coma proportionally to obtain a 'corrected' redshift for Virgo of $z = .00395 \pm 5\%$. Using the 53 million lightyears distance to Virgo determined by Cepheids, this gives a Hubble constant of $H_0 = 980 \times .00395/.053 \approx 73$ km/s/Mpc.

Many cosmologists find this value a trifle on the high side, despite the $\pm 8\%$ error margin that is claimed. They argue that there may be systematic errors (an offset) in the process used to determine the corrected redshift of the Virgo cluster. The obtained value for $H_0$ does however overlap with the presently accepted range of 55 to 75 km/s/Mpc, which is about 65 km/s/Mpc $\pm$ 15%.

So it appears that when measuring large cosmological distances absolutely, using the redshift only, we can presently do no better than $\pm 15\%$. Add to this the fact that we do not know the precise expansion curve, and the precision becomes even worse.

Forewarned by the above information, let us use the simple linear Hubble law and calculate a few characteristic distances for the universe. First let us determine how far away the object of figure 16.1, with a redshift of 1 is, using the same $H_0 = 65$ that we used in the graphs.

As any radar engineer will know, the relative velocity between the target and the radar is determined by half the fractional Doppler shift (half, because the radar signal travels to the target and back again and is thus Doppler shifted twice). The fractional Doppler shift is thus twice the relative velocity, expressed as a fraction of the speed of light: $\Delta\lambda/\lambda = 2v/c$, where $c$ is the speed of light.

In cosmology, we measure redshift only one way and therefore $\Delta\lambda/\lambda = v/c$. This is however only an approximation for relative velocities that are much,

much lower than the speed of light. Stated in another way, this is valid for fractional Doppler shifts which are much, much smaller than 1.

In the case of a larger fractional Doppler shift, the relationship of Special Relativity must be used. This gives the fractional Doppler shift as

$$\frac{\Delta\lambda}{\lambda} = \frac{1 + v/c}{\sqrt{1 - v^2/c^2}} - 1 = z,$$

where $z$ is the cosmological redshift. Due to other cosmological factors, this is not quite a valid interpretation of redshift, but it is good as a first approximation.

Note that if $v \ll c$, then $v^2/c^2$ becomes negligible and the equation reduces to the approximation $\frac{\Delta\lambda}{\lambda} \approx v/c$, as mentioned before. To extract the velocity for a given Doppler shift out of the full relativistic equation is a messy operation, but the result can be expressed quite simply as follows:

$$v/c = \frac{(z+1)^2 - 1}{(z+1)^2 + 1}. \tag{16.1}$$

This tells us the approximate recession velocity of an object with a measured redshift of $z$. In cosmology, what we really want is the relationship between the redshift $z$ and the 'look-back' distance $s$, which is essentially the time it took the light to have reached us from the time that it left the object.

If we assume a linear Hubble law $v/c = H_0 \times s$ (where $s$ is the 'look-back distance'), then $s = \frac{v/c}{H_0}$, which after substituting $v/c$ from equation 16.1 gives

$$s = \frac{(z+1)^2 - 1}{(z+1)^2 + 1} \times \frac{1}{H_0}. \tag{16.2}$$

Now we can calculate a first order distance to our galaxy at redshift $z = 1$, using a Hubble constant $H_0 = 65$ km/s/Mpc, or 65/980 per Gly in geometric units, giving

$$s = \frac{4-1}{4+1} \times \frac{978}{71} \cong 8.3 \text{ Gly.}$$

This value is a bit outside the values obtained from the graphs in figure 16.1, but not grossly so. Remember, we found 7.7 Gly for the 73% vacuum energy case, about a 7% error, showing that distance errors made by assuming a linear Hubble law out to large distances are comparable to the ± 5% error in the value of the Hubble constant itself.

The other interesting distance is the radius of the observable universe, i.e., the distance to the origin of the cosmic microwave background (CMB). The redshift of the CMB is $z \approx 1000$, making the ratio

$$\frac{(z+1)^2 - 1}{(z+1)^2 + 1} = \frac{1,002,000}{1,002,002} \cong 1$$

and the distance to the edge of the observable universe

$$s \approx 1 \times \frac{978}{71} \cong 13.8 \text{ Gly.}$$

This is just the Hubble radius for $H_0 = 71$ km/s/Mpc, which is a 'characteristic' value for the radius of the observable universe. This result is in almost perfect agreement with the 73% vacuum energy expansion curve of figure 16.1, which gave an observable universe radius of 13.7 Gly.

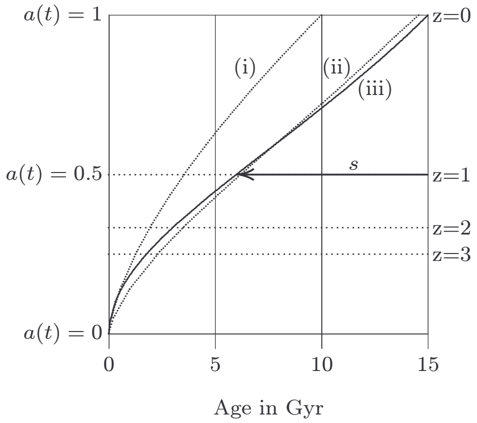


**Figure 16.2:** The solid curve (iii) is drawn for a linear Hubble law: $\beta = H_0 \times s$, where $\beta$ is the apparent recession speed and $s$ is the 'look-back' distance. The dotted curves are as for figure 16.1—(i) for no vacuum energy and (ii) for 70% vacuum energy. The correlation between (ii) and (iii) is remarkable, although it is probably purely coincidence, because a perfectly linear Hubble law all the way to the edge of the observable universe has no physical justification.

Figure 16.2 shows an expansion curve for a linear Hubble law, plotted together with the two curves of figure 16.1 (page 203). Amazingly, an expansion curve that produces a precisely linear Hubble law looks very much like the curve for 73% vacuum energy, complete with accelerating expansion in the second half of the history! The correlation is to all likelihood a pure coincidence because there is no physical justification for a precisely linear Hubble law right to the end of the observable universe.

Some readers may find it surprising that a linear Hubble law ($v/c = H_0 \times s$) does not translate to a linear expansion curve, but rather to the curved expansion of figure 16.2 (iii). The reason is that a linear expansion curve translates to a linear relationship between **redshift** $z$ and **distance** $s$ and not to a linear relationship between apparent recession velocity $v$ and distance $s$ ($v/c$ and $z$ are not the same, as shown by equation 16.1).

## 16.3 The distance ladder

To conclude this chapter on cosmological measurements, a brief overview of the main 'rungs' in the so called 'distance ladder' of astronomy is given. We will discuss each element, from the closest to the farthest briefly.

### 16.3.1 Parallax

The baseline for parallax measurement is the average radius of Earth's orbit (one astronomical unit or AU), where 1 AU $\approx 150$ million km $\approx 1.6 \times 10^{-5}$ lightyears. (Engineers would like to call this 16 $\mu$l.y.!) Annual parallax must be seen as the (plus and minus) peak deviation from the mean angle of a star during one orbit of the Earth around the Sun as shown in figure 16.3.

This deviation can be measured by ground based telescopes with an accuracy of about $1.5 \times 10^{-8}$ radians ($0.015\mu$rad) and by space based equipment, like *Hipparcos* (for Hight Precision Parallax Collecting Satellite), to some 5 times better, or $0.003\mu$rad. These accuracies are not absolute, but rather relative to some very distant ("fixed") stars, where the parallax approaches zero.

To obtain a $\pm 10\%$ accuracy, the smallest parallax that can be measured is about 10 times the accuracy of the equipment, so a practical limit for ground based equipment is $0.15 \pm 10\%$ $\mu$rad and for space based equipment about $0.03 \pm 10\%$ $\mu$rad (or $3 \times 10^{-8}$ radians).

Further, proper motions of stars in the transverse (angular) direction must also be accounted for, which is obtained by also measuring the positions of stars at the beginning and the end of one complete orbit of the Earth around the Sun. The true parallax is then zero and any deviation must be due to transverse motion of the star relative to Earth. Half of this deviation is then subtracted from the "6 months deviation" to obtain the true parallax.

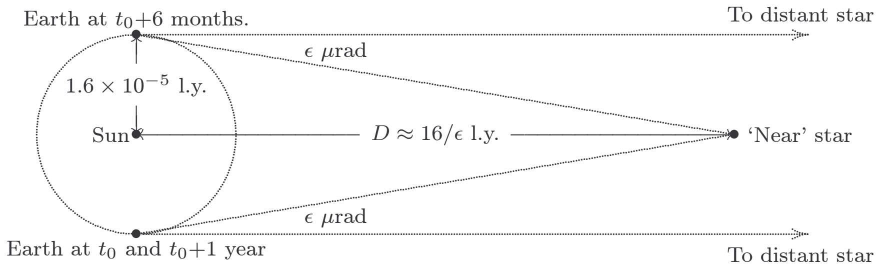


**Figure 16.3:** Schematical representation of parallax. In 6 months the parallax shift due to the Earth's orbit is actually $2\epsilon$ $\mu$rad, but astronomers prefer to relate the parallax angle to the average radius of Earth's orbit, or 1 astronomical unit (AU), which equals about 150 million km, or $1.6 \times 10^{-5}$ lightyears.

With a baseline of $1.6 \times 10^{-5}$ lightyears (the radius of Earth's orbit) and the 10% accuracy parallax limit of $\epsilon_{min} \approx 3 \times 10^{-8}$ radians peak (Hipparcos), parallax is therefore usable out to a distance of about $\frac{1.6 \times 10^{-5}}{3 \times 10^{-8}} \approx 500$ lightyears to a $\pm 10\%$ accuracy level. At double this range the accuracy would deteriorate to some $\pm 20\%$.

Special space-based equipment of the future is likely to at least double these ranges, which we will soon see to be quite important. A very long baseline array (VLBA) project has reported some parallax measurements at 100 times the Hipparcos accuracy (Harvard-Smithsonian Centre for Astrophysics, Dec 2005), but the jury is still out on the validity. If it turns out to be correct, the range of parallax measurements might be extended to some 5,000 lightyears.*

> *The problem might be that the 'right' stars may not be measurable with VLBA. See the 'standard candles' subsection below.

In engineering terms, we can say that the distance to a star is approximately $16/\epsilon$ lightyears. Here $\epsilon$ is the peak (not peak-to-peak) parallax in $\mu$rad.

Astronomers and cosmologists express parallax in *arcseconds* where 1 arcsecond $\approx 5\mu$rad. The observational limit is about $\epsilon_{min} = 0.006$ arcsecond peak.

Astronomers like to work in distance units of *parsecs*, which is simply the inverse of the parallax in arcseconds. So the equivalent distance of the furthest usable object is some $1/.006 \approx 160$ parsecs for $\pm 10\%$ accuracy. One parsec is about 3.26 lightyears.

### 16.3.2 Proper motion

The name 'proper motion' is a bit of a misnomer, because astronomers mean by that the angular change caused by the relative movement between Earth and the star during one year. It is measured at the same time on consecutive years and depends on the transverse movement of a star relative to Earth during the period of one year, as shown in figure 16.4.

The "real" motion of a star will usually include a radial and a transverse component, of which only the radial component is directly measurable by means of the Doppler shift of it's light spectrum. The angular motion of a star is also directly measurable, but to obtain transverse velocity from angular motion, one needs the distance—and that is the 'independent' variable that we are looking for!

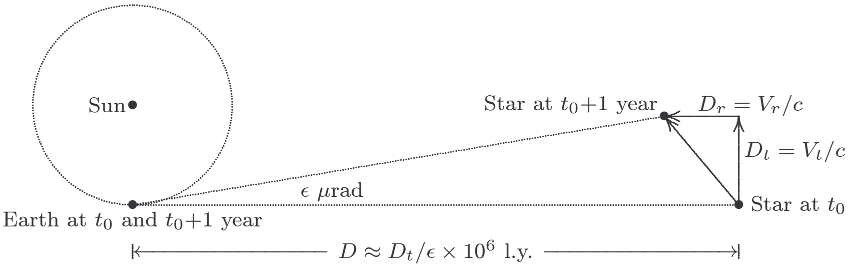


**Figure 16.4:** The geometry of proper motion distance measurement. $V_r$ is the radial velocity component and $V_t$ the transverse velocity component of a star (in km/s), $D_t$ is how far the star has moved across out line of sight in one year and $c$ is the speed of light ($3 \times 10^5$ km/s). The factor $10^6$ in the calculation of $D$ comes from the angular movement units shown ($\mu$rad).

The transverse velocity of a star is obtained by one of two methods: the *moving cluster* or the *statistical parallax* method. In the moving cluster method, astronomers obtain the precise radial velocities of the stars in an open cluster* by means of their individual Doppler shifts.

> *The nearest open cluster is the Hyades, part of the constellation Taurus, the bull. The best known one is perhaps the Pleiades or 'seven sisters', also part of Taurus.

The transverse velocities are obtained indirectly from their annual angular movement across the sky. By measuring the angular movement over many years, a "vanishing point" (as in a perspective drawing) is calculated from the combination of radial speed and angular movement.

This makes it possible to determine the approximate transverse velocity of individual stars. Once the transverse velocity $V_t$ (relative to Earth) and the proper motion $\epsilon$ per year are known, it requires only simple geometry to calculate the distance from Earth, as shown in figure 16.4.

The second method, *statistical parallax*, does not depend on the availability of open clusters, but uses the radial velocities, as obtained by Doppler shift and the proper motion angles of a large sample of stars at various distances and directions. Sophisticated statistical methods allow the most probable transverse velocity of an individual star to be determined from it's measured radial velocity. The distance calculation then proceeds as for the moving cluster method, as discussed above.

The proper motion method improves the range of the standard parallax distance measurement, because relative to Earth, stars move much faster than Earth's orbital speed around the Sun. Add to this the fact that we have a full year's movement as a baseline and not just six months as for direct parallax.

The baseline for the annual angular displacement increases by at least as factor 10, which limits the usable range to about 5000 l.y. The relationship between transverse speed and distance is $D = D_t/\epsilon \times 10^6$ lightyears, where $D_t = V_t/c$, $V_t$ the transverse speed in km/s, $c$ the speed of light ($3 \times 10^5$ km/s) and $\epsilon$ the annual angular movement in $\mu$rad.

Although less accurate than the direct parallax method, the importance of proper motion is that the usable range overlaps well with the distances where one of the "standard candles" of astronomy is found, as discussed next.

Figure 16.5 shows how the proper motion distances are connected to the parallax distances on the ladder. This is a 'minimal distance ladder'—there are many more 'rungs', but this serves to illustrate the principle.

### 16.3.3 The 'standard candles'

There are quite a few sources of radiation that have characteristic that make them good as standard sources. We will discuss only two of the most important ones here.

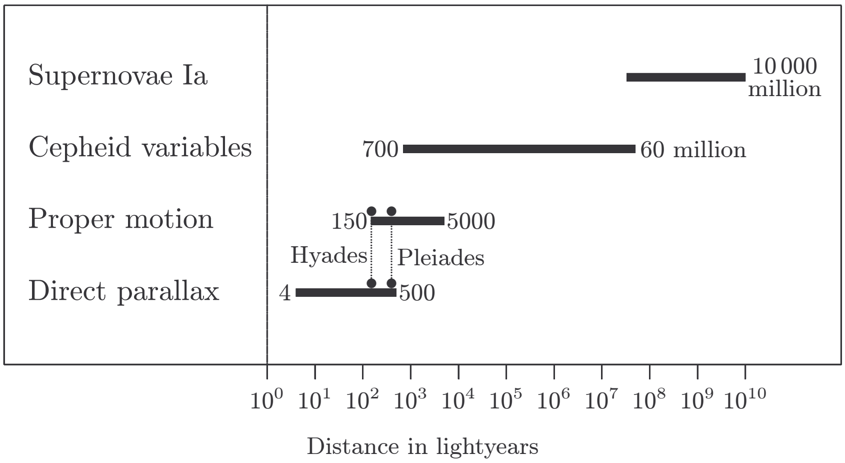


**Figure 16.5:** The minimal distance ladder showing only the 'rungs' for the proper motion/parallax connection.

**Cepheids** From a distance of 700 lightyears upwards, many so called *Cepheid variable stars* are found—the North Star (Polaris) is one of the closest known Cepheids. A typical Cepheid is a yellow supergiant star, a thousand times more radiant than our Sun.

The apparent (observed) brightness of a Cepheid varies periodically with a period that is a function of it's brightness. The brighter, the longer the period, ranging from days for a feint Cepheid to several months for the brightest ones.

Apparent brightness and distance are related by an inverse square law. So if you know the distance to a Cepheid with a specific period, you can calculate the distance to another Cepheid with that same period by comparing the brightness ratio of the two.

Further, if you know the law of absolute brightness versus the period of variation, you can in principle calculate the distance to any Cepheid from it's observed period and apparent brightness. Using Cepheids in the small Megallanic cloud (SMC), which are all at roughly the same distance from us, the law for period against brightness was found, at least in a relative sense.

Then the distance to Cepheids nearer to us (inside the Milky Way) was determined via the proper motion method. This allowed an absolute calibration for these 'standard candles' as distance measurement tools.

Due to their great brightnesses, the apparent brightness and period (and thus the distance) of Cepheids are measurable by the Hubble space telescope up to sixty million lightyears away. Figure 16.6 shows the ladder connections.

**Supernovae SNe Ia** The next rung in the distance ladder is the type Ia supernova, called SN Ia for short (SNe Ia in the plural). These monsters are thousands of times more luminous than the Cepheids and can be observed at distances hundreds of times further than Cepheids.

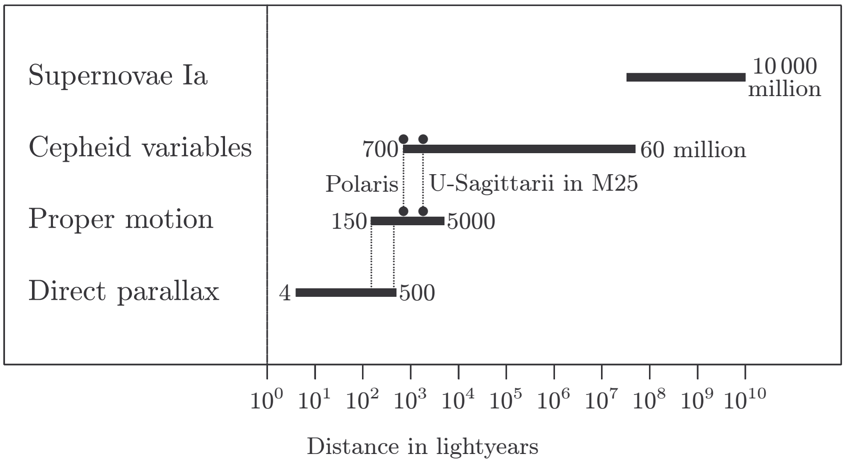


**Figure 16.6:** The minimal distance ladder with the Cepheid/proper motion connection added.

SNe Ia are thought to originate from white dwarf stars in binary systems that has nibbled away at the mass of it's companion star until the white dwarf's mass exceeds a critical value of about 1.4 times the mass of the Sun, causing it to go supernova and blow itself apart. As a result, all SNe Ia have a similar initial mass and thus a similar absolute luminosity.

By measuring the maximum apparent luminosity and the time constant of the flare up curve, astronomers and cosmologists can compensate for differences in mass, which influence the absolute brightness. Once one SN Ia is detected in a galaxy for which the distance is known through the Cepheid method, the type Ia supernova can be 'calibrated' and used as the 'standard candle' virtually up to the limit of the observable universe.

The accuracies of the SN 1a and Cepheid methods are relatively good, but at the time of this writing, they still depend for their 'calibration' on the proper motion method with it's uncertainties.

This gives an overall accuracy of no better than 20%. This is why the direct parallax measurement at larger distances by future space equipment is very important. If the parallax to a number of Cepheids can be measured directly, the accuracy of the whole distance ladder will improve. Figure 16.7 shows this last connecting rung in the distance ladder.

Some readers may wonder why the Hubble constant does not feature in the distance ladder. Is it not so that the Hubble constant can be used to determine the distance of objects for which the redshift is known?

The answer is yes, but only to an accuracy that is much less than what the distance ladder can provide. The present Hubble constant is known with an accuracy of about $\pm 15\%$, about the same as the distance ladder accuracy.

To use the Hubble constant as a distance measuring tool, one needs two parameters: the redshift, which can be measured pretty accurately, plus the shape of the expansion curve, which is not known with any certainty. This uncertainty causes additional distance measurement errors that are roughly

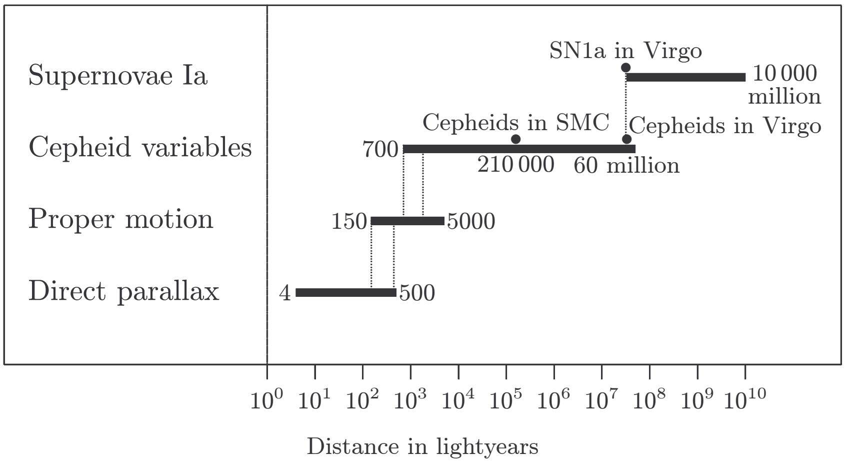


**Figure 16.7:** The completed (minimal) distance ladder, showing the last 'rung', the SN 1a/Cepheid connection in the Virgo cluster.

as large as the uncertainty of the Hubble constant itself.

In actual fact, any improvement in the distance ladder is used to pin down the Hubble constant more accurately and also to improve our knowledge of the shape of the expansion curve.

Cosmologists normally don't attach a distance to a specific redshift, but simply says that a newly discovered supernova or quasar is at a 'distance' of such and such a redshift.

At the start of the twenty-first century, the physical object with the largest measured redshift weighed in at about $z = 6$. Cosmologists can say with fair confidence that this object is at a distance of xxx from the edge of the observable universe.

However, to say how far this object is from us (based on light travel time), we need to know how far the edge of the observable universe is from us. In order to know this, we need the value of the Hubble constant and also the shape of the expansion curve, neither of which is known to great accuracy.

# Appendix A

# The Twin Paradox

where one
twin ages less
than the other one

This "perennial" is almost as old as Einstein's special theory of relativity itself. Einstein did not "invent" the paradox, although it was stimulated by his 1905 paper, where he spoke about the fact that two clocks that were separated, one staying inertial and the other one being moved away from the first one and then brought back again, will not read the same time.

The popular press explained it in terms of twins of closely the same age, where one twin sets off on a long, fast journey and eventually returns home. Special relativity then predicts that the "away twin" will be younger than the "at home" twin.

The "paradox" arises out of an abuse of special relativity's freedom to choose any inertial frame as the reference frame and make all calculations relative to it. The "abuse" on it's part arises out of choosing a non-inertial frame (the away twin) as reference and then making wrong conclusions—like that the home twin can just as well be considered as in motion relative to the away twin.

This then means that the home twin should therefore suffer the same amount of time dilation as has been calculated for the away twin and could thus be considered to end up the younger one (or at least both twins still being the same age).

To confuse the issue even further, many explanations of the difference between the two reference frames are very confusing and unconvincing—even some given in reputable technical books. Search the web for "Too Many Explanations: a Meta-Objection" (the author found it on "http://math.ucr.edu") and see for yourself.

None of the one's discussed there is fully convincing either—which may be just another "meta-objection" (what does "fully convincing" mean anyway?) Nevertheless, this book will attempt to give three explanations: (i) a simple "hand waving" argument; (ii) a very relativistic calculation and (iii) a more engineering-like representation and calculation.

## A.1 Hand-waving argument

The simplest and also the best explanation is the one based on spacetime events, like we considered before. There are four events involved: event 1 - when the away twin leaves home; event 2 - when the away twin arrives at the turn-around point; event 3 - when the away twin leaves the turnaround point on the way home; event 4 - when the away twin arrives home again.

Only the away twin is present at all four events and must therefore measure a shorter time interval between the events than the home twin. So, no paradox. True, the away twin was not in one inertial frame all the time, but the trip can in principle be made arbitrarily long, so that the turnaround time (the non-inertial part, which can be of fixed length) becomes negligible.

## A.2 Relativistic argument

The very relativistic argument views the situation quantitively from both the home and the away twin's point of view. Say the away twin flies off at $v = 0.6c$ for a coordinate (home twin frame) time of 5 years, then turns around in a negligibly short time and returns home at the same speed, for a total voyage of about 10 years of coordinate time.

The relativistic time dilation factor is $d\tau/dt = \sqrt{1 - 0.6^2} = 0.8$. This means, according to special relativity, the away twin would age only 8 years during the voyage, as illustrated in figure A.1 below.

This was the easy part. To analyse the situation from the perspective of the away twin is more difficult, since two different inertial frames are involved (actually three, if we count in the home twin's inertial frame—in a sense the turn-around phase is another frame, but it is not inertial and we choose to ignore it as insignificant).

The most comprehensible analysis is done in one of the away twin's two inertial frames, specifically the outbound frame. The home twin must then fly from the origin at a speed $v = 0.6c$ (to remain compatible with the previous analysis). The away twin remains stationary in the reference frame for 4 years, then quickly accelerate to a speed $v'$ that will allow catching the home twin at a relative speed $v = 0.6c$.

This means that, in the reference frame, we must use the relativistic addition of velocities rule to find the coordinate speed of the away twin,

$$v' = \frac{0.6 + 0.6}{1 + 0.6} \times 0.6c = 0.8824c.$$

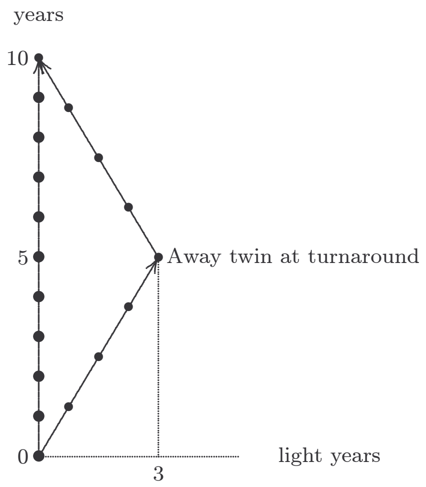


**Figure A.1:** A Minkowski spacetime diagram with the home twin's inertial frame as reference. The spacing between the bullets on the worldlines represent one year intervals according to the respective twin's clocks.

The speed difference in the reference frame is then $0.8824c{-}0.6c = 0.2824c$. The home twin will have a $0.4 \times 0.6 = 2.4$ light year head start, so the coordinate time it will take the away twin to catch up is $2.4/0.8824 = 8.50$ years, which brings the total trip time in the reference frame to 12.50 years.

Relative the reference frame, the away twin will suffer time dilation of $\sqrt{1 - 0.8824^2} = 0.4706$ during the catch-up phase. Multiply this by the 8.50 years and we get 4.0 years. So the away twin has aged a total of 8 years (4 before moving and 4 during the catch-up phase), as determined before.

The home twin, moving steadily for 12.5 years at $v = 0.6c$ relative to the reference frame, suffered a time dilation factor of 0.8 and ends up aging $12.5 \times 0.8 = 10$ years as before. Still no paradox. Figure A.2 illustrates this scenario.

## A.3 Engineering argument

The "engineering-like" calculation use electromagnetic signals and relativistic Doppler shift. Each twin sends pulsed electromagnetic signals to the other at a pulse period of one year (according to their own respective clocks). The away twin would have received 10 signals sent by the home twin, while the home twin would have received only 8 signals sent by the away twin.

This scenario is illustrated in figure A.3, a split image for clarity. This shows a crucial non-symmetry in the rates that the home twin and the away twin receives signals. Although the amount of stretch and shrinkage of the received periods are the same, the amount of time that the signals are stretched and shrunk is very different between the respective twins.

The Doppler shift ratio (period of received signal $T_r$ to period of transmitted

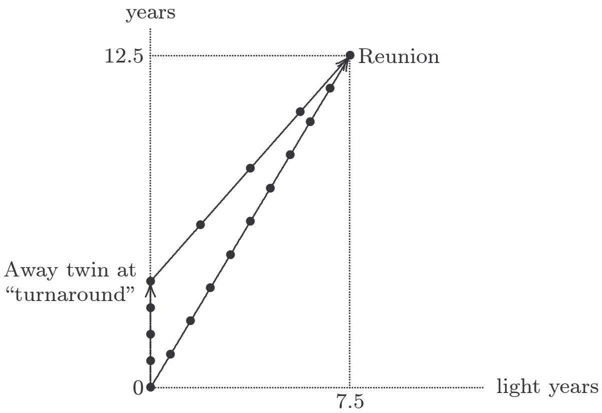


**Figure A.2:** A Minkowski spacetime diagram where the away twin is at rest in the reference frame until "turnaround". The spacings between bullets on the worldlines represent one year of propertime in the respective inertial frames. There are 10 spacings on the home twin's worldline and 8 spacings on the away twin's worldline.

signal $T$) for the outbound leg (opening velocity) is

$$T_r/T = \sqrt{1.6/0.4} = 2.0$$

and the same ratio for the inbound leg (closing velocity) is

$$T_r/T = \sqrt{0.4/1.6} = 0.5,$$

as can be clearly seen in figure A.3. The non-symmetry comes from the fact that the away twin receives compressed period signals immediately after turnaround, while the home twin has to wait until the first signal after turnaround arrives at home before noticing the event.

One can also draw the same sort of picture in one of the two inertial reference frames of the away twin. In figure A.4, the home-bound leg of the away twin has been chosen as reference. The results are exactly the same as before. There is no paradox in the twin paradox!

## A.4 Bizarre arguments

To close this chapter on the twin paradox, we will briefly look at some of the more bizarre explanations that have been offered in the literature. It arrives at the correct answer, but is somewhat meta-physical! The problem arises when the situation is analyzed from the away twin's point of view, but the two different inertial frames are pictured on the same spacetime diagram. This causes a quite understandable discontinuity in the diagram.

If you choose a reference frame in which the away twin is permanently at rest, it appears as if the home twin flies away, turns around and heads back to the away twin again. Now the away twin can 'quite rightly' assume that the home twin's clock runs slow during the outbound leg and also during the inbound leg of the journey.

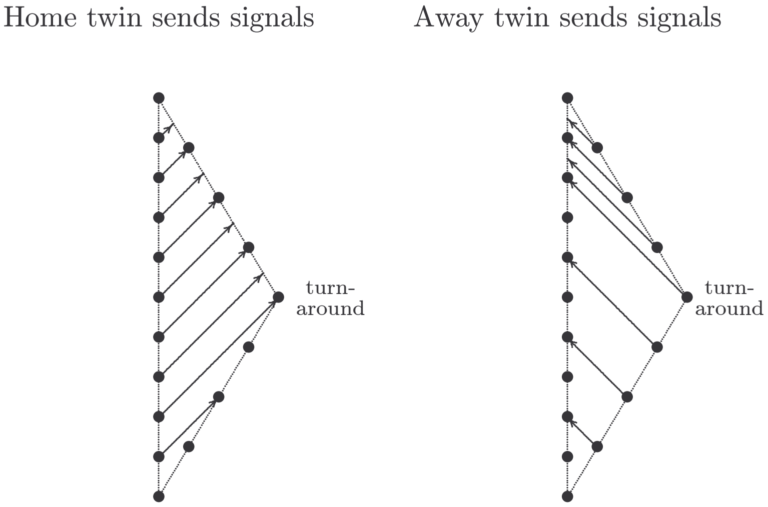


**Figure A.3:** Minkowski spacetime diagrams showing the paths of electromagnetic signals sent between the twins at yearly intervals according to their respective clocks. Note the non-symmetry between the received signals of the home twin and the away twin.

This may create the paradoxical impression that the home twin must end up younger than the away twin. The solution is that the combination of two different inertial frames onto one spacetime diagram creates discontinuity in the worldline of the home twin, as pictured in figure A.5. The "strange" part is how this is sometimes explained.

If the home twin's worldline is simply connected where the two sections cross, it appears as if the home twin has only aged 6.4 years ($8 \times 0.8$, since the time dilation is 0.8). Then there is 3.6 years "missing" in the calculation of the home twin's age.

Perhaps the worst piece of relativity that can be done is to call on general relativity's gravitational redshift/blueshift to come to the rescue. The "worst piece" goes something like this: during the away twin's turnaround, there needs to be acceleration for a finite time. Due to the equivalence principle, the away twin can be considered to be in a gravitational field.

Just like clocks far from a massive body runs faster than clocks closer to it, the home twin's clock can be considered as running faster than the away twin's clock during the acceleration (effectively a gravitational blueshift). Multiply the ratio of the clocks by the time needed for the away twin to turn around and the home twin's clock gain exactly the 3.6 years "missing" time on the "corrupted" spacetime diagram!

So what is wrong with this explanation? Well, it is not general relativistic at all—acceleration cannot be used to calculate gravitational time dilation. Gravitational time dilation scales with $\sqrt{M/r}$, while gravitational acceleration scales with $M/r^2$, where $M$ is mass and $r$ is distance. You cannot convert acceleration to time dilation, because there are many combinations of $M$ and $r$ that will give the same time dilation but different gravitational accelerations.

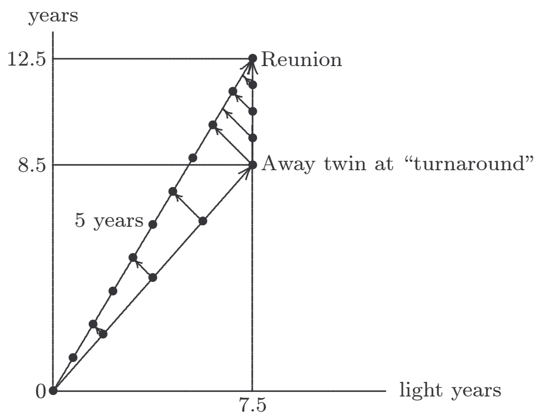


**Figure A.4:** A Minkowski spacetime diagram showing the paths of an electromagnetic signal send from the away twin to the home twin once every year of "away-time" Here the home-bound leg of the away twin has been chosen as inertial reference frame.

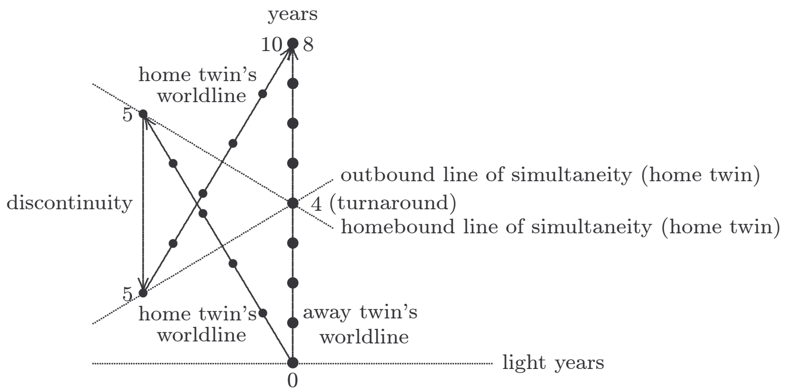


**Figure A.5:** A Minkowski diagram combining the two inertial frames of the away twin into one. The worldline of the home twin ends up with a discontinuity, as indicated.

What is true, is that the explanation simply boils down to the standard desynchronization of clocks of special relativity, based upon the relativity of simultaneity—and this cannot be used to explain differences in the ages of the twins.

Age is not something you can synchronize or adjust. It simply happens progressively and it happens differently for different inertial frames. Nevertheless, something like this very explanation is found in some reputable books, e.g., [Peacock]. As long as such explanations are offered, the "controversy" over the apparent twin paradox will probably never die.

# Appendix B

# The Sagnac effect

Refuting an
all too common attack
on relativity

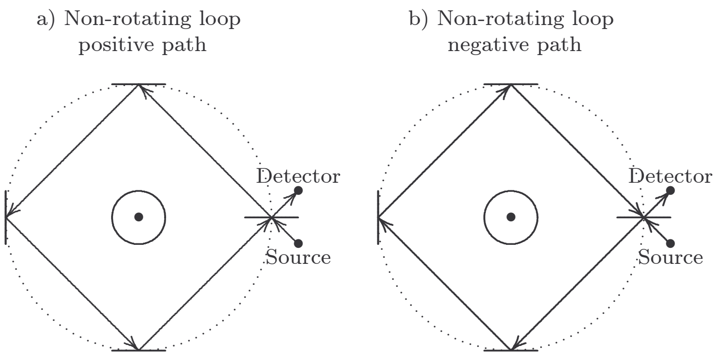
An argument that is sometimes fielded against Special Relativity, is the case of the Sagnac effect, where light is reflected (or guided by optical fiber) in both directions around a closed loop. If the loop is not rotating, the path lengths are the same both ways and there will be a specific interference pattern at the detector. Figure B.1 shows a setup for a non-rotating loop.

## B.1 The rotating loop

If the loop is rotating, there is a phase shift between the two light beams (or bundles), depending on the rotation rate. This phase shift can be measured by means of interferometry and it is the underlying principle used in fiber-optical gyros. This effect is often put forward as "...proof that light does not have a constant speed relative to the loop, as required by Special Relativity...".

The answer is that Special Relativity requires no such thing, because a rotating loop is not an inertial frame of reference. The inertial frame can only be the one in which the centre of the rotating loop is at rest—and inertial frames of reference do not rotate. It is easy to show that the phase shift comes from **different path lengths measured in the inertial frame**, as depicted in figure B.2.

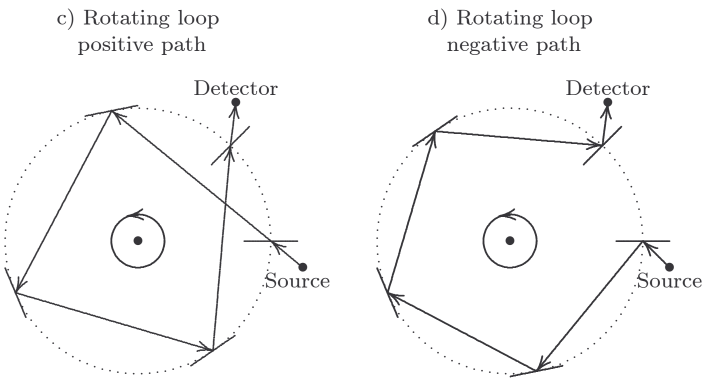
If you are still uncertain about the validity of the argument, consider this: Let the loop be non-rotating, but place (stationary) mirrors in the positions indicated in diagrams c) and d) as superimposed. Now everything is at rest relative to the inertial frame, and the path length difference is the same as



a) Non-rotating loop
positive path

b) Non-rotating loop
negative path

**Figure B.1:** The Sagnac experiment in schematic form when the apparatus is non-rotating. When superimposed a) and b) have the same path lengths and the light at the detector will register a specific interference pattern. The mirror at the source and detector is half reflective so that half the light is transmitted straight through and half is reflected into the perpendicular direction.



c) Rotating loop
positive path

d) Rotating loop
negative path

**Figure B.2:** The Sagnac experiment in schematic form when the apparatus is rotating (rotation exaggerated). When superimposed, c) and d) have different path lengths in the inertial frame and the light at the detector will register an interference pattern that is shifted relative to the pattern formed by a) and b) of figure B.1.

for the rotating loop, meaning the interference pattern will be the same as for the rotating loop.

The Sagnac apparatus simply measures absolute rotation relative to the universe at large. This rotation can be measured by any number of means, including gyroscopes and accelerometers.

The setup does not measure absolute linear motion though. The whole apparatus can be moved uniformly relative to some reference frame, at whatever speed one chooses, without rotating it. By definition of 'uniformly', it must also not be accelerated during the experiment.

The observed interference pattern will remain as for the stationary apparatus. This is because the path lengths in the two directions changes by exactly the same amount, as measured by an observer stationary in the reference frame. Actually, relativistically one should say that the path lengths do not change at all, because an observer riding on the moving apparatus cannot detect any change in either path length.

## B.2 The 'infinite' loop

Another 'good' argument that must be answered, is that one can make the size of the Sagnac apparatus arbitrarily large, and rotate it arbitrarily slowly, so that the centrifugal forces at the mirrors are negligible. Now the light travels in an 'inertial frame' (both ways).

We will however still be able to detect the rotation by means of the shift in the interference pattern. It looks like this 'proves' that light travels 'slower' (relative to the 'inertially moving' mirrors) in the direction with rotation than against the rotation.

The answer is, firstly, that the centrifugal forces at the mirrors have nothing to do with the measurement—not in Newton dynamics, neither in relativistic dynamics; secondly, the inertial, non rotating observer, sitting at the centre of the rotation, still measures a path length difference, as soon as there is any form of rotation, no matter how slow.

The bottom line still is: Special Relativity predicts the outcome of the Sagnac experiment exactly as measured, which is not surprising—there is nothing relativistic about the experiment and in such cases, Special Relativity and Newton dynamics agree.

# Appendix C

# Quasi-Newtonian acceleration

In the body of this text, various references have been made to this appendix, in support of formulas given for relativistic acceleration. In the first main section of this appendix, a (hopefully) fairly readable discussion of the formulas is given.

For those readers that like reading mathematics, the second main section derives most of the formulas from relativistic principles. The treatment does not solve the field equations, but is rigorous as far as it goes, using the Schwarzschild metric and exact solutions to the Einstein geodesic equations.

Because of the relatively complex mathematics in this appendix, fully geometrized units will be used for economy, where $c = G = 1$. This means that the $c$ will be dropped in $v/c$, so that $v$ is a dimensionless speed parameter, ranging from 0 to 1. Further, the normalized mass $\bar{M} = GM/c^2$ will be used throughout.*

> *The reader is also reminded that in this text $g_{tt} = 1 - 2\bar{M}/r = -g_{00}$ and $g_{rr} = 1/g_{tt} = -1/g_{00}$. The meaning of $g_{tt}$ is the 'time-time' coefficient and $g_{rr}$ is the 'radial-radial' coefficient of the spacetime metric.

## C.1 General discussion

Consider a test particle with rest mass $m_0$ that is accelerated by a uniform force $F$ in gravity-free space, as measured in some inertial frame. Let the particle have an instantaneous speed parameter $v$ in the direction of the force.

From Newtonian considerations, we can say that the particle accelerates at $a_0 = F/m_0$. However, according to special relativity, the particle will accelerate at

$$a_v = \frac{F(1-v^2)}{m_0} = a_0(1-v^2), \tag{C.1}$$

as is shown in more detail in section C.2.1 below. Firstly, the particle appears to have a 'moving mass' of $m_v = m_0/\sqrt{1-v^2}$, which obviously diminishes the acceleration $a_0$ by a factor $\sqrt{1-v^2}$. Secondly, the resultant acceleration is diminished by a further factor $\sqrt{1-v^2}$ due to the apparent slow running of the clock of the test particle (velocity time dilation), so that the resultant acceleration is

$$a_v = a_0(1-v^2).$$

Thus the effective acceleration achieved by the force is 'dilated' by two factors of velocity time dilation. The resultant acceleration can also be written as

$$a_v = a_0 - a_0 v^2, \tag{C.2}$$

so that viewed from a Newtonian perspective, it appears as if an *apparent* acceleration component of $a_{(opp)} = -a_0 v^2$ is opposing the expected acceleration $a_0$.

### C.1.1 Acceleration and gravity

This same situation arises when a radially moving object is accelerated by a gravitational field, as discussed in chapter 4. As measured by a local observer, an object moving with a radial velocity $v_\varphi$ relative to a mass $\bar{M}$, at distance $r$ from this mass, accelerates under gravity by

$$a_g = -\sqrt{g_{rr}}\frac{\bar{M}}{r^2}(1-v_\varphi^2) = a_0 - a_0 v_\varphi^2, \tag{C.3}$$

where here $v_\varphi^2$ is the locally measured radial velocity, $a_0 = -\sqrt{g_{rr}}\bar{M}/r^2$, the local acceleration of an object at rest relative to $\bar{M}$. According to the local observer, there is thus an apparent opposing acceleration in the radial direction of $a_0 v_\varphi^2$, due purely to the radial velocity. If we transform this apparent opposing acceleration to the distant observer, we get

$$a_{r(opp)} = g_{tt}\frac{\bar{M}}{r^2}v_r^2, \tag{C.4}$$

having multiplied by the usual 'three factors of gravitational redshift' and using the relationship $v_\varphi = g_{rr} v_r$, with $v_r$ the radial velocity measured by the distant observer (Schwarzschild coordinates).

We will see later in this appendix that the distant observer measures an additional apparent opposing radial acceleration of twice the above value, due to space curvature.

It may seem a bit surprising that the situation in a gravitational field is equivalent to that of a particle under a uniform force in gravity-free space. One must remember that we are working with the *instantaneous* velocities and accelerations here, so that the changing 'force' of gravity due to changing radial distance does not enter the arguments.

### C.1.2 Acceleration and transverse movement

The above treatment holds when the velocity $v$ is along the direction of the force, for argument's sake, both along the $x$-axis, or in a purely radial direction in a gravitational field. Now suppose that the particle is moving at an angle $\alpha$ relative to the $x$-axis, while the accelerating force $F_0$ remains along the $x$-axis, as in figure C.1. The moving mass of the particle is now determined by the velocity $v = \sqrt{v_x^2 + v_y^2}$, while the effective force is determined by the velocity $v_x$, giving the $x$-acceleration as

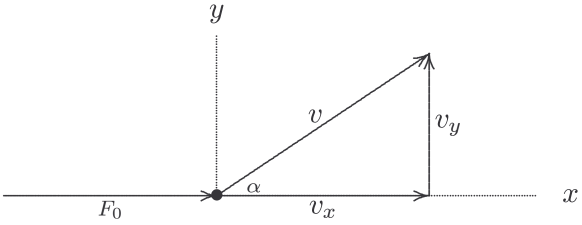


**Figure C.1:** Force $F_0$ is applied along the $x$-axis, while the velocity $v$ is at an angle $\alpha$ to the $x$-axis.

$$a_x = \frac{F_o\sqrt{1-v_x^2}}{m_o/\sqrt{1-v^2}} = a_o\sqrt{1-v_x^2}\sqrt{1-v^2}. \tag{C.5}$$

Subtract form this the Newtonian acceleration $a_o$ and we have the apparent opposing acceleration as

$$a_{opp} = -a_o(1 - \sqrt{1-v_x^2}\sqrt{1-v^2}). \tag{C.6}$$

The acceleration in the $y$ direction is zero and $v_y$ remains constant. When working with accelerations in a gravitational field, it is convenient to work in polar coordinates, with radial and transverse velocity components. In a non-circular orbit, both $v_r$ and $v_t$ are changing along the orbit, in contrast to the case above, where $v_x$ was changing, but $v_y$ remained constant.

As we have seen above, changing velocities go hand in hand with apparent opposing accelerations. The simplest way to treat this in a quasi-Newtonian fashion is to consider the component of the gravitational acceleration in the direction of movement.

If the velocity vector makes an angle $\alpha$ with the positive radial direction (along which the negative gravitational acceleration is working), the effective acceleration in the direction of movement is $-a_g \cos\alpha$, where $a_g = -\bar{M}/(r^2\sqrt{g_{tt}})$, the local static gravitational acceleration. This causes an apparent 'opposing' acceleration of

$$a_{opp} = -a_g v^2 \cos\alpha$$

in the direction of movement, similar to the opposing acceleration in the $x$ direction above. As shown in figure C.2, the projections of $a_{opp}$ onto respectively the radial and transverse axes are, in the frame of the local observer,

$$a_{\varphi(opp)} = -a_g\ v^2 \cos^2\alpha = -a_g\ v_\varphi^2 \quad \text{and} \tag{C.7}$$

$$a_{\vartheta(opp)} = -a_g\ v^2 \cos\alpha\ \sin\alpha = -a_g\ v_\varphi v_\vartheta. \tag{C.8}$$

This result may be somewhat surprising to the reader. How can a spherically symmetrical gravitational field produce a transverse acceleration? The answer lies in Kepler's second law of planetary motion: in order to sweep out equal areas in equal time, the transverse velocity must be inversely proportional to $\sqrt{r}$, where $r$ is the instantaneous orbital radius.

If there is a radial velocity component present, the orbital radius changes and so does the transverse velocity component, meaning there is an effective acceleration in the transverse direction. This is normal Newtonian orbital mechanics. The relativistic opposing components is a deviation from the Newtonian case in so far as that the faster an object goes, the more it resists the change in velocity.

When these local opposing accelerations, coming purely from velocity, are transformed to the frame of the distant observer, they become

$$a_{r(opp)} = g_{rr}\frac{\bar{M}}{r^2}v_r^2, \tag{C.9}$$

$$a_{t(opp)} = g_{rr}\frac{\bar{M}}{r^2}v_r v_t. \tag{C.10}$$

The above is very loose discussion of the transverse acceleration effect. A more formal discussion is given in section C.2.4 below. It will be shown below that space curvature produces additional apparent opposing accelerations.

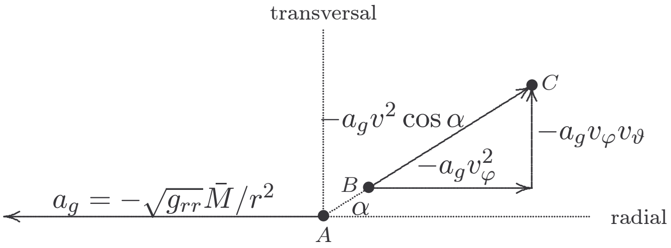


**Figure C.2:** When the velocity vector lies at an angle $\alpha$ from the radial axis, a gravitational acceleration $a_g$ along the radial produces a Newtonian acceleration of (say) vector $C \to A$ along the direction of movement. The relativistic 'opposing' acceleration is the vector $B \to C$ (not to scale), which decomposes into the radial and transverse components shown.

## C.1.3 Acceleration and space curvature

Let an object at a distance $r$ from a spherically symmetrical mass $\bar{M}$, maintain a *constant* positive local radial velocity $v_\varphi$ over a small radial distance $\Delta r$. It means that some form of propulsion is required, i.e., the object cannot be in free-fall.

If we transform this constant local radial velocity to the distant coordinate system, the result is a changing 'distant' velocity. This is caused by the spacetime curvature over the distance $\Delta r$. Recall that local radial velocities transform to the distant frame by a factor $g_{tt}$, which is less than unity in a gravitational field, but approach unity at large distances.

This means that, to the distant observer, the constant local radial velocity appears to be an acceleration in the radial direction. To obtain this apparent acceleration, we first determine the difference in radial velocity over distance $\Delta r$:

$$\begin{aligned}
\Delta v_r &= \left(1-\frac{2\bar{M}}{r+\Delta r}\right)v_\varphi - \left(1-\frac{2\bar{M}}{r}\right)v_\varphi \\
&= \left(\frac{2\bar{M}}{r}-\frac{2\bar{M}}{r+\Delta r}\right)v_\varphi \\
&= \frac{2\bar{M}\Delta r}{r(r+\Delta r)}\,v_\varphi \\
&= \frac{2\bar{M}}{r^2}\,v_\varphi\Delta r \quad (\text{if } \Delta r \to 0). \qquad (C.11)
\end{aligned}$$

If the object traveled the distance $\Delta r$ in time $\Delta t$, as measured by the distant observer, the apparent acceleration is

$$\begin{aligned}
\frac{\Delta v_r}{\Delta t} &= \frac{2\bar{M}}{r^2}\,v_\varphi\frac{\Delta r}{\Delta t} \\
&= \frac{2\bar{M}}{r^2}\,v_\varphi v_r \quad (\text{when } \Delta r, \Delta t \to 0) \\
&= \frac{2\bar{M}}{r^2} g_{rr} v_r^2 \quad (\text{since } v_\varphi = g_{rr} v_r). \qquad (C.12)
\end{aligned}$$

Since this is always positive, it means that the apparent acceleration is 'repulsive' in nature, i.e., it works in a direction opposite to normal radial gravitational acceleration. Recall that there is an apparent opposing radial acceleration of $(\bar{M}/r^2)(v_r^2)$ due to velocity alone (in the absence of space curvature). Thus the total apparent opposing radial acceleration due to radial velocity in curved space is

$$a_{r(rep)} = \frac{3\bar{M}}{r^2} g_{rr} v_r^2. \qquad (C.13)$$

Add to this the static gravitational acceleration $-\bar{M}/r^2\ g_{tt}$ and the total radial acceleration (in the absence of transverse velocity) is

$$a_r = \frac{-\bar{M}}{r^2}\left(g_{tt} - 3 g_{rr} v_r^2\right), \qquad (C.14)$$

as stated in chapter 4. Section C.2.2 below gives an alternative, rigorous derivation of the last equation above.

When the velocity is not purely radial, i.e., there are radial and transverse components present, one can do an analysis of the effects of space curvature similar to the one for radial velocities.

In imitation of the case for radial velocities, we will now assume that a constant local velocity, including radial and transverse components, is maintained over a small radial distance $\Delta r$. Along the same lines as in the case of opposing radial acceleration above, we end up with an apparent transverse acceleration due to space curvature of

$$a_t = \left(\sqrt{1 - 2\bar{M}/(r + \Delta r)} - \sqrt{1 - 2\bar{M}/r}\right) \frac{v_\theta}{\Delta t}, \tag{C.15}$$

where $v_\theta$ is the local transverse velocity, taken as positive for positive angular movement in the chosen coordinate system. It can be shown that when $\Delta r \to 0$, the factor in brackets approaches $\frac{\bar{M}}{r^2}\sqrt{g_{rr}}\Delta r$. The apparent transverse acceleration caused by space curvature is therefore

$$a_t = \frac{\bar{M}}{r^2} \frac{\sqrt{g_{rr}}\Delta r v_\theta}{\Delta t} = \frac{\bar{M}}{r^2}\, g_{rr} v_r v_t, \tag{C.16}$$

because $\Delta r/\Delta t = v_r$ and $v_\theta = \sqrt{g_{rr}} v_t$. This adds to the apparent transverse acceleration due to velocity alone, which was shown above to have the same value, yielding a total apparent transverse acceleration of

$$a_t = \frac{2\bar{M}}{r^2}\, g_{rr} v_r v_t, \tag{C.17}$$

as is more formally derived in section C.2.4 below.

It is clear that this apparent acceleration vanishes when either $v_r$ or $v_t$ (or both) are zero, as is to be expected. Otherwise, it has the same sign as the product $v_r v_t$, meaning it works in the direction of $v_t$ when $v_r v_t > 0$ and against $v_t$ when $v_r v_t < 0$. It can be shown that this transverse acceleration is the major cause of the difference between Newtonian and relativistic orbit shapes.

This is however not the full story of the effect of transverse velocity on gravitational acceleration. An object with transverse velocity ‘feels’ more ‘gravitational pull’ than an object momentarily stationary in the gravitational field.

The static value $-g_{tt}\bar{M}/r^2$ is modified by a factor $(1 + 2g_{rr}v_t^2)$, giving a radial acceleration component, due purely to the transverse velocity, of

$$a_{r(v_t)} = \frac{-\bar{M}}{r^2} 2 g_{tt} g_{rr} v_t^2 = \frac{-\bar{M}}{r^2} 2 v_t^2. \tag{C.18}$$

The proof of this relatively simple expression is rather involved and for those readers curious enough about such things, it is shown in section C.2.3 below.

In principle it involves finding the rate of change of radial velocity at the periapsis or apoapsis of a relativistic orbit (when $v_r = 0$ and $v_t \neq 0$) and then subtracting the centrifugal acceleration from this value.

In the case of Newtonian gravity, this gives the radial acceleration of a test object at that point as

$$a'_r = \frac{d^2 r}{dt^2} - \frac{v_t^2}{r} = \frac{-\bar{M}}{r^2}. \tag{C.19}$$

In the relativistic case, the radial acceleration is obtained as

$$a_r = \frac{d^2r}{dt^2} - \frac{v_t^2}{r} = \frac{-\bar{M}}{r^2}(g_{tt} + 2v_t^2). \tag{C.20}$$

As is shown in the referenced section, this equation is derived from the Schwarzschild metric, through a rather tedious manipulation.

We can now sum all the various acceleration terms to get, in the radial and transverse directions respectively,

$$\begin{aligned} a_r &= \frac{-\bar{M}}{r^2}(g_{tt} - 3g_{rr}v_r^2 + 2v_t^2), && \text{(C.21)} \\ a_t &= \frac{2\bar{M}}{r^2}\, g_{rr}v_r v_t. && \text{(C.22)} \end{aligned}$$

As shown in chapter 6, this set of acceleration equations can be used in a quasi-Newtonian way to find the trajectory (or orbit) of any small object in the gravitational field of an isolated static, non-rotating and uncharged black hole. It works just as well for orbits around any spherically symmetrical mass, which can be represented by a point mass.

It is easy to verify that the accelerations do become Newtonian in the low velocity, weak field limit, where $g_{tt} \to 1$ and $v_r,\ v_t \ll 1$. The transverse component becomes vanishingly small as $v_r v_t \to 0$, which applies to most of the orbits in our solar system.

Even in the case of planet Mercury, the maximum value of $v_r v_t$ is of the order $10^{-8}$. With $2\bar{M}/r^2$ of the order $10^{-18}$ m$^{-1}$, the opposing transverse acceleration is of the order $10^{-26}$ m$^{-1}$ maximum, and when multiplied by $c^2$, it translates to only $10^{-9}$ m/s$^2$. This is a real tiny opposing transverse acceleration (some 0.1 nano-g).

It can however be shown that it is the main contributor to the perihelion shift. In the simulation of the 'poor man's orbit' in chapter 6, the opposing transverse acceleration alone produces a periapsis shift of about $1.5\pi$ radians, out of a total periapsis shift of $2\pi$ radians (obviously only correct for the initial conditions used in the example). The other $\pi/2$ radians comes from the opposing radial acceleration.

Since the author could provide little or no references to support this work, the reader must wonder whether the equations are correct or whether the author is waffling.

There are two ways of increasing the confidence in the equations—the first is to study the derivations in the next section; the second is to note that the above accelerations can be used in a quasi-Newtonian way to yield orbits that are virtually indistinguishable from those obtained using the formal relativistic orbital equations, as shown in chapter 6.

Does the latter prove the equations right? Not quite. But they seem to pass the 'engineering test': *close enough for (most) practical purposes.*

The author has to admit that this is a quite unorthodox treatment of relativistic acceleration and may perhaps be frowned upon. It may seem to be even more complex than the formal relativistic method used in chapter 6. However, from an engineer's perspective, there is something 'tangible' about this treatment that is somewhat lacking in the formal treatment, which is not a very intuitive process, to say the least.

The quasi-Newtonian approach given here is not offered as an alternative to the formal theory, but simply as a 'crutch' towards better comprehension of relativity. For those readers interested, a more formal treatment of various relativistic accelerations is given in the remaining part of this appendix.

## C.2 Issues of a more technical nature

This discussion assumes, as a starting point, the validity of the relativistic addition of velocities rule, the Schwarzschild metric and some of the solutions to the geodesic equations. Therefore there is no tensor algebra involved.

### C.2.1 Acceleration caused by a constant force

Let a test particle of rest mass $m_0$ move at velocity $v$ and be accelerated by a constant force $F_0$, working in the direction of movement, all as measured in some inertial frame of reference. If we use Newton dynamics, then in a time interval $\Delta t$, the particle will undergo a speed increment of $\Delta v' = F_0/m_0 \Delta t$, with a new velocity of $v_1' = v + \Delta v'$.

In relativistic dynamics, we cannot simply add the velocity increment to the original velocity—we have to use the relativistic addition of velocities rule, giving a velocity of

$$v_1 = \frac{v + \Delta v'}{1 + v\Delta v'}. \tag{C.23}$$

Hence, the velocity increment is

$$\Delta v = v_1 - v = \frac{v + \Delta v'}{1 + v\Delta v'} - v = \frac{\Delta v'(1 - v^2)}{1 + v\Delta v'}. \tag{C.24}$$

The acceleration $a = \Delta v/\Delta t$ as observed in the inertial frame is, after substituting $\Delta v' = F_0/m_0 \Delta t$ and then dividing by $\Delta t$,

$$a = \frac{F_0}{m_0} \frac{(1 - v^2)}{(1 + v\Delta v')}. \tag{C.25}$$

Now let $\Delta t \to 0$, so that $\Delta v' \to 0$ and we have

$$a = \frac{F_0}{m_0}(1 - v^2). \tag{C.26}$$

It is as if the effect of force $F_0$ decreases by a factor $\sqrt{1 - v^2}$ and the rest mass $m_0$ increases by a factor $1/\sqrt{1 - v^2}$. Note that the sign of $v$ does not play a role, as long as the velocity is in line with the force vector. So the same effect will be observed if the velocity vector is in the opposite direction to the force vector.

One can say that the expected (Newtonian) acceleration is decreased by *two factors of velocity time dilation*. Alternatively, one can say that there appears to be an opposing acceleration of $\frac{-F_0}{m_0}v^2$ working against the expected Newtonian acceleration of $\frac{F_0}{m_0}$.

This is only valid if the velocity vector is parallel to the force vector. If the velocity vector is normal to the force vector, the Newtonian acceleration will only be decreased by *one* factor of velocity time dilation ($\sqrt{1-v^2}$), because then only the increase in mass plays a role.

### C.2.2 Gravitational acceleration with pure radial velocity

As a solution to the geodesic equations, the conservation of total energy of a radially free-falling object can be expressed as (chapter 6)

$$\tilde{E}^2 = \frac{g_{tt}}{1-v_\varphi^2}, \tag{C.27}$$

where $\tilde{E}$ is the total energy parameter (a constant), $v_\varphi$ the local radial velocity and $g_{tt} = 1 - 2\bar{M}/r$. From this it is easy to extract

$$v_\varphi^2 = 1 - \frac{1-2\bar{M}/r}{\tilde{E}^2}. \tag{C.28}$$

Since the square of the distant radial velocity is $v_r^2 = (1-2\bar{M}/r)^2 \ v_\varphi^2$,

$$v_r^2 = \left(\frac{dr}{dt}\right)^2 = (1-2\bar{M}/r)^2 - \frac{(1-2\bar{M}/r)^3}{\tilde{E}^2}. \tag{C.29}$$

Now differentiate $r$ with respect to $t$ and simplify to get

$$\begin{aligned}
\frac{d^2r}{dt^2} &= \frac{2\bar{M}(1-2\bar{M}/r)}{r^2} - \frac{3\bar{M}(1-2\bar{M}/r)^2}{r^2\tilde{E}^2} \\
&= \frac{2\bar{M}}{r^2}g_{tt} - \frac{3\bar{M}}{r^2}g_{tt}^2\left(\frac{1-v_\varphi^2}{g_{tt}}\right) \\
&= \frac{\bar{M}}{r^2}g_{tt}[2-3(1-v_\varphi^2)] \\
&= \frac{-\bar{M}}{r^2}g_{tt}(1-3v_\varphi^2) \\
&= \frac{-\bar{M}}{r^2}\left(g_{tt}-3g_{rr}v_r^2\right),
\end{aligned} \tag{C.30}$$

the gravitational acceleration of a radially free falling object as measured by the distant observer (recall that $v_\varphi = g_{rr}v_r$ and $g_{rr} = 1/g_{tt}$.) This is the same value of radial acceleration as obtained before through the 'opposing' accelerations due to velocity and space curvature.

### C.2.3 Gravitational acceleration at the peri- or apoapsis

The Schwarzschild metric can be reworked into the form (similar to [Faber] eq. 162)

$$g_{tt}\left(\frac{\tilde{E}}{g_{tt}}\right)^2 - \frac{1}{g_{tt}}\left(\frac{dr}{d\phi}\frac{\tilde{L}}{r^2}\right)^2 - r^2\left(\frac{\tilde{L}}{r^2}\right)^2 = 1, \tag{C.31}$$

where $g_{tt}$ takes the place of Faber's $\gamma$, $\tilde{E}$ the place of $b$ and $\tilde{L}$ the place of $h$. Expand $g_{tt} = 1 - 2\bar{M}/r$ and rearrange, to get

$$\frac{dr^2}{d\phi^2} = (\tilde{E}^2 - 1)r^4/\tilde{L}^2 + 2\bar{M}r^3/\tilde{L}^2 - r^2 + 2\bar{M}r. \tag{C.32}$$

Since $d\phi^2 = v_t^2 dt^2/r^2$, we can state

$$\frac{dr^2}{dt^2} = \left((\tilde{E}^2 - 1)r^2/\tilde{L}^2 + 2\bar{M}r/\tilde{L}^2 - 1 + 2\bar{M}/r\right) v_t^2. \tag{C.33}$$

If we work at the periapsis or apoapsis of an orbit, $dv_t/dt = 0$,* so that

*The transverse acceleration $dv_t/dt$ smoothly changes sign there, so it must go through zero.

after differentiating $r$ with respect to $t$ and rearranging, we get

$$\frac{d^2r}{dt^2} = \left((\tilde{E}^2 - 1)r/\tilde{L}^2 + \bar{M}/\tilde{L}^2 - \bar{M}/r^2\right) v_t^2. \tag{C.34}$$

Since the radial acceleration $a_r = d^2r/dt^2 - v_t^2/r$, we can write

$$a_r = \left((\tilde{E}^2 - 1)r/\tilde{L}^2 + \bar{M}/\tilde{L}^2 - \bar{M}/r^2 - 1/r\right) v_t^2. \tag{C.35}$$

Substitute $\tilde{E}$ and $\tilde{L}$ (for $v_r = 0$) as defined in chapter 6 and simplify

$$\begin{aligned}
a_r &= \frac{g_{tt}^2 - g_{tt} + v_t^2}{r} + \frac{\bar{M}}{r^2}(g_{tt} - v_t^2) - \frac{\bar{M}}{r^2}v_t^2 - \frac{v_t^2}{r} \\
&= \frac{g_{tt}^2 - g_{tt}}{r} + \frac{\bar{M}}{r^2}(g_{tt} - v_t^2) - \frac{\bar{M}}{r^2}v_t^2 \\
&= \frac{-2\bar{M}}{r^2}g_{tt} + \frac{\bar{M}}{r^2}(g_{tt} - v_t^2) - \frac{\bar{M}}{r^2}v_t^2 \quad \text{(see * below)} \\
&= \frac{-\bar{M}}{r^2}(g_{tt} + 2v_t^2). 
\end{aligned} \tag{C.36}$$

* It is easy to show that: $(g_{tt}^2 - g_{tt})/r = (1 - 2\bar{M}/r - 1)g_{tt}/r = -2\bar{M}g_{tt}/r^2$.

### C.2.4 Angular acceleration in an orbit

The angular acceleration in a Newtonian orbit is obtained from the Newtonian equations of motion, e.g., [Faber] eq. 109, here reworked as

$$\dot{\omega}_N = \frac{d^2\phi}{dt^2} = \frac{-2}{r}\frac{dr}{dt}\frac{d\phi}{dt} = \frac{-2v_r v_t}{r^2}, \tag{C.37}$$

since $dr/dt = v_r$ and $d\phi/dt = v_t/r$, the latter taken as always positive.

Similarly, the relativistic angular acceleration is obtained via the geodesic equations as

$$\dot{\omega}' = \frac{d^2\phi}{d\tau^2} = \frac{-2}{r}\frac{dr}{d\tau}\frac{d\phi}{d\tau} = \frac{-2}{r}\frac{dr}{dt}\frac{d\phi}{dt}\frac{dt^2}{d\tau^2} = \frac{-2v_r v_t}{r^2}\frac{dt^2}{d\tau^2}, \tag{C.38}$$

a rework of [Faber] eq. 159c. This 'angular acceleration' in terms of coordinate angles and proper time differential ($d\tau$) is now to be transformed to the coordinate clock differential ($dt$).

The process is quite complex (see the next section), but the result is rather simple:

$$\dot{\omega} = \frac{d^2\phi}{dt^2} = \frac{-2v_r v_t}{r^2}\left(\frac{r-3\bar{M}}{r-2\bar{M}}\right). \tag{C.39}$$

This gives another clue as to why an orbiting object passing within $r = 3\bar{M}$ of a black hole will be dragged into the hole. If $v_r < 0$ and $2\bar{M} < r < 3\bar{M}$, the angular acceleration becomes negative, which is in conflict with a stable orbit on its inbound leg.

The previous result can be rewritten as

$$\dot{\omega} = \frac{-2v_r v_t}{r^2} + \frac{2\bar{M}}{r^3} g_{rr} v_r v_t, \tag{C.40}$$

where $g_{rr} = 1/(1-2\bar{M}/r)$. This shows that there is an apparent opposing angular acceleration of

$$\dot{\omega}_{opp} = \frac{2\bar{M}}{r^3} g_{rr} v_r v_t \tag{C.41}$$

working against the Newtonian angular acceleration $\dot{\omega}_N$. The opposing linear transverse acceleration is then

$$\dot{v}_{t(opp)} = r\,\dot{\omega}_{opp} = \frac{2\bar{M}}{r^2} g_{rr} v_r v_t. \tag{C.42}$$

As shown before, half of this opposing transverse acceleration comes from velocity effects and half from space curvature.

### C.2.5 Angular acceleration derived for the distant observer

What follows is rather tedious and adds no real relativistic insight, so it may be skipped without loosing anything. The reason for the complexity is that we need to work from $d\phi/dt$, which is not easily expressible in terms of $\phi$ or $t$, but rather as a function of radial distance $r$.

One needs to do quite a bit of juggling with differentials of 'functions of functions'.* The following geodesic relations (from chapter 6) are used in

*If a mathematically minded reader knows a more comprehensible procedure, the author would be delighted to hear about it.

this derivation:

$$\frac{d\phi}{d\tau} = \frac{\tilde{L}}{r^2} \text{ and } \frac{dt}{d\tau} = \frac{\tilde{E}}{1-2\bar{M}/r}, \text{ where}$$

$$\tilde{L} = \frac{rv_t}{\sqrt{g_{tt} - g_{rr}v_r^2 - v_t^2}} \text{ and } \tilde{E} = \frac{g_{tt}}{\sqrt{g_{tt} - g_{rr}v_r^2 - v_t^2}}.$$

From the above, it is easily derived that

$$\frac{d\phi}{dt} = \frac{d\phi}{d\tau}\frac{d\tau}{dt} = \frac{\tilde{L}}{r^2}\frac{1-2\bar{M}/r}{\tilde{E}} = \frac{1-2\bar{M}/r}{r^2}\frac{\tilde{L}}{\tilde{E}}. \qquad (C.43)$$

Since both $\tilde{L}$ and $\tilde{E}$ are constants in an orbit, it follows that $\tilde{L}/\tilde{E}$ is also a constant, which we will shall call $K$ for brevity. Hence we can write

$$\left(\frac{d\phi}{dt}\right)^2 = K^2\left(\frac{1-2\bar{M}/r}{r^2}\right)^2, \qquad (C.44)$$

which is to be differentiated against $t$. To make differentiation simpler, define

$$\mu = \frac{1}{r}, \text{ so that } \frac{d\mu}{dr} = \frac{-1}{r^2} = -\mu^2, \text{ and also define}$$

$$\nu = \left(\frac{1-2\bar{M}/r}{r^2}\right)^2 = (1-2\bar{M}\mu)^2\mu^4, \text{ so that}$$

$$\frac{d\nu}{d\mu} = 4\mu^3(1-2\bar{M}\mu)(1-3\bar{M}\mu),$$

by the 'product rule' and factorization. Other useful relations in terms of $\mu$ are

$$\frac{d\mu}{dt} = \frac{d\mu}{dr}\frac{dr}{dt} = -\mu^2 v_r \text{ and } \frac{d\phi}{dt} = \frac{v_t}{r} = \mu v_t.$$

Equation C.44 above can now be simply written as

$$\left(\frac{d\phi}{dt}\right)^2 = K^2\nu, \qquad (C.45)$$

which, with the above definitions, is easy to differentiate against $t$, i.e.,

$$2\frac{d\phi}{dt}\frac{d^2\phi}{dt^2} = K^2\frac{d\nu}{dt} = K^2\frac{d\nu}{d\mu}\frac{d\mu}{dt}. \qquad (C.46)$$

After substitution and division by $2\mu v_t$, we have

$$\begin{aligned}\frac{d^2\phi}{dt^2} &= 4K^2\mu^3(1-2\bar{M}\mu)(1-3\bar{M}\mu)\frac{-\mu^2 v_r}{2\mu v_t} \\ &= -2K^2\mu^4\frac{v_r}{v_t}(1-2\bar{M}\mu)(1-3\bar{M}\mu). \qquad (C.47)\end{aligned}$$

Finally, substitute $\mu = 1/r$ and the constant $K^2 = \tilde{L}/\tilde{E} = r^2v_t^2/(1-2\bar{M}/r)^2$ to arrive at

$$\frac{d^2\phi}{dt^2} = \frac{-2v_rv_t}{r^2}\left(\frac{1-3\bar{M}/r}{1-2\bar{M}/r}\right) = \frac{-2v_rv_t}{r^2}\left(\frac{r-3\bar{M}}{r-2\bar{M}}\right), \qquad (C.48)$$

as used in the previous section.

## C.3 Summary of Quasi-Newtonian acceleration

This then wraps up engineering relativity. The subject is complex, but it has been shown that for moderately simple cases, it can be reduced to almost Newtonian form. What is more, the 'almost Newtonian form' has an engineering 'look and feel'!

The most important part to understand is that objects at high velocity resist being accelerated more than what they do according to Newton. This is the opposing acceleration components for flat space, where gravity is absent.

Bring in the curved space of gravity and it appears as if there is even more resistance to be accelerated. This is caused by the coordinate transformations between different regions in curved space.

May the complexities of objects moving fast in a gravitational field be more comprehensible to the engineer that has read this!

# Appendix D

# The Titius-Bode law

the very interesting
spacing of the
planets

In 1766, Johann Daniel Titius of Wittenberg in Germany formulated the modern version of the empirical rule now known as the *Titius-Bode law.* It was apparently done as a note that he added while translating a work from the French natural philosopher Charles Bonnet titled "Contemplation de la Nature" into German. The note, translated into English reads:

> "Take notice of the distances of the planets from one another, and recognize that almost all are separated from one another in a proportion which matches their bodily magnitudes. Divide the distance from the Sun to Saturn into 100 parts; then Mercury is separated by four such parts from the Sun, Venus by 4+3=7 such parts, the Earth by 4+6=10, Mars by 4+12=16.
>
> "But notice that from Mars to Jupiter there comes a deviation from this so exact progression. From Mars there follows a space of 4+24=28 such parts, but so far no planet was sighted there. But should the Lord Architect have left that space empty?
>
> "Not at all. Let us therefore assume that this space without doubt belongs to the still undiscovered satellites of Mars, let us also add that perhaps Jupiter still has around itself some smaller ones which have not been sighted yet by any telescope.
>
> "Next to this for us still unexplored space there rises Jupiter's sphere of influence at 4+48=52 parts; and that of Saturn at 4+96=100 parts. What a wonderful relation!"

In 1768, astronomer Johann Elert Bode published more or less the wording of Titius in the second edition of his introduction to astronomy: "Anleitung zur Kenntniss des gestirnten Himmels". He originally did not give credit to Titius, so the 'law' came to be known as Bode's law in the astronomy world. He did however, in later editions of his book, mention Titius as his source, hence the modern name of the law: the Titius-Bode law.

Bode did correctly note that the idea of Titius that the 'missing' planet may be a moon of Mars is incorrect and stated that a planet may yet be discovered at that distance. This eventually lead to the discovery of the largest asteroid, Ceres, at the predicted distance.*

> *All the above information extracted from "BODE'S LAW AND THE DISCOVERY OF CERES" by Michael Hoskin, Churchill College, Cambridge.

Bode related the mean distances of the then known planets from the Sun to a simple mathematic progression of numbers. Start with the following sequence of numbers: 0, 3, 6, 12, 24, 48, 96, 192, 384, 768. Add 4 to each number, giving 4, 7, 10, 16, 28, 52, 100. Then divide each number by 10, giving 0.4, 0.7, 1.0, 1.6, 2.8, 5.2, 10.0. This gives the Bode distances from the Sun to successive planets in astronomical units (a.u.).*

> *One astronomical unit (a.u.) = 149,597,870 km, the mean distance between Earth and the Sun.

The Titius-Bode law for distance $D$ can also be formulated as follows:

$$D = 0.4 + (0.3 \times N) \text{ a.u.},$$

where $N =$ integer$(2^{n-2})$ and $n$ means the $n^{th}$ planet from the Sun, counting in the asteroid Ceres. The series $N$ works out to be 0, 1, 2, 4, 8, 16, 32, 64, 128, 256,.... The empirical law holds up well for planets up to Uranus, but breaks down for Neptune and Pluto, as is evident from table D.1.

| **Some Planetary Data in the Solar System** | | | | | |
|---|---|---|---|---|---|
| n | **Planet** | $D_m$ **(a.u.)** | $D$ **(a.u.)** | $P$ **(years)** | $V_o$ **(km/s)** |
| 1 | Mercury | 0.39 | 0.40 | 0.24 | 47.89 |
| 2 | Venus | 0.72 | 0.70 | 0.62 | 35.03 |
| 3 | Earth | 1.00 | 1.00 | 1.00 | 29.79 |
| 4 | Mars | 1.57 | 1.60 | 1.88 | 24.13 |
| 5 | Ceres | 2.77 | 2.80 | 4.61 | 17.90 |
| 6 | Jupiter | 5.20 | 5.20 | 11.86 | 13.06 |
| 7 | Saturn | 9.54 | 10.00 | 29.46 | 9.64 |
| 8 | Uranus | 19.19 | 19.60 | 84.01 | 6.81 |
| 9 | Neptune | 30.06 | 38.80 | 164.79 | 5.43 |
| 10 | Pluto | 39.53 | 77.20 | 248.54 | 4.74 |

**Table D.1:** Comparison between the correct mean distances of planets ($D_m$) and the predictions of Bode's law ($D$). $P = D_m^{1.5}$ is the period and $V_o = \frac{29.79}{\sqrt{D_m}}$ is the mean orbit velocity in a reference frame where the Sun is at rest.

Although Bode's law is merely empirical, there is today some serious considerations that it might have an underlying physical principle. No physical theory for the formation of the solar system could thus far derive such a principle in a rigorous way, but observations of planetary systems around other stars seems to suggest that they also follow something similar to Bode's law.

The following quote by John Gribbin from THE GUARDIAN, LONDON, illustrates the point: "The discovery of three planets orbiting a pulsar known as PSR B1257+12 has revealed a system with properties that almost exactly match those of the Inner Solar System, made up of Mercury, Venus and Earth. The similarities are so striking that it seems there may be a law of nature which ensures that planets always form in certain orbits and always have certain sizes; and it leads credence to the significance of a mathematical relationship that relates the orbits of the planets in our Solar System, which many astronomers have dismissed as mere numerology."

It is today accepted that Pluto is part of the Kuiper Belt, a huge disk-like plane of planetoids and comets that circle the Sun outside of Neptune's orbit. In 1951, astronomer Gerard Kuiper postulated the existence of such a 'belt', containing left-over debris from the formation of the solar system.

In 1992, a 150-mile wide body, called 1992QB1, was detected at the predicted distance of the Kuiper Belt. It was quickly followed by the detection of several similar-sized objects, confirming that the Kuiper Belt was real.

The Hubble space telescope detected many more smaller planetoids and also confirmed that many comets are circling the Sun in the Kuiper belt. This explained the existence of short-period comets (with periods of less than 200 years).*

*The Kuiper Belt is not to be confused with the Oort Cloud, which is postulated to be a spherical shell of comets at some 5000 a.u. It could be the source of the long-period comets.

Another interesting relationship between the periods of the planets is shown in figure D.1. The logarithms of the planetary periods against their sequential positions produce a nearly linear slope. The rotation period of the Sun near it's equator closely fits the position of 'planet zero'!

Although it may not look like it, for the planets up to Uranus, this method is far less accurate than Bode's law—the 'close fit' is an illusion caused by the logarithmic vertical scale. The actual deviation (in days) from the straight line 'law' is generally quite large, peaking at around $\pm 30\%$.

The relationship in figure D.1 also holds for the major moons of the larger planets, at least for those that have a reasonable number of major moons, like Jupiter and Uranus. Saturn has only one major moon (Titan) and likewise, Neptune has only one of any significance (Triton). Figures D.2 and D.3 show the plots for the periods of the major moons of Jupiter and Uranus respectively.

What does all this mean? Simulations of planetary formation are extraordinarily complex. It is thought that the solar system originated from a 'proto-solar nebula', a 'cocoon' of gas and dust that contracted under gravitational forces (perhaps aided by a nearby supernova explosion) to become denser at the center, forming the Sun. Such proto-solar nebulae are observed around other stars in our Galaxy.

The contraction caused the nebula to rotate and flatten out to form what is known as a 'proto-planetary disk'. Out of this disk the planets formed through accretion of gas and dust, passing through various phases.

It is thought that the process started with "dirty gas balls" that collided frequently to form 'planetesimals' (minute planets). Through a complex process of collisions, the planetesimals finally formed the planets as we know them.

This is an extreme oversimplification and as they say, 'the devil is in the detail'. Present day models produce neither the planetary spacings, nor the planetary sizes and compositions that we observe, so the 'right' model still needs to be found.

The computing power required to run more advanced models is exorbitant, so a representative model is probably still quite a distance into the future. The Internet contains many short articles about old and new theories of planetary formation, e.g., [space.com]. Just do a search on that website.

The log-linear plots are on the next page...

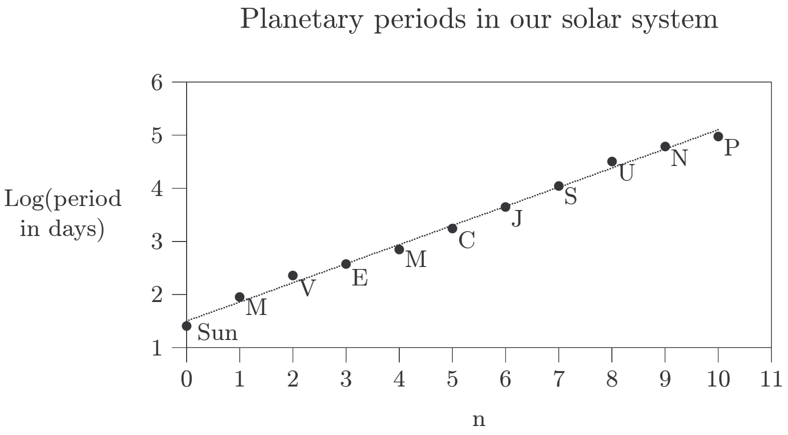


**Figure D.1:** A $\log_{10}$ plot of the planetary periods against the sequential positions of the planets. The Sun's surface is taken as 'planet 0', with the 'orbital period' equal to the rotation period at the Sun's equator. The straight line is the equation log(period)=1.5 + 0.35n.

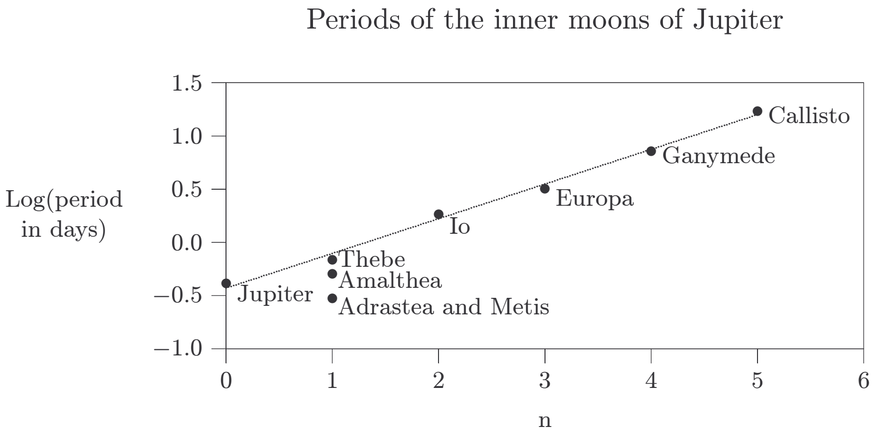


**Figure D.2:** A plot of the periods of the inner moons of Jupiter. Between Jupiter and Io, four asteroid-sized moons are found, but with irregular spacing. They do not fit the linear logarithmic 'law', while the four large 'Galilean' moons fit quite closely. Outside of Callisto, there are another eight small asteroid-like moons. They may possibly be passing objects that were captured by Jupiter.

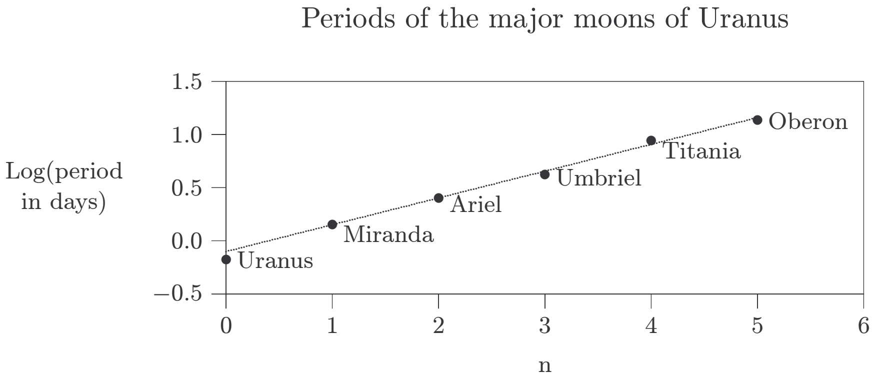


**Figure D.3:** The (log) periods of the major moons of Uranus. There are at least 10 very small moons between Uranus and Miranda (detected in 1986 by the spacecraft Voyager 2), but they seemingly do not require a separate interval.

# Bibliography

[Coord.] Hamilton A. 1998. *Schwarzschild Coordinates*. Internet: http://casa.colorado.edu/

[Einstein] Einstein A. 1950. *Essays in Physics*. New York: Philosophical Library, Inc. Republished in 1996 as *The Theory of Relativity and other essays*. New York: CITIDAL Press.

[Faber] Faber, R.L. 1983. *Differential Geometry and Relativity Theory*. New York: Marcel Dekker, Inc.

[Ferguson] Ferguson K. 1999. *Measuring the Universe*. London: Headline Book Publishing.

[GP-B] NASA 2004. *A Pocket of Near-Perfection: Gravity Probe B*. Internet: http://science.nasa.gov/headlines/y2004/26apr-gpbtech.htm

[Gribbin] Gribbin J. and Rees M. 1991. *Cosmic Coincidences*. London: Black Swan.

[Guth] Guth A. and Steinhardt P. 1990. *The Inflationary Universe*. Scientific American Jan. 1990.

[Ibison] Ibison M., Puthoff H. and Little S. *The speed of gravity revisited*. Austin: Institute for Advanced Studies (also available on Internet).

[Mitton] Mitton, J. 1991. *A Concise Dictionary of Astronomy*. Oxford University Press.

[MTW] Misner C.W., Thorne K.S. and Wheeler J.A. 1973. *Gravitation*. San Francisco: W.H. Freeman and Company.

[Pathria] Pathria R.K. 2003. *The Theory of Relativity, 2nd ed.*. New York: Dover Publications Inc.

[Peacock] Peacock J.A. 1998. *Cosmological Physics*. Edinburgh University Press.

[Peebles] Peebles P.J.E. 1993. *Principles of Physical Cosmology*. Princeton University Press.

[Sigg] Sigg D. 2001. *Gravitational Waves.* Internet: www.ligo.caltech.edu/

[Smoot] Smoot G. and Davidson K. 1993. *Wrinkles in Time.* London: Abacus.

[space.com] Internet: http://www.space.com/scienceastronomy/solarsystem/ planetformation020709-1.html.

[Thorne] Thorne K.S. 1995. *Black Holes and Time Warps.* London: Papermac.

[Tides] NOAA/NOS. *Our Restless Tides.* Internet: www.co-ops.nos.noaa.gov/

[Will(a)] Will C.M. 1981,1993. *Theory and Experiment in Gravitational Physics.* Oxford University Press.

[Will(b)] Will C.M. 1988. *Was Einstein Right?.* Oxford University Press.

[Will(c)] Will C.M. Internet: http://www.livingreviews.org/Articles/ Volume4/2001-4will/

# Index