General Relativistic Dynamics

Extending Einstein's Legacy Throughout the Universe

Fred I. Cooperstock

World Scientific

General Relativistic Dynamics

Extending Einstein's Legacy Throughout the Universe

This page intentionally left blank

General Relativistic Dynamics

Extending Einstein's Legacy Throughout the Universe

Fred I. Cooperstock
University of Victoria, Canada

World Scientific
NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

Published by

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

Library of Congress Cataloging-in-Publication Data

Cooperstock, F. (Fred)

General relativistic dynamics : extending Einstein's legacy throughout the universe / Fred I Cooperstock.

p. cm.

Includes bibliographical references and index.

ISBN-13: 978-981-4271-16-5 (hardcover : alk. paper)

ISBN-10: 981-4271-16-0 (hardcover : alk. paper)

1. General relativity (Physics) 2. Gravity. 3. Gravitational fields. 4. Galaxies. I. Title.

QC173.6.C67 2009

530.11--dc22

2009012008

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

Einstein illustration by *Travis Morgan*, morgantj@gmail.com

Galaxy illustration by *Robert Gendler*, www.robgendlerastropics.com

Printed in Singapore.

In memory of Sima, Tevya and Aryeh

This page intentionally left blank

Contents

... and the slavery of fear had made men afraid to think. But such is the irresistible nature of truth, that all it asks, and all it wants, is the liberty of appearing.

Thomas Paine in "The Rights of Man" [1]

The only justification for our concepts and systems of concepts is that they serve to represent the complex of our experiences; beyond this, they have no legitimacy.

Albert Einstein in "The Meaning of Relativity" [2]

Chapter 1

Introduction

One often reads about the beauty and importance of general relativity (GR). It is viewed almost universally as our premier theory of gravity, superseding Newton's theory of gravity in a profound way. We share this view whole-heartedly. However, while there have been a very large number of papers written in the field, it is disconcerting to note how little there is in the way of a connection of this large accumulated effort with the great body of the world's physicists and astronomers. Many of these papers are of great mathematical complexity, decipherable only by the experts in the field. Even when decoded, many, if not most, have little relevance to the important issues of physics and astronomy. This is unfortunate as gravity is assuming a greater role of significance in science and the best theory of gravity should be readily accessible to the non-experts.

Our goal in this book is multi-fold: first, we wish to present the essentials of general relativity in a simple way so that any physicist who might have missed training in the field, once having digested the primary concepts and equations, will not have his or her eyes glaze over when confronted by some reference to or equation for general relativity. To do so, we will also cover the essentials of special relativity to provide a smooth transition to the general theory. For a detailed development of special relativity, the reader is directed to the truly classic work of L. D. Landau and E. M. Lifshitz [3] and the very intuitive approach of H. Bondi [4]. The special relativistic treatment in this book has been particularly influenced by these excellent thinkers.

The reader who is comfortable with the standard development of special relativity but who is unfamiliar with Bondi's intuitively appealing approach might wish to start with Chapter 3. In the standard approach, we discuss how the physical equivalence of all inertial reference frames plus the experimental result that the velocity of light is an invariant form the foundation of special relativity, Einstein's theory of space and time in the absence of gravity. These two cornerstones show us that the old transformations of the spacetime coordinates with the Cartesian coordinates with which we are familiar, no longer hold and that time is no longer an absolute. The correct transformations using Cartesian coordinates preserve both the value and the form of the spacetime interval, the important measure linking time and space that will lead us into general relativity. Most importantly, we will show that the lengths of bodies and the intervals of time are seen to vary in remarkable ways when viewed in different frames of reference. We will get our first taste of the importance of extrema in physics, how the minimum possible time interval between the ticks of a clock is read in the rest frame of that clock (the "proper time interval") whereas the maximum possible length of a body is the length that is measured in the rest frame of that body (the "proper length"). Being extrema of opposite sense underlines the reciprocal nature of time and space.

In the development of special relativity, we have our first contact with the vectors and tensors of four-dimensional spacetime. These mathematical structures are used with the important Principle of Least Action, which forms the basis of the fundamental equations of physics and relativistic dynamics.

In Chapter 3, Bondi's approach to special relativity builds upon the Doppler factor between two observers in relative motion. Some of the basic results of special relativity are re-derived using Bondi's simple and appealing framework. It enables us to resolve the so-called twin or clock paradox, the asymmetric aging of twins who reunite after a voyage of separation. We will see that there is nothing mysterious that occurs at the turnaround point in the voyage of the accelerating twin. Different paths in spacetime track different spans of time in analogy with the different paths in space which track different spans of distance.

Having covered the essentials of special relativity, we proceed into

the next major step in Chapter 4, the development of general relativity, Einstein's theory of gravity. (The reader who is comfortable with the basics of general relativity might wish to proceed to Chapter 5 or 6, depending upon his/her familiarity with the subject.) We first discuss the Principle of Equivalence, the local equivalence between accelerated reference frames and gravitational fields that was the guiding light for Einstein in his quest for a relativistic theory of gravity. We also focus on the importance of the approximate aspect of the Equivalence Principle, how it is spacetime curvature rather than accelerated reference frames that constitutes true gravity, a point well-articulated by J. L. Synge. That being said, the usefulness of the Equivalence Principle remains, as illustrated in the lead-in to the spacetime metric tensor as describing a gravitational field.

We then proceed with the basic mathematics, tensor calculus, that is required for technical work in general relativity. (Depending upon the degree to which the reader may wish to follow technical aspects, he or she may wish to skim over the sections that follow in Chapter 4 and then concentrate more carefully beginning with Chapter 5.) With the basic aspects of general coordinate transformations covered, we proceed to explore the nature of curved spaces, introducing the important Riemann tensor, which characterizes gravity in an invariant manner. We then focus upon a key GR departure from Newtonian gravity, the removal of gravity from the category of forces. A freely gravitating body is seen to move with zero intrinsic acceleration, following the extremal paths, the geodesics of the spacetime that the body occupies. Motion under gravity as a force is replaced with free motion following the special paths in curved spacetime in analogy with airline pilots who follow the geodesics, the great circles on the globe, to minimize distance between two points.

The energy and momentum conservation laws of special relativity are first generalized to arbitrary coordinate systems using the new generalized derivative, the "covariant derivative". Guided by the Principle of Equivalence, we are led to the conservation laws for general relativity. Consistency with these laws and the demand for melding with Newtonian gravity under appropriate conditions brings us to the Einstein tensor. This incorporates gravity on the left hand side of the Einstein field equations to equal the source, the energy-

momentum tensor, on the right hand side. Having established the Einstein field equations with the basic background structures and concepts, we are prepared to study the relativistic world of gravity.

We first study the simple important and very interesting Schwarzschild solution of the Einstein equations in Chapter 5, the gravitational field in vacuum under the condition of spherical symmetry. To do so, we require the concepts of proper distance and proper time in general metrics, including gravity. The study of the Schwarzschild spacetime introduces interesting issues, event horizons, black holes and singularities, issues that have ignited the imagination of the general public for decades. As compared to special relativity, there is a paucity of experimental corroboration for the theory of general relativity. We outline the various issues regarding these tests.

As there are waves of electromagnetic nature emanating from the acceleration of charges, Einstein's general relativity predicts that there must be waves of gravitational nature arising from the acceleration of masses. In Chapter 6, we discuss these waves briefly. Gravity waves have never yet been observed directly but their existence is inferred from the motion of certain sources such as the binary pulsar, PSR1913+16. However, issues concerning energy and momentum for such waves are not clear-cut. Energy localization has been an enduring controversy in general relativity. We will bring forward our hypothesis that energy, including the contribution from gravity, is most logically localized in the regions of the energy-momentum tensor. This has the unsettling implication that gravitational waves, assuming the reality of their existence to which we certainly subscribe, are not carriers of energy in vacuum.

Part of the natural appeal of physics is that it encompasses all scales of dimension, from the very tiniest size, a size so small as to challenge the imagination, to the universe itself which may in fact be infinite. In Chapter 7, we wind our way through the hierarchy of scales in nature, connecting them to the fundamental forces. We know that the macroscopic physics of our everyday experience breaks down completely when we reach the quantum scale of atomic physics at dimension 10^{-10} m. At this level and smaller, the world as we know it is gone. New forces come into play, the "strong force" binding the nucleus and the "weak force" responsible for the decay of particles such as the neutron. Even the familiar electromagnetic force from

macroscopic physics must be "quantized", taking on a new character.

It has been almost universally assumed that gravity must also be quantized at a certain stage. We have argued that this is not necessarily the case since gravity is fundamentally different: all particles and fields other than gravity exist *within* spacetime whereas gravity *is* spacetime, i.e. its curvature. In our view, this aspect sets gravity apart as the enveloper of all the rest of physics and removes the necessity for its quantization. However, it is well to ask whether gravity might play a non-classical role at a scale that arises from equating the Compton wavelength of a particle where a quantum duality sets in, to the dimension of contraction at which a body exhibits an event horizon, bringing into play the full nonlinearity of general relativity. This occurs at the Planck scale, of dimension 10^{-35} m. While we can look at this number with its long chain of zeros, we cannot begin to visualize it as an actual length in any normal sense.

In his earliest work in atomic physics, Bohr had set into motion the quantization of the hydrogen atom with *ad hoc* rules that were remarkably successful for their time. We were inspired by the work of Bohr to attempt an *ad hoc* quantization at the Planck scale, adding spin and charge to the Planck mass.

Interestingly, at the extreme of Planck quantum states, a new level of the dimensionless fine structure constant α arises, namely $1/128$ as opposed to the approximate value of $1/137$ of atomic physics. The $1/128$ value has a serendipitous aspect as this is almost precisely the α value governing high energy radiation in Z-boson production and decay. We know of no particular reason for these numbers to match. Perhaps it is a sheer coincidence. However, R. P. Feynman used to impress upon us that it is the confluence of numbers that can foreshadow important truths in physics.

The preceding focused upon the smallest of scales. Proceeding to the very largest series of scales beginning in Chapter 8, we first take a brief excursion into cosmology, the very largest scale in nature. We provide some perspective by building the image of the vast dimensions that we will encounter by the use of scaling, reducing the size of the Sun to that of the head of a pin. From there, we can better picture the vastness of empty space between the stars and then the immensity of a galaxy. We will discuss some of the early ideas about the cosmos and how the modern picture developed.

There has been much recent interest in the idea that the universe is currently in a state of *accelerated* expansion. We will discuss this aspect briefly in conjunction with the cosmological term in the Einstein equations. We will argue that this term should be viewed most logically as another form of matter, albeit exotic, and not a "geometrical" adjunct to the theory.

As a primary goal, in Chapters 9 and 10, we will show why general relativity must be brought into the greater sweep of dynamical problems in the universe. These entail the motions of stars in the galaxies and the motions of galaxies within clusters. Until now, it was believed that Newtonian gravity was the appropriate theory for these scales and that general relativity came into significant play only in situations of ultra-strong (or at least very strong) gravity or for the dynamics of the universe as a whole. This is a bizarre view: overall, the gravity of the universe is weak. Since general relativity is seen as necessary to describe this largest scale, why would one expect GR to be unnecessary for the second and third largest scales in nature, those of the clusters of galaxies and of the galaxies themselves? To this point, when one encountered the expression "general relativistic dynamics", the reference was to those very special situations in nature such as the case of a closely orbiting pair of neutron stars where very strong gravitational fields and very high velocities prevailed. These cases are certainly very interesting but they are of very limited range. It is worth repeating: the new reality to be faced is that general relativity reaches into the dynamics of essentially all of the key basic building blocks in nature, the arrays of the billions of stars in the galaxies and the clusters of the galaxies themselves. In the vast majority of these cases, the gravity is not very strong and the velocities are not very high by standard relativity measure.

The importance in following this new path is immediate: without general relativity, one is left with serious issues first brought to bear from the work of F. Zwicky and V. Rubin in having to account for the higher-than-expected velocities of stars within galaxies and of galaxies within clusters, velocities having to be rationalized on the basis of Newtonian gravity. This had led to the belief that the normal baryonic matter that we see is but a small fraction of the total matter in the universe, that there is an immense quantity of so-called "dark matter" that is required to drive these anomalous velocities. This

matter was seen to reveal its existence solely by gravitation. Clearly if the gravitational laws that underpin this belief are removed, the paradigm shifts. Many would argue that the galactic motions are no longer the main reasons for belief in dark matter, that the primary reason is now the need for the extra matter to quickly form the galaxies and their clusters shortly after matter decoupled from radiation in the early universe. While early universe studies are the glamorous focus of current interest, some perspective is useful here: progress in early universe study has been impressive, but firm pronouncements as to what is required to explain COBE and WMAPs strike us as unjustified. Alternative scenarios should be, and surely will be explored in the years to come.

In Chapter 9, we will describe the paradigm shift due to general relativity. We will describe how S. Tieu and the author accounted for the high velocities of stars in spiral galaxies by the application of general relativity and without the requirement for the vast stores of exotic dark matter as is the case when Newtonian gravity is taken as the underlying theory of gravity. This raises the phenomenon of *general* relativistic velocity to a level of fundamental importance as an observational tool for situations of weak gravity and velocities far below that of light, domains previously reserved for Newtonian gravity. An essential point is this: where general relativity gives a result for a given physical system different from that provided by Newtonian gravity, we choose the general relativistic result. There is nothing controversial about this choice—the great majority of the world's physicists would do likewise as general relativity is regarded as our premier theory of gravity. The issue is the generally prevailing belief that the galactic systems that we studied should have given the same descriptions with the applications of both Newtonian gravity and general relativity. While our approach and departure from expectations have received a great deal of attention world-wide in the media and by many physicists and astronomers in hundreds of communications, it has also been opposed by an interesting variety of critics. We have answered their critiques in detailed papers and in Appendix A, we present a simpler account of both the nature of the criticisms and our replies to them.

In Chapter 10, we will then proceed to the next size scale to present an account of how we can rationalize the relatively large ve-

locities of galaxies within clusters without the aid of dark matter, again using general relativity. We will focus on our study of the spherical collapse under gravity of a pressureless ball of fluid as an idealized model of a freely gravitating cluster of galaxies. We will present the contrast between the local and asymptotic measures of the velocities of the particles, the former that would be perceived by observers in the vicinity of the galaxies and the latter that astronomers would actually perceive over the vast distances separating us from the populations of galaxies under study. It is well understood that even in classical physics, velocity is an observer-dependent quantity. In special relativity, that dependence assumes a more complicated form and it is even more complicated in general relativity. While these aspects of velocity have been understood, they have been under-appreciated in the case of general relativity. Our approach in taking these aspects into careful account leads to an alternative to dark matter as an explanation for the high galactic velocities seen in clusters of galaxies.

A key theme of this book is one of repositioning general relativistic as opposed to Newtonian dynamics as an essential tool for the complete description of the motions of bodies under gravity. Our work with Tieu on galactic dynamics focuses on the hitherto unjustified neglect of the application of general relativity and the interesting effects its incorporation produces. By contrast, in Chapter 11 we will also discuss our work on the *misuse* of general relativistic dynamics in the notion of "closed timelike curves", often interpreted as time machines, and how they can be seen most logically as mere mathematical rather than physical constructs.

Finally, in Chapter 12 we will discuss the overall direction of current theoretical physics research, its fixation on the need for unification of the fundamental forces with gravity. We will argue that there are reasons for regarding gravity as fundamentally different from the other forces in nature, in fact that it is not a force at all. Rather than the standard forces that serve as mechanisms coupling objects to each other as ingredients *within* spacetime, gravity is a property of spacetime itself, its curvature. This is the essence of gravity. While most relativists, if pressed, would acknowledge this to be the case, the focus has been lost (or never fully appreciated) by many. One aim of this book is to restore the realization of gravity's essence.

This book will have accomplished its mission if readers will appreciate the wonders that Einstein has brought to our recognition of relativity in a now much broader framework as one of the cornerstones of modern physics.

This page intentionally left blank

Chapter 2

Essentials of Special Relativity

2.1 Basic principles

Special relativity is Einstein's theory of space and time in the absence of gravity. It has at its base, Newton's principle of relativity which states that physics is the same in all inertial reference frames. An inertial reference frame is one in which a body moves with a constant velocity in the absence of an unbalanced force.

As well, special relativity incorporates the experimental fact that there exists a maximum velocity for the propagation of a signal in nature, namely c, the speed of light in vacuum. Simply put: there is a speed limit for interactions in nature.[a] Given this experimental fact and given that all inertial frames are physically equivalent, it follows that the speed of light in vacuum (the speed limit) must be the same, i.e. c, in all such frames. This is certainly counter-intuitive: if A throws a ball to B at speed V, B will measure the ball speed as V if B is at rest relative to A but B will measure a slower speed if B is running away from A and a faster speed if B is running

[a]The existence of a maximum speed for the propagation of interactions leads to the conclusion that there can be no truly rigid bodies in nature. For example, if one were to believe that a steel rod is truly rigid, a push leading to motion at one end would require an *instantaneous* pressure wave to flow to every other part so that each part would move in step. However such an infinite speed pressure wave does not exist and hence there is some buckling of the rod, however small [3]. Note the discussion of how close one can come towards perfect rigidity in [5].

towards A. If photons, the quanta of light, behaved like ordinary macroscopic balls, the velocity measurements for them would follow in the same manner. However, what works for balls does not work for light. We must constantly remind ourselves, as Feynman had done with such gusto, that in physics, it is the experiment that is the ultimate arbiter. While the injection of common sense reasoning is very important for our understanding of physical reality, it is the experiment that can over-ride our most cherished presuppositions, and when this occurs, we must yield.[b]

The constancy of the speed of light leads to the necessity that the notion of absolute time, a prime example of a cherished presupposition of classical Newtonian physics, must be abandoned. Intervals of time as well as intervals of space, the measured lengths of bodies, are seen to be relative to one's frame of reference. This in turn leads to the necessity of focusing upon "events" in space and time, not just where some event E has occurred, but where *and when* that event has occurred relative to the frame of reference that is being used to label it mathematically. It is essential because, as we shall see, the temporal relationship between events is no longer an absolute in relativity but rather it is dependent upon the reference frame of the observer. It should be noted that in normal parlance, the word "event" usually refers to some happening of significance. In physics, events need not have any particular significance and are generally defined simply as points in a spacetime plot, with time plotted on one axis and space plotted on axes perpendicular to the time axis.

A particularly shocking consequence of the invariance of the speed of light in vacuum is the necessity to abandon the seemingly commonsense notion that the simultaneity of events is absolute. To illustrate this, consider a railway car moving along the track at speed V. In the middle of the car, a match is struck and receptors fixed at both ends of the car record the arrival of the first photons from the flash. Clearly they record the arrivals at the same time as they are totally equivalent receptors. Within the frame of the car, there is no physical significance to their being at the back or the front of the car; the distance from the match is the same for both and the speed of the photons is the same in both directions.

[b]This applies with particular force in the bizarre domain of quantum phenomena.

However, observers standing along the track have a different picture of the events. They see the receptor at the back of the car moving towards the source of the light and the receptor at the front moving away from the source. While the photons are traveling to the back end, relative to the reckoning of the outside observers, the back end receptor has a *decreased* distance for the photons to reach it while the front end receptor has an *increased* distance for the photons to reach it. Since the speed of the photons as viewed by the outside observers along the track is the same value c in both directions, just as was the case for the rest frame of the car, clearly they will say that the event of photons reaching the back end of the car *preceded* their arrival at the front end. Simultaneity relative to the rest frame of the car does not translate into simultaneity in the frame of the outside observers along the track.

Had the outside observers calculated according to the classical compounding of velocities, they would have viewed the photon velocities towards the rear and front ends as $V-c$ and $V+c$ respectively rather than $-c$ and c and they would have deduced simultaneity in their frame as well. Counter to our intuition and experience with material objects, the speed V of the light source has no effect on how the outside observers gauge the velocity of the light. The key is the invariance of the speed of light in relativity in abolishing the absoluteness of simultaneity.

2.2 The spacetime interval and the Lorentz transformation

Consider two events: a photon emitted at (x_1, y_1, z_1) at time t_1 and absorbed at (x_2, y_2, z_2) at time t_2 relative to a reference frame K. Since the photon speed is c, the distance traversed can be expressed as $c(t_2-t_1)$ or as $\sqrt{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2}$. For any two events, we define the square of the spacetime interval s_{12} between them as

$$s_{12}^2 = c^2(t_2-t_1)^2 - (x_2-x_1)^2 - (y_2-y_1)^2 - (z_2-z_1)^2. \qquad (2.1)$$

Its value is seen to be 0 for the events on the path of the photon given the two ways of expressing the same distance traversed by the photon. Note that this interval squared has been defined with

differing signs, a positive time interval squared and a negative space interval squared. This is crucial in what follows and it incorporates the essential difference between time and space.

We could equally well consider the photon events from the point of view of a frame K^* and we would find (2.1) again with the symbols starred, apart from c which would be redundant since $c = c^*$ by experiment. Thus we have

$$s_{12}^2 = s*_{12}^2 = 0 \tag{2.2}$$

for the spacetime interval squared between photon events. This can also be expressed for infinitesimally separated photon events as

$$ds^2 = ds*^2 = 0 \tag{2.3}$$

where

$$ds^2 = c^2dt^2 - dx^2 - dy^2 - dz^2 \tag{2.4}$$

is now the expression for the infinitesimal spacetime interval squared.

It can be shown [3] that the invariance of the spacetime interval for events on the path of a light ray leads to the invariance of the spacetime interval for events that are *not* on the path of a light ray, i.e. when the interval as expressed in either (2.1) or (2.4), is not zero. *The invariance of the spacetime interval between events is an essential fact in relativity.*

Let us label an event as $E : (t, x, y, z)$ relative to observer frame K. Let an observer K^*, moving at speed V along the x direction carrying axes (x^*, y^*, z^*), label this same event. The K^* observer will assign spacetime coordinates to this event as $E : (t^*, x^*, y^*, z^*)$. It is straightforward to show that the familiar classical transformation between the coordinate systems $x = x^* + Vt^*, y = y^*, z = z^*, t = t^*$ incorporating the absoluteness of time, will not correctly relate the coordinates of the event in the two frames because this transformation will not preserve spacetime intervals in the form of (2.1) or (2.4) between distinct events. The transformation leads to an unwanted cross-term in dx^*dt^* for infinitesimally separated events. It is simple to verify that the correct transformation that preserves the form is the Lorentz transformation

$$x = \frac{x^* + Vt^*}{\sqrt{1 - V^2/c^2}}, y = y^*, z = z^*, ct = \frac{ct^* + Vx^*/c}{\sqrt{1 - V^2/c^2}}. \tag{2.5}$$

In Section 3.4, we will derive this transformation using measurements with beams of light.

The V/c term in the new time transformation and the square root factors display the smallness of the effect of relativity when the relative velocity between the frames is much less than c and the criticality of the effect when V is close to c. The transformation also contrasts the stark change that relativity imposes upon the relationship between t and t^* as compared with the Newtonian absolute time relation $t = t^*$.

With frame K^* moving with velocity V relative to K, it follows that K moves with velocity $-V$ relative to K^*. Therefore the inverse transformation is effected by replacing every un-starred coordinate with its starred counterpart while letting V go into $-V$. The result is

$$x^* = \frac{x - Vt}{\sqrt{1 - V^2/c^2}}, y = y^*, z = z^*, ct^* = \frac{ct - Vx/c}{\sqrt{1 - V^2/c^2}}. \qquad (2.6)$$

2.3 Lorentz contraction and time dilation

From the Lorentz transformation, it is easy to show that the length of an object as measured in its rest frame, the "proper length", is maximal: an observer moving relative to the object measures a shorter length, the phenomenon called "Lorentz contraction". Consider a meter stick at rest in the K frame along the x axis, its left end at x_L and its right end at x_R at all times t. The K frame observers measure its length l as $l = x_R - x_L = 1$ meter. Using the first of (2.5), we will deduce the length l^* of the stick as observed in the frame K^*. We have, for the relationship between the coordinates,

$$x_L = \frac{x_L^* + Vt_L^*}{\sqrt{1 - V^2/c^2}}, \quad x_R = \frac{x_R^* + Vt_R^*}{\sqrt{1 - V^2/c^2}}. \qquad (2.7)$$

By choosing different combinations for the values of t_L^*, t_R^*, we can realize any length that we wish for the stick length $l^* = x_R^* - x_L^*$ as viewed in K^*. This being the case, a logically useful *definition* of length is called for. This is the value of $x_R^* - x_L^*$ when taken at a *common* time t^* for the left and right end x^* values. Then, subtracting the first from the second of (2.6), the time terms cancel

and we have

$$x_R - x_L = \frac{x_R^* - x_L^*}{\sqrt{1 - V^2/c^2}} \tag{2.8}$$

Thus $l^* = l\sqrt{1 - V^2/c^2}$ and since $\sqrt{1 - V^2/c^2} < 1$ for non-zero V, we see that $l^* < l$, i.e. the stick appears shorter than 1 meter.

One might ask if this kind of derivation could have been applied just as well to the inverse transformation, the first of (2.6), to deduce the opposite, that it is the frame in which the stick is seen to be moving that shows a bigger length. This is not the case. Recall that the left end of the stick is at $x = x_L$ for arbitrarily chosen t and similarly for the right end. Thus, attempts to use this equation merely reveals that any length can be attributed to the stick in the frame K^* by choosing values of t appropriately.

A key aspect is the lack of symmetry in the comparison. The stick is at rest in one special frame and this sets it apart from all other frames in which it is seen to be moving. However, another kind of comparison does have symmetry. If observers in K measure the lengths of sticks that are at rest in K^*, the K observers will say that these sticks are shorter and if observers in K^* measure the lengths of sticks that are at rest in K, the K^* observers will say that these sticks are shorter. This is a comparison with perfect symmetry: each stick has its proper length in its rest frame and each stick is being compared symmetrically with length as judged in the adjacent moving frame. This reflects the total physical equivalence of the two frames of reference and the objects being compared.

We now consider how lengths of intervals in time compare. A time interval read in the rest frame of a clock, what we define as the "proper time", is minimal: observers in a frame moving relative to the clock would deduce that the time interval is longer, the phenomenon generally referred to as "time dilation". To see this, we now consider a clock at rest in the K frame and compare an interval of time that is read in the clock's rest frame K to the corresponding interval read in the frame K^* in which the clock is seen to be moving. Let the clock's position in K be x_C which is the case for all times t. Let the first click demarking the start of the interval be at time t_1 and the final click demarking the end of the interval be at time t_2. The length of the interval Δt is equal to $t_2 - t_1$. From the second of

(2.6), we have

$$ct_1^* = \frac{ct_1 - Vx_C/c}{\sqrt{1 - V^2/c^2}}, \quad ct_2^* = \frac{ct_2 - Vx_C/c}{\sqrt{1 - V^2/c^2}} \tag{2.9}$$

for the two ticks of the clock demarking the time interval. Subtracting the first equation from the second in (2.9), we have $t_2^* - t_1^* = \Delta t^* = \Delta t/\sqrt{1 - V^2/c^2}$. Since the square root factor is less than 1, we see that $\Delta t^* > \Delta t$. Again we see the reciprocal nature of space and time: lengths shrink and time intervals expand. Proper lengths are maximal and proper time intervals are minimal. Note that since ds^2 in (2.4) is an invariant and since $dx = dy = dz = 0$ in the rest frame of a clock, the proper time interval can also be characterized by ds/c.

2.4 Causality

Due to the presence of both positive and negative signs in ds^2, the spacetime interval squared, the values of ds^2 can be positive, negative or zero. Spacetime intervals that are zero in value are called "null" or "lightlike" intervals. Intervals whose squares are positive are called "timelike" and intervals whose squares are negative are called "spacelike". For discussions involving intervals and the important issue of causality, it is useful to define the "light cone". This is the geometric figure that the totality of possible light rays intersecting the base event traces out in a spacetime diagram. It is simplest to display this with two spatial dimensions in the x, y plane with time running in the axis perpendicular to this plane. The photons that are confined in space to the x, y plane and that intersect the spacetime point $(t, x, y, z) = (0, 0, 0, 0)$ trace out trajectories in this picture that are straight lines through the $(0, 0, 0, 0)$ origin point and are at constant slope with respect to the t axis. Clearly the assembly of all such photons traces out two cones, the one above the (x, y) plane displaying the photon events in the future of the base event and the cone below displaying the past photon events.

Events that are connected to a base event by a timelike interval, when displayed in a spacetime diagram are seen to be connected by a line that lies within the light cone[c] and hence can be causally re-

[c]For example, the history of a body at rest at the spatial origin $x = y = z = 0$

lated with an absolute sense of which event came "before" and which event came "after" the base event. By contrast, events separated by a spacelike interval to the base event are connected by a line outside of the light cone and hence can never be causally related to the base event. (No signal can exceed the speed of light in vacuum.) For such events, it is easy to show that one event can occur before, after or even simultaneous with the base event, depending upon the frame of reference chosen to label them. This is not a cause for consternation as such outside-the-light cone events can never be causally related to the base event. Such events can never occur at the same spatial position as the base event and hence are said to be "absolutely separated" from the base event.

This is to be compared to timelike separated events which can be seen to have occurred at the same spatial position by the right choice of reference frame but which can never be seen to occur at the same time. As well, for two events A and B that are timelike separated, if A occurs before B in one frame, it will be seen to be so in any other frame. For such events, there is a definite ordering in time. Let us say that event B occurred after event A in some reference frame. Then it will be so in all frames. This is so because to realize a reference frame in which there is a reverse order in time, one would have to proceed through a succession of transformations passing through the frame in which the events are simultaneous. But this is impossible, given that the events are timelike separated.

Having a definite ordering in time is consoling as the causal ordering connection that could link such events in a given frame must retain that ordering in all frames to avoid a logical inconsistency. In graphic terms for example, if event A were a mosquito biting one's arm and a later event B would be the slap of the mosquito into paste, it would be a physical absurdity were the events to be seen in reverse order in a different frame. While a film run backwards would show it as such, it would not be a realization of unfolding events in nature. The laughter that such reverse-run films provoke is a testament to their non-physicality.

With time intervals being non-absolute in relativity, changing from frame to frame as do the space intervals, it is useful to regard

traces out the t axis as its spacetime trajectory, the axis of the light cone, which is, of course, within the light cone.

the set of spacetime coordinates (ct, x, y, z) as the components of a four-dimensional vector $x^i = (x^0, x^1, x^2, x^3)$ with $i = 0, 1, 2, 3$, whose "length" squared is defined as $(x^0)^2 - (x^1)^2 - (x^2)^2 - (x^3)^2 = (ct)^2 - x^2 - y^2 - z^2$. This is a dimension step up from elementary three-dimensional mathematics with the added twist of a different sign for the square of the new fourth coordinate $x^0 = ct$ relative to that of the three space coordinates. It is necessary to incorporate this difference in keeping with the invariance of the spacetime interval under a transformation between coordinate systems. The difference in sign is the characteristic that makes time essentially different from space.

One often reads that relativity places space and time on the same footing with time being just another coordinate like the space coordinates. However, this is misleading: one should say that space and time, while being of equal significance in their alterability, are in a sense placed on a *reciprocal* rather than an equal footing. As well, it is essential to emphasize that our experience in nature is that while we can fix our spatial position, we cannot stop the flow of time.

Actually, x^i is more precisely referred to as a Lorentz "contravariant" four-vector. This name distinguishes it from a slightly different four-vector (with a subscript rather than superscript index) x_i, the Lorentz "covariant" four-vector. Its components are related to x^i as $x_0 = x^0, x_\alpha = -x^\alpha, \alpha = 1, 2, 3$. This is very useful as it enables us to express the invariant $c^2t^2 - x^2 - y^2 - z^2$ very succinctly as $x_i x^i$ with the repeated index denoting a summation over $0, 1, 2, 3$. Four-dimensional scalar products must always be formed with repeated indices where one is a subscript (covariant) and the other a superscript (contravariant).

2.5 Transformation of velocity and the aberration of light

We now consider how the motion of an object appears from the vantage point of the frames K and K^*. Let $v_x = dx/dt$ be the x-component of the velocity of the object and let v_y, v_z be the y and z components, all as viewed in K. Similarly, let $v_x^* = dx^*/dt^*$ be the x^* velocity component seen in K^* and v_y^*, v_z^* be the y^*, z^* components as seen in K^*. To compare the velocities seen in the two frames,

we take differentials of (2.5), keeping V, the given relative velocity between K and K^*, constant:

$$\begin{aligned} dx &= \frac{dx^* + V dt^*}{\sqrt{1 - V^2/c^2}}, \\ dy &= dy^*, \quad dz = dz^*, \\ cdt &= \frac{cdt^* + V dx^*/c}{\sqrt{1 - V^2/c^2}}. \end{aligned} \tag{2.10}$$

Dividing dx, dy, dz by dt in (2.10), we find the relationship between the velocity components in the two frames as

$$\begin{aligned} v_x &= \frac{v_x^* + V}{1 + V v_x^*/c^2}, \\ v_y &= \frac{v_y^* \sqrt{1 - V^2/c^2}}{1 + V v_x^*/c^2}, \\ v_z &= \frac{v_z^* \sqrt{1 - V^2/c^2}}{1 + V v_x^*/c^2}. \end{aligned} \tag{2.11}$$

This rather complicated transformation reverts to the simple Newtonian connection

$$v_x = v_x^* + V, \quad v_y = v_y^*, \quad v_z = v_z^* \tag{2.12}$$

in the Newtonian limit of velocities much smaller than c.

Using (2.11), we can relate the directions of the velocities of objects relative to the coordinate axes. With no loss in generality, we consider the motion of the object to be in the x, y plane with θ being the angle between the velocity vector (v_x, v_y) and the x-axis as viewed in K and θ^* being the corresponding angle as viewed in K^*. Thus

$$\begin{aligned} v_x &= v \cos\theta, \qquad v_y = v \sin\theta, \\ v_x^* &= v^* \cos\theta^*, \quad v_y^* = v^* \sin\theta^* \end{aligned} \tag{2.13}$$

where

$$v = \sqrt{v_x^2 + v_y^2}, \quad v^* = \sqrt{v_x^{*2} + v_y^{*2}}. \tag{2.14}$$

Thus, from (2.13) and (2.11),

$$\tan\theta = \frac{v^* \sqrt{1 - V^2/c^2} \sin\theta^*}{v^* \cos\theta^* + V}. \tag{2.15}$$

By setting $v^* = c$, the object becomes a photon and the formula in (2.15) becomes the relativistic expression for the aberration of light, the change in the direction of the propagation of light in changing to a new moving reference frame. In the Newtonian limit ($V \ll c$), the aberration formula reduces to the familiar expression

$$\Delta\theta = \theta^* - \theta = (V/c)\sin\theta^*. \tag{2.16}$$

2.6 Four-vectors and four-tensors

Just as x^i transforms to x^{i^*} as in (2.6), so too it is useful to define any set A^i of four functions of the coordinates x^i as the components of a general Lorentz four-vector. This is provided that these functions transform to new functions A^{i^*} in the same manner as x^i (i.e. as in (2.6) when one changes to a new coordinate frame x^{i^*}. Lorentz four-vectors (and four-tensors) are the basic mathematical objects in special relativity. The essence of what makes a vector or a tensor lies in the nature of the transformation of the object (the set of functions) when expressed in a different reference frame. For vectors, once we designate four arbitrary functions A^i of the coordinates x^i as the components of a Lorentz four-vector in the x^i coordinate frame, the new form of this four-vector A^{i^*} in any other Lorentz frame x^{i^*} is completely determined by the transformation (2.6) with A^0, A^1, A^2, A^3 replacing ct, x, y, z (as well as their starred counterparts) in (2.5).

A vector can also be described as a tensor of first rank. A set of 16 functions B^{ik} with i and k each taking on the values $0, 1, 2, 3$ are said to form the components of a second rank Lorentz four-tensor provided this set transforms as does A^i for each of the two indices in B^{ik}; similarly for third and higher rank tensors. The rank of a tensor is given by the number of indices attached to it. Thus, a vector is also a tensor of first rank. A scalar is a tensor of zero rank (no indices).

Since velocity[d] $v^\alpha = dx^\alpha/dt, \alpha = 1, 2, 3$ is of great importance in physics, we require the four-vector generalization, the four-velocity u^i. One might first be tempted to define this as dx^i/dt. But recalling the relativity of time intervals, while dx^i is a four-vector, dt is not a

[d] We adopt the convention of [3] that Greek indices range over the space indices.

scalar and hence the product of dx^i with $1/dt$ is not a four-vector. However ds is a scalar, and therefore four-velocity is defined as[e]

$$u^i = dx^i/ds. \tag{2.17}$$

Also of importance is four-acceleration w^i which in special relativity is naturally defined as

$$w^i = du^i/ds. \tag{2.18}$$

However, we will see in what follows that this procedure for acceleration is inadequate in general relativity.

Another important four-vector is the energy-momentum four-vector p^i,

$$p^i = mcu^i. \tag{2.19}$$

where m is the mass. The spatial components $p^\alpha = (p_x, p_y, p_z)$ constitute the relativistic components of the linear momentum

$$p^\alpha = mv^\alpha/\sqrt{1 - v^2/c^2}. \tag{2.20}$$

and the time component p^0 is E/c where E is the relativistic energy

$$E = mc^2/\sqrt{1 - v^2/c^2}. \tag{2.21}$$

From this, we have the base level of energy of a body, the energy as measured in the body's rest frame ($v = 0$), to be the very important and familiar mass-energy identification

$$E = mc^2. \tag{2.22}$$

Straightforward calculations with (2.20) and (2.21) reveal the important connections beween relativistic energy, momentum and velocity

$$E^2 = p^2c^2 + m^2c^4 \tag{2.23}$$

and

$$p = Ev/c^2. \tag{2.24}$$

where $p = \sqrt{{p_x}^2 + {p_y}^2 + {p_z}^2}$, the magnitude of the three-momentum.

[e]Note that u^i thus defined is dimensionless. It could have been defined with ds replaced by ds/c to recover the usual dimensionality for velocity. However, it is customary not to do so and it becomes irrelevant in the increasingly common usage of coordinates for which c is taken to be 1, as we will adopt in later chapters.

2.7 Special relativistic dynamics

Having the basic elements in place, we turn to the question of how bodies move within the theory of special relativity. There is a basic principle that deals with this and much more. It is called the "Action Principle" or, more formally, the "Principle of Least Action". Many have come to regard this principle as the most basic in all of physics. In the broadest terms, it states that for any physical system, there exists an invariant integral called the "action" such that when the system evolves from state A to state B, it does so in such a manner as to minimize the action. The integral is evaluated over the region in which the system evolves. In the case of a body, it is over its path (a curve) in spacetime between the initial and final spacetime points. If we are considering the evolution of a field such as an electromagnetic field, it evolves over its domain, three-dimensional space between an initial and a final time. Accordingly, the action integral for this case is over spacetime, covering all of three-dimensional space and over time between the limits of the initial and final states. The invariant action integral is expressed in generality for arbitrary assumed evolutions and the physically correct evolution is the one that minimizes this integral. Just as in simple calculus, the minimum is found by setting the first derivative to zero but now, the derivative must be evaluated over a path rather than at a point. Performing these calculations requires a knowledge of the calculus of variations which is beyond the scope of this book. Readers who wish to follow this mathematics should consult [3].

In their most elegant form, the great equations of physics, the Maxwell equations of electromagnetism, the field equations of general relativity, even quantum mechanics can be realized through the Action Principle. Fermat's Principle in optics, describing the path of a light ray through a medium of varying refractive index as one that minimizes the time, is a realization of the Action Principle. It is interesting to contemplate the fact that the most basic phenomena of nature are tied to extrema. The challenge is to discover the appropriate Lagrangians to create the action integrals and for this purpose, there is no magic wand available. However, there is a widespread sentiment that all *bona fide* physical theory emerges from an action integral, that the Action Principle is paramount. Feynman [6] provides a highly recommended lively account of the Action Principle

in his characteristic style.

Richard P. Feynman (1918–1988) was one of the most distinguished and accomplished theoretical physicists of the 20th Century. His contributions to quantum electrodynamics earned him the Nobel Prize along with J. Schwinger and S. I. Tomonaga. His three-volume "Lectures on Physics" [6] continue to be a great resource for beginning students, graduate students and researchers in physics. One of the most colorful and inspiring lecturers, his interests and contributions spanned a variety of fields beyond physics.

For a free body in special relativity, the invariant available to us is the spacetime interval and hence the action integral S is simply

$$S = K \int ds, \quad K = \text{constant} \tag{2.25}$$

taken between the limit spacetime points $A = (t_1, x_1, y_1, z_1)$ and $B = (t_2, x_2, y_2, z_2)$ representing the spacetime points from which the body begins its motion to where it ends its motion.

The action integral is a solely mathematical construct. The actual physical path is revealed in setting the variation of the action to zero, in searching out from the infinite number of imaginable unphysical types of evolutions, the unique physical path that constitutes the minimum, i.e.

$$\delta S = 0. \tag{2.26}$$

The δ is the variational calculus derivative representing the first derivative of this integral taken not at a single point as in ordinary calculus but over the entire path between A and B.

Using the calculus of variations, the result is

$$\frac{du^i}{ds} = 0 \tag{2.27}$$

which tells us that the four-velocity u^i of the body is a constant. This comes as no surprise but it is useful to follow through this simplest application of the very important Action Principle (see [3]).

The constancy of u^i over the path implies, through (2.19) that the four-momentum p^i is constant, and hence the energy and linear momentum are constant.

While the choice of action integral for a free body is quite straightforward, the choice for more complicated systems is correspondingly more complicated. To be noted is that when the action is expressed as an integral over time, the integrand is the Lagrangian, L, the function that we first encounter in classical mechanics,

$$S = \int L dt. \tag{2.28}$$

It can be shown [7] that the dynamics embodied in the key physics equation (2.26) is equivalent to the Lagrangian satisfying Lagrange's equations

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}^\alpha}\right) = \frac{\partial L}{\partial x^\alpha}. \tag{2.29}$$

However, while the Lagrangian that we first encounter in classical mechanics is the difference between the kinetic (T) and potential (U) energies,

$$L = T - U = \frac{1}{2}mv^2 - U \tag{2.30}$$

the Lagrangian in relativistic physics applies to fields as well as particles and assumes a variety of complicated forms. As the simplest example, for a free body, the potential energy is a constant which can be taken to be zero for convenience. The comparison of (2.25) and (2.28) using

$$ds = cdt\sqrt{1 - \frac{v^2}{c^2}} \tag{2.31}$$

[which follows from (2.4)] yields the relativistic Lagrangian for a free body

$$L = Kc\sqrt{1 - \frac{v^2}{c^2}}. \tag{2.32}$$

For small velocities, the relativistic Lagrangian (2.32) must approach the classical mechanics Lagrangian (2.30) (with $U = 0$). We make this comparison by expanding the square root in (2.32) and retain up to the lowest non-vanishing order term in the velocity,

$$L \approx -Kv^2/2c. \tag{2.33}$$

Comparing the v^2 terms in (2.33) and (2.30) gives $K = -mc$ and hence the relativistic Lagrangian for a free body is[f]

$$L = -mc^2\sqrt{1 - \frac{v^2}{c^2}}. \tag{2.34}$$

With the relativistic Lagrangian (2.34), we apply the standard formalism to derive the relativistic momentum and energy:

$$p^\alpha = \partial L/\partial v^\alpha \tag{2.35}$$

or

$$p^\alpha = \frac{mv^\alpha}{\sqrt{1 - v^2/c^2}}. \tag{2.36}$$

The energy is[g]

$$E = p^\alpha v^\alpha - L = mc^2/\sqrt{1 - v^2/c^2}. \tag{2.37}$$

We see that the energy and momentum derived in this way match the expressions that we had found in (2.21) and (2.20) by using (2.19). However, the latter have the added virtue of showing us that $(E/c, p^\alpha)$ constitute a Lorentz four-vector. As a result, given the energy and momentum in one Lorentz frame x^i, by the Lorentz transformation, we know how to find these quantities in any other Lorentz frame x^{i^*} with the transformation format of (2.6):

$$\begin{aligned} p_x^* &= \frac{p_x - VE/c^2}{\sqrt{1 - V^2/c^2}}, \\ p_y &= p_y^*, \quad p_z = p_z^*, \\ E^* &= \frac{E - Vp_x/c}{\sqrt{1 - V^2/c^2}}. \end{aligned} \tag{2.38}$$

2.8 Relativistic Doppler shift

The power of Lorentz covariance is well-illustrated in the relativistic Doppler formula, displaying the difference in the frequency of light

[f] An additive constant could be retained but it has no physical significance and is most conveniently set to zero.

[g] Here α is summed over $1, 2, 3$.

perceived by an observer in motion relative to that observed in the rest frame of the source of light. The derivation requires the introduction of a new four-vector called the "wave four-vector" with symbol k^i. The time component k^0 is taken to be the angular frequency[h] ω divided by c and the three-vector k^α is defined by

$$k^\alpha = \frac{\omega}{c} n^\alpha \tag{2.39}$$

where n^α is a unit vector in the direction of wave propagation. We recall that while we can designate any set of four components k^i to constitute a Lorentz four-vector, the above formulated definition has value only if its components in a new frame x^{i^*} have the same significance in terms of frequency and wave propagation direction as they had in the original frame x^i. That they do indeed have this significance is seen from the fact that the inner product

$$k^i x_i = \omega t - k^\alpha x^\alpha \tag{2.40}$$

is the phase of the wave which is a scalar. By the "quotient rule" (see [8]), since x_i is an arbitrary Lorentz four-vector and the phase of the wave is a scalar (see [9]), the wave four-vector as defined above for all frames is truly a *bona fide* Lorentz four-vector.

Having this established, the work is essentially done; we know how Lorentz four-vectors transform, in the manner of (ct, x, y, z) in (2.5), i.e.

$$\begin{aligned} k^0 &= \frac{k^{0^*} + (V/c)k^{1^*}}{\sqrt{1 - V^2/c^2}}, \\ k^1 &= \frac{k^{1^*} + (V/c)k^{0^*}}{\sqrt{1 - V^2/c^2}}, \\ k^2 &= k^{2^*}, \quad k^3 = k^{3^*}. \end{aligned} \tag{2.41}$$

Consider the source of light to be at rest in x^i with the observer whose frame is x^{i^*} moving with velocity V relative to the source and in the x direction as usual. Let α be the angle between the direction of wave propagation and the x^* axis as viewed in the x^{i^*} frame. Then

$$k^{1^*} = (\omega^*/c)\cos\alpha. \tag{2.42}$$

[h] $\omega = 2\pi\nu$ where ν is the actual frequency.

Substituting (2.42) into (2.41) and using (2.39), we find

$$\omega^* = \frac{\omega\sqrt{1 - V^2/c^2}}{1 + (V/c)\cos\alpha}. \tag{2.43}$$

This is the Doppler shift formula in all accuracy, even for relativistic velocities. Note that with $V > 0$ and α between 0 and $\pi/2$, the observed frequency ω^* is less than ω, the red-shift. Also, for $V \ll c$ and α not close to $\pi/2$, (2.43) yields the familiar classical expression for the fractional change in frequency

$$\frac{\omega^* - \omega}{\omega} = \frac{\Delta\omega}{\omega} = -(V/c)\cos\alpha. \tag{2.44}$$

Returning to (2.43), consider the special case where $\alpha = \pi/2$, i.e. the observer sees the light arriving perpendicular to his[i] direction of motion.

In that case, $\cos\alpha = 0$ and

$$\omega^* = \omega\sqrt{1 - V^2/c^2}. \tag{2.45}$$

Thus, the observer sees a red-shift with the same value (as required by symmetry) whether he moves to the right ($V > 0$) or to the left ($V < 0$). It is very small (of order V^2/c^2) compared to the usual shifts for non-relativistic velocities. This is a purely relativistic effect, stemming from the non-absoluteness of time, without even an infinitesimal residue in pre-relativity physics. It is often referred to as the "transverse Doppler shift". By contrast, in classical physics, a frequency shift can only occur if there is a component of the light propagation velocity in the direction of motion of the observer.

To make sense of sources moving at very high velocities, astronomers must use the relativistic equations above to properly interpret their observations.

We have covered the essential aspects of special relativity in this chapter. For further details, the reader is directed to the classic treatise of Landau and Lifshitz [3].

[i]For this and subsequent references to a person in the abstract, "his" means "his/her" and "he" means "he/she".

Lev D. Landau (1908–1968) is generally regarded as one of the most distinguished theoretical physicists of the 20th Century. His major works spanned several areas of physics and he was awarded the Nobel Prize for his important development of the theory of superfluidity. His ten-volume Course on Theoretical Physics with E. M. Lifshitz remains a key research tool for physicists world-wide. Tragically, Landau suffered major injuries in a traffic accident from which he never fully recovered.

Before proceeding to general relativity, we revisit special relativity from the ingenious approach of Bondi.

This page intentionally left blank

Chapter 3

Bondi's k-Calculus Approach to Special Relativity

3.1 Introduction

Bondi [4] has developed a unique highly intuitive approach to special relativity that displays some of the essential characteristics with spacetime diagrams.

> Sir Hermann Bondi (1919–2005) was a distinguished mathematician and cosmologist. With T. Gold and F. Hoyle, he developed the Steady State theory of the universe, widely believed but eventually discarded upon the discovery of the cosmic microwave background radiation. As well as his development of the k-calculus approach to special relativity, Bondi wrote several important papers on a variety of subjects in general relativity.

It makes extensive use of the relativistic Doppler factor k that relates inertial reference frames in relative motion and the fundamental new aspect of relativity vis-a-vis Newtonian physics, the invariance of the speed of light. We will see how clearly this approach resolves the so-called twin paradox, a key stumbling block for many who are first introduced to relativity. The word "calculus" to describe the Bondi method is misleading as his method relies upon elementary

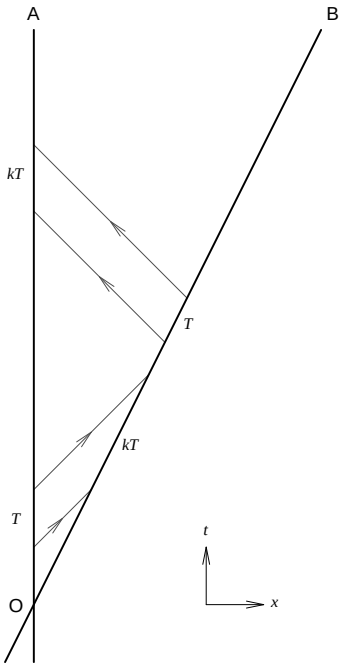

Figure 3.1: Light shone from Ava to Beta and vice-versa with intervals displaying their complete equivalence.

algebra and could be taught to junior high school students.

We consider two inertial observer twins Ava (A)[a] and Beta (B)[b] in relative motion. In Figure 3.1, the horizontal axis is x and the vertical axis is t. The diagram is displayed from the perspective of Ava who is shown to be at rest in (x, t) and Beta is seen to move in the positive x direction.

The spacetime diagrams that follow have the closest connections to simple plane geometry from the vantage point of the rest observer.

Some time after Ava meets Beta (the point at which their lines cross in the diagram and at which point they both set their clocks to 0), Ava shines light toward Beta for a period T by her (Ava's) clock. Beta receives the light for a proportional amount of time kT

[a] Ava–Greek–"An eagle".

[b] Beta–Czech–"Dedicated to God".

where the k constant symbol is the Doppler factor connecting the observers. The lines with arrows indicate light rays.

Later, Beta sends light toward Ava for the same period T and since Ava and Beta are totally equivalent physically, Ava observes the light for the same period as Beta had observed the earlier light from Ava, namely kT. Note that the geometrical appearance aspects such as the slopes of outgoing light rays, are necessarily biased in favor of the rest observer. Only relative to the rest observer are the light ray slopes always at 45 deg.

3.2 Velocity–Doppler factor connection

Ava wishes to determine Beta's speed relative to herself. For this, a different process is required, as shown in Figure 3.2. Immediately upon meeting Beta at O, Ava shines light to Beta for a period T and as in Figure 3.1, Beta receives the light for a period kT. Ava and Beta agree that as long as Beta receives light from Ava, she (namely Beta) sends light back to Ava.

Since this emission period from Beta is kT, the reception of Beta's light by Ava is the Doppler factor k times the emission period, i.e. $k(kT) = k^2T$.

From here it is a simple procedure to deduce the relative velocity between Beta and Ava. All that is required is to determine the distance between Ava and Beta at a convenient point and how much time has elapsed since their meeting at O to achieve that separation. The point P that is chosen is where Beta has received the last photon from Ava and has sent her last photon back to Ava. Ava reckons that she has sent her last photon reaching P at time T and has received the last photon back from Beta at time k^2T. Thus, the time for the photon to reach P and return is $k^2T - T$. Multiplying this by c, the photon speed, we get twice the OP separation. Therefore the OP separation is $D = c(k^2T - T)/2$. Ava also reckons that the last photon reached P at the half-way point in time from the emission at time T and the reception at time k^2T, namely at time $T^* = T + (k^2T - T)/2 = (k^2 + 1)T/2$. Finally, the relative velocity v is D/T^* or

$$v = c\frac{k^2 - 1}{k^2 + 1}. \tag{3.1}$$

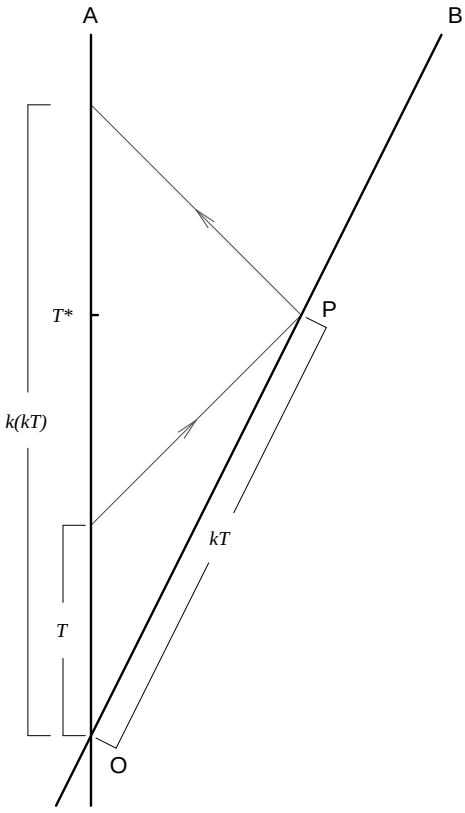

Figure 3.2: Determination of the speed of Beta relative to Ava.

Solving for k yields the relativistic Doppler factor k in terms of relative velocity[c]

$$k = \sqrt{\frac{1 + v/c}{1 - v/c}}. \tag{3.2}$$

The following properties are noted:

1) $k = 1$ for $v = 0$ which is logical since there is no Doppler shift when there is no relative velocity.

2) $k > 1$ if $v > 0$ which is logical since a relative recession entails an increase in period, decrease in frequency, increase in wavelength, or red-shift.

3) Similarly, $k < 1$ for $v < 0$. In this case there is a relative approach and hence a blue shift.

4) If $v \longrightarrow -v$ then $k \longrightarrow 1/k$.

3.3 Composition law for velocities and Doppler factors

We now determine the relativistic composition law for velocities. Consider a third observer Cayla (C),[d] introduced as in Figure 3.3. Let Ava emit light to Beta for a period T. Beta receives the light for a proportional time period $k_{\mathrm{AB}}T$ where k_{AB} is the Doppler factor between Ava and Beta. For as long as Beta receives the light, she transmits to Cayla who receives it for a period $k_{\mathrm{BC}}(k_{\mathrm{AB}}T)$ where k_{BC} is the Doppler factor between Beta and Cayla. At this point we invoke the key feature of special relativity, the invariance of the speed of light: we can view the direct transmission of light from Ava to Cayla for a period T being received by Cayla for a time k_{AC} with the same photon lines that were already used for the previous two-step process. This is because the light speed does not get boosted in the two-step process as compared to the direct transmission. As a result, we can equate the reception times by Cayla of the one-step and two-step process. Canceling the common factor T, we have

$$k_{\mathrm{AC}} = k_{\mathrm{AB}}k_{\mathrm{BC}}. \tag{3.3}$$

[c]The positive root solution is chosen to maintain the same direction of flow of time for the two observers.

[d]Cayla–"pure".

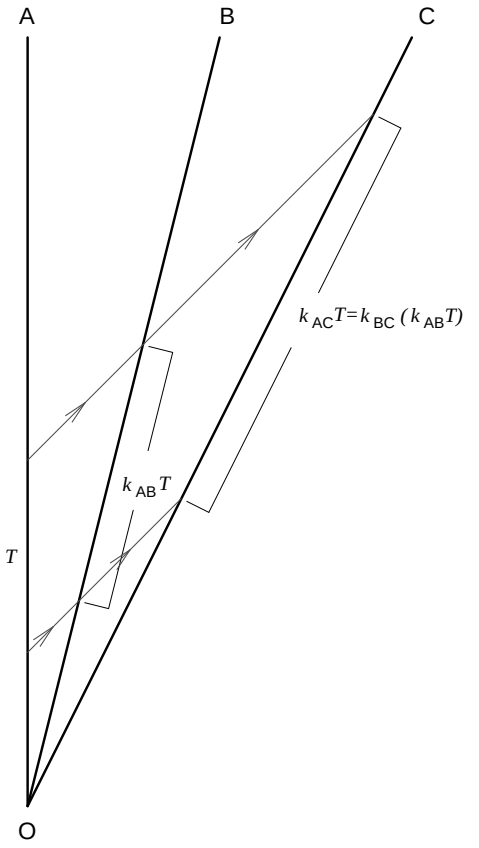

Figure 3.3: Determination of the compounding velocities in relativity.

Similarly, if we were to introduce a fourth observer Della (D)[e], we would have

$$k_{\mathrm{AD}} = k_{\mathrm{AB}} k_{\mathrm{BC}} k_{\mathrm{CD}} \tag{3.4}$$

and so on for an arbitrary sequence of observers. Let us now work in units where $c = 1$ (we will restore the symbol later). Returning to the three observers in Figure 3.3, from (3.1) and (3.3) we have

$$v_{\mathrm{AC}} = \frac{{k_{\mathrm{AC}}}^2 - 1}{{k_{\mathrm{AC}}}^2 + 1} = \frac{{k_{\mathrm{AB}}}^2 {k_{\mathrm{BC}}}^2 - 1}{{k_{\mathrm{AB}}}^2 {k_{\mathrm{BC}}}^2 + 1}. \tag{3.5}$$

Now using (3.2) repeatedly in (3.5) for the k factors in terms of relative velocities, after simplification we find (with the explicit c now restored)

$$v_{\mathrm{AC}} = \frac{v_{\mathrm{AB}} + v_{\mathrm{BC}}}{1 + v_{\mathrm{AB}} v_{\mathrm{BC}}/c^2}. \tag{3.6}$$

This is the familiar relativistic law for the composition of velocities. If we let c approach infinity, we retrieve the usual Newtonian velocity composition law

$$v_{\mathrm{AC}} = v_{\mathrm{AB}} + v_{\mathrm{BC}}. \tag{3.7}$$

If the relative velocities of the observers are small compared to c, the effect of using (3.6) instead of (3.7) is small. However, for "relativistic" velocities, the effect is dramatic. For example, with $v_{\mathrm{AB}} = v_{\mathrm{BC}} = 3c/4$, the correct composition law (3.6) yields $v_{\mathrm{AC}} = 24c/25$, a velocity less than c as must be the case and considerably different from $1.5c$ that would result from (3.7).

Of particular interest is the elegant simplicity of the composition law for Doppler factors as a simple multiplicative sequence in (3.4) in contrast to the awkward composition law for velocities in relativity (3.6). The simplicity of the former is understandable as it reflects the invariance of the speed of light and it is the Doppler factor that characterizes the connection between observers vis-a-vis light propagation. Note, however, that the Newtonian composition law, (3.7) (and its familiar extension to more observers, even non-collinearly) for velocities does have the simplicity that the Doppler factor composition displays. This is a reflection of the absoluteness of time in Newtonian physics. The Newtonian composition proceeds pictorially

[e]Della-English-"A woman from the island of Delos".

with vectors pasted end-to-end, having each observer's clock in step with all the rest.

In the next section, we display the non-absoluteness of time as an algebraic formula.

3.4 Derivation of the Lorentz transformation

The Lorentz transformation is the recipe for the labeling of event coordinates by one inertial observer relative to another. To derive this recipe, we return to our two twin observers Ava and Beta and consider an event E which Ava labels (t, x) and Beta labels (t^*, x^*). Ava and Beta synchronize their clocks when they meet at O and Ava decides to send a photon so that it arrives in coincidence with the event E. At that point, it is reflected back to Ava. Since for Ava, the event is at position x, the photon must have been sent from Ava at time $t - x/c$ to arrive at time t at E. It will arrive back to Ava a time x/c later, i.e. at time $t + x/c$.

Since the speed of the photon is also c for Beta, the photon intersects with Beta at time $t^* - x^*/c$ for the inbound path and at time $t^* + x^*/c$ for the outbound path. From our earlier discussion regarding the relationship between time intervals, we see from Figure 3.4 that

$$t^* - x^*/c = k(t - x/c) \tag{3.8}$$

where k is the Doppler factor between Ava and Beta.

Similarly we can focus on an emission time interval of $t^* + x^*/c$ from Beta to Ava with a reception time interval $t + x/c$ by Ava's reckoning. Therefore we have

$$t + x/c = k(t^* + x^*/c) \tag{3.9}$$

Eliminating x^* between (3.8) and (3.9), we find

$$2t^* = t(k + 1/k) - (x/c)(k - 1/k) \tag{3.10}$$

$$2x^* = x(k + 1/k) - (ct)(k - 1/k) \tag{3.11}$$

Substituting the expression for k from (3.2) into (3.10) and (3.11), we find as in (2.5), the important Lorentz transformation

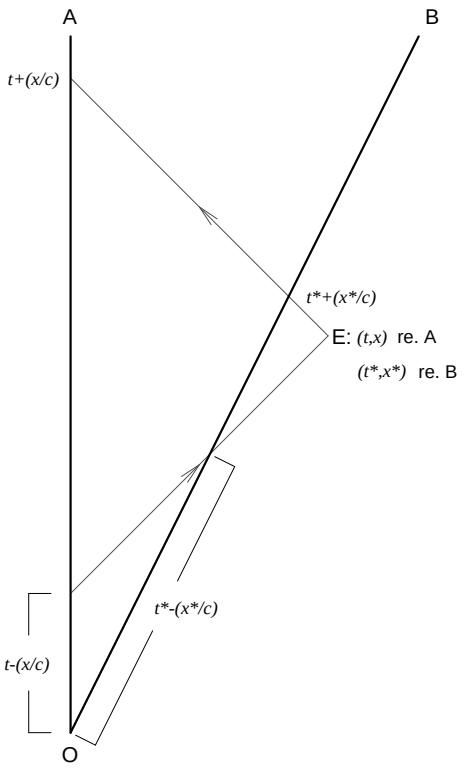

Figure 3.4: Derivation of the Lorentz transformation by the Bondi method.

$$x^* = \gamma(x - vt), ct^* = \gamma(ct - vx/c) \tag{3.12}$$

where $\gamma = (1 - v^2/c^2)^{-1/2}$.

In the second of (3.12), we see the complexity of the relationship between times read in the two frames in relative motion.

3.5 The twin or clock paradox

Because of the nature of time in the theory of relativity, it turns out that we can describe a very interesting scenario involving our pair of twins, Ava (A) and Beta (B). Ava stays home while Beta, the adventurer, takes off on a long journey at high speed, turns around eventually from homesickness and heads back to reunite with her twin sister Ava. If the conditions of time interval and speed are sufficient, Beta could return aged by let us say one year while Ava is long gone, unavailable for the planned reunion. Instead, Beta finds that she is meeting Ava's great-great-grandchildren. The supposed paradox consists of considering Beta to have been at "rest" while Ava is to have made the long journey and returned. Then it might at first glance appear from "relativity" that Ava should have aged only one year while Beta's space ship should have the future descendents of Beta emerging for the reunion. Logically it cannot be both. Which is correct?

The paradox is resolved using the k-calculus with some additional logical arguments. There is an essential asymmetry between Ava and Beta in this exercise in that while Ava follows an inertial spacetime trajectory, Beta undergoes a period of travel where she undergoes deceleration followed by acceleration. Acceleration and deceleration in her spaceship are translated into sensations that Beta experiences in her spaceship such as the variations in pressure against her seat. Such physical manifestations are not felt by Ava. Thus Beta's journey, unlike that of Ava, cannot be wholly one of following an inertial spacetime trajectory. This is best illustrated by bringing in our third observer Cayla (C) as shown in Figure 3.5.

At a velocity v, Beta leaves Ava at O where they synchronize their clocks. Immediately upon separation, Beta sends light back to Ava for a time T by Beta's reckoning and Ava receives this emission for a time kT. At the time T by Beta's clock, she meets Cayla who is

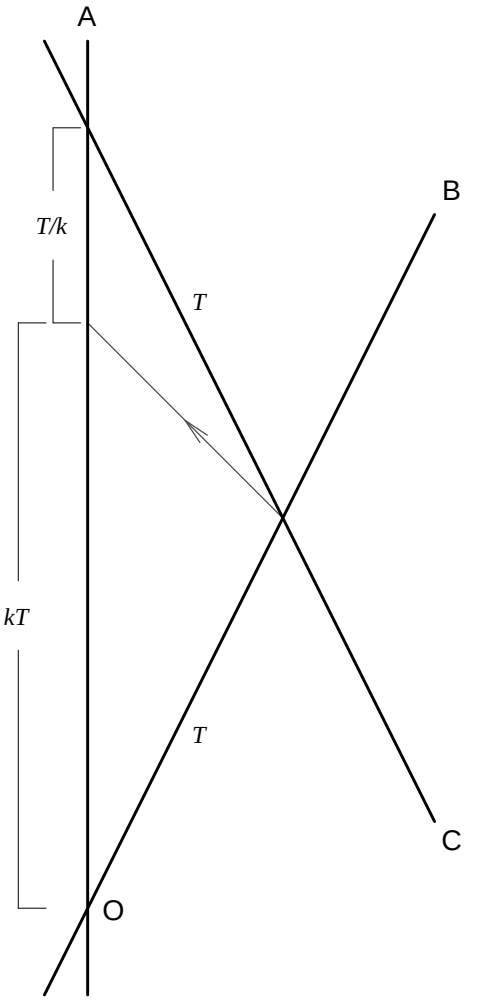

Figure 3.5: Comparison of times with three inertial observers.

traveling towards Ava with velocity $-v$. Beta and Cayla synchronize their clocks to the time T when they cross at which point Cayla begins to beam light to Ava until Cayla meets Ava.

The mathematics is very simple. By symmetry, Cayla beams her light to Ava for the same period T as Beta had beamed to Ava until she met Cayla. Thus when Cayla meets Ava, Cayla notes that her clock reads $T + T = 2T$ o'clock. We recall that when v goes to $-v$, k goes to $1/k$. Therefore the period of light reception by Ava of Cayla's transmission to her is $(1/k)T$. From Figure 3.5, we see that Ava has witnessed a period of time $kT + (1/k)T = (k + 1/k)T$ from the time she said farewell to Beta until she met Cayla. Since $k + 1/k$ is greater than 2 unless $k = 1$ (in which case there would not have been any relative motion), we see that Ava concludes that more time has elapsed for her than has elapsed for the combined journeys of Beta and Cayla between the three meetings.

Rather than introduce the third person Cayla, we could have considered Beta to have undergone a short deceleration period just before the point of Beta meeting Cayla followed by a short acceleration period. In this manner, her spacetime journey relative to Ava closely approximates the three-person plot of Figure 3.5. In this case, we have simulated the picture described earlier of Beta making a return trip and meeting Ava's great-great-grandchildren upon returning home. It is the deceleration/acceleration phase of Beta that is not present in the entirely inertial spacetime trajectory of Ava that makes Ava and Beta physically non-equivalent.

Some have argued that the periods of deceleration and acceleration will always compensate to remove the time difference and make Beta return to Ava at the same time. This is an untenable argument. Consider a sequence of journeys by Beta of *different* durations with the same velocities and with the identical reversals at the turnaround points as shown in Figure 3.6. Since the spacetime itself does not evolve in time, the physical effects that accrue at each one of the turnarounds must be identical because the turnarounds were identical. However, the return journeys of different durations require *different* amounts of compensation at the turnarounds to have the twins always unite with the same clock readings. Therefore the assumption that the non-inertial periods of travel will compensate and remove the time difference, is faulty.

It is also to be emphasized that while it is the period of acceleration by Beta that breaks the otherwise physically equivalent inertial observer symmetry of the journeys of Ava and Beta, it is the lengths in time of Beta's segments before and after the acceleration period that determine the extent to which their clock readings differ at the time of reunion.

Even more significantly, experiments with atomic clocks taken on return flights have displayed the effect. One might have objected that since Ava is always an inertial observer and Beta is an inertial observer for all but the very small spacetime trajectory segment near the turnaround point, it would seem that they should read essentially the same time upon reunion, i.e. that the spacetime segments are all straight lines apart from a very tiny segment. Bondi provides a very astute analogy as a counter to this argument. He considers journeys of Ava and Beta in the $x - y$ plane as shown in Figure 3.7, a plot in two *spatial* dimensions as in a conventional map. Ava's journey is a straight line and Beta's journey is almost a straight line apart from the kink that changes Beta's direction at the extreme point. It is that kink at R that breaks the symmetry and renders Beta's spatial distance covered longer than that covered by Ava. *The essential point is that in relativity, time is a route-dependent quantity just as distance is a route-dependent quantity.* It is also to be noted that it is the lengths of the segments before and after the kink that determine just how much longer Beta's journey will be than that of Ava. The existence of the kink makes the journeys non-equivalent but it does not determine how non-equivalent they are in distance covered. This extends the analogy with the twins regarding the degree of time difference between the clocks of the twins.

Another point to note is that while the greater *distance* covered is pictorially greater for Beta than for Ava in the space-space diagram of Figure 3.7 (the shortest distance between two points is a straight line), the shorter *time* for Beta is pictorially longer than that for Ava in the space-time diagram of Figure 3.6. This should come as no surprise. Time is not just another dimension like x, y or z. Space and time are different concepts.

Space and time are unified in Einstein's special relativity but they are not equivalent. They have a reciprocal connection, identified from the outset by the differing signs in the spacetime interval. We have

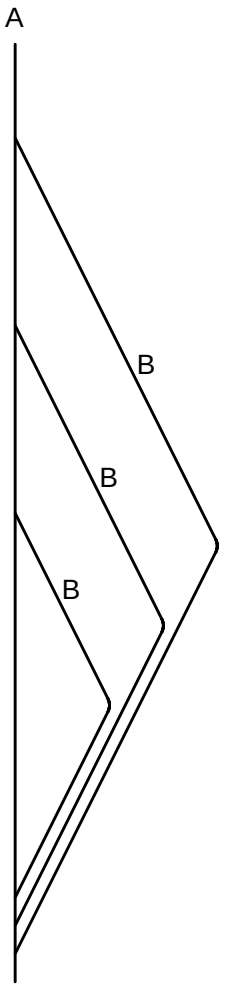

Figure 3.6: Comparison of times for two observers with journeys of different durations.

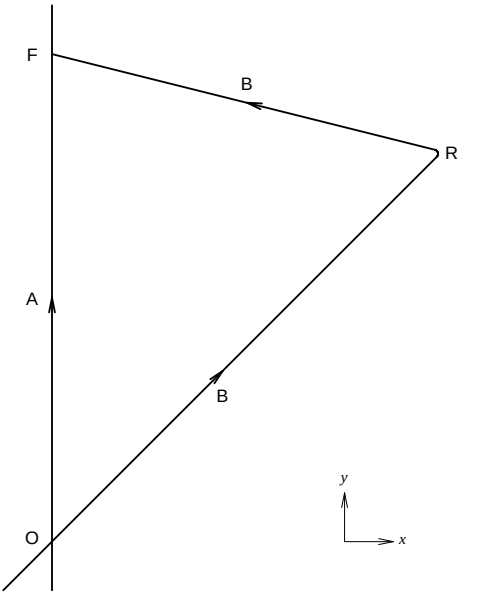

Figure 3.7: Comparison of distance covered for two observers over different routes.

outlined the essentials of special relativity, Einstein's theory of space and time in the absence of gravity. The incorporation of gravity into the relativistic framework is our primary focus in the chapters to follow.

Chapter 4

Essentials of General Relativity

4.1 The need for a new theory of gravity

Special relativity excludes gravity from its consideration. General relativity introduces gravity into the fold of relativistic physics. Indeed, *general relativity is Einstein's theory of gravity.* Its underlying basis is so novel that even from those who are versed in the theory, one frequently hears the old Newtonian ideas and prejudices about gravity brought into discussions involving general relativity. We will be returning to this issue in the chapters that follow.

Once special relativity was in place with its limit on the speed with which interactions could propagate, Einstein knew from the outset that Newton's theory of gravity would have to be replaced. In the sections that follow, we will trace through the logic leading to the new theory of gravity and its essential features. Newton's field equation for the gravitational potential ϕ

$$\nabla^2\phi(x,y,z,t) = 4\pi G\rho(x,y,z,t) \tag{4.1}$$

being an elliptic differential equation, necessitated that any change in the mass density distribution ρ would change the gravitational potential ϕ and hence the gravitational "force" (in the language of Newtonian gravity) on any distant object, with no time delay whatsoever, contrary to the essential dictum of relativity. If one were to move one's arm, in principle the entire universe of Newtonian physics

would be affected by this action instantaneously through the change in the gravitational field so produced.

4.2 The Principle of Equivalence

In the process of rectifying the problem of the instantaneous propagation of information in Newton's theory of gravity, Einstein was strongly influenced by the connection between the gravitational force and the so-called "inertial" forces, the centrifugal and Coriolis forces. These have the property, like the gravitational force, of being proportional to the mass. As a result, with the mass canceling out of Newton's Second Law $F = ma$, the acceleration is independent of the mass. (Recall the experiment of dropping a feather and a rock in an evacuated cylinder: both hit the bottom together.)

In fact Einstein imagined an experimenter in an elevator whose support cable had snapped, sending it into free-fall. The experimenter would now have the interesting sensation of weightlessness, as do the astronauts far from Earth after their vehicle's rockets are no longer firing. If the experimenter were to release objects from his pockets, they would be seen by the experimenter to hang suspended next to him. This is because they have the same acceleration *locally* relative to the Earth as does the person himself, and relative to his local (elevator) frame, he, along with the objects, is at rest. In the same vein, a rocket ship with this experimenter inside, and rockets firing, accelerating at 9.8 $\mathrm{m/s^2}$ would feel as if he were at rest on the surface of the Earth, now with his sensation of weight restored. Objects released from his pockets would "fall" relative to the rocket ship with an acceleration of 9.8 $\mathrm{m/s^2}$.

These observations are encapsulated into a general principle that is known as the Principle of Equivalence: "a gravitational field is locally equivalent to an accelerated reference frame". This property was retained by Einstein as one to be incorporated into a revised theory of gravity. The importance of the local nature of this property is frequently overlooked or minimized. This issue will arise in later chapters.

4.3 The metric tensor

The brilliance of the next step is easily underappreciated as we become so accustomed to its use. Within the framework of special relativity, i.e. spacetime *without* gravity, if we were to change from an inertial reference frame to an accelerated reference frame such as the frame of an accelerating rocket ship or of a rotating merry-go-round, we would be using a coordinate transformation that, unlike the Lorentz transformation (2.5), does not reproduce the metric form of (2.4). Rather, it imposes a coordinate-dependent structure upon the spacetime interval that we express in a very useful general form as

$$ds^2 = g_{ik}dx^i dx^k \tag{4.2}$$

where the "metric tensor" g_{ik} is in general dependent on all of the spacetime coordinates.[a] Since the coordinate-dependent metric tensor carries within it the fact that the coordinate system is now accelerated and since gravity is locally equivalent to an accelerated reference frame, it was natural (in hindsight!) to make the deduction that gravity itself should be characterized by the metric tensor, in concert with the Equivalence Principle.

However, here we must discuss an important issue concerning the essence of gravity. In the above description, while the transformation to an accelerated reference frame simulates gravity, the inverse transformation removes it and hence the actual spacetime is physically devoid of gravity. Gravity has only been simulated. It is then well to ask what constitutes *real* gravity. Is gravity nothing more than an artifact of an observer's reference frame or does gravity have an intrinsic nature, an intrinsic connection with spacetime?

For this, we consider a simple concrete example of real gravity produced by an actual source, the Earth. If we re-examine the freely falling observer and the objects that he releases in the elevator with its cables having been severed, we realize that those objects will not actually remain precisely at rest relative to the observer. Objects displaced in a vertical direction will separate in the course of time because those closer to the Earth will have a slightly greater ac-

[a] As before, a repeated index with Latin indices denotes a summation from 0 to 3. This is usually referred to as the "Einstein summation convention". Later in this section we will provide the mathematical framework for the metric tensor.

celeration relative to the Earth than those above them. Similarly, objects displaced in a horizontal direction will approach each other in the course of time during the free-fall because each one will fall towards the center of the Earth, i.e. on converging as opposed to parallel lines. This is in contrast to the situation of having the objects released within an accelerating rocket ship where all the objects undergo the same acceleration relative to the rocket ship, i.e. along parallel lines. Thus, the detailed picture is more complicated: the field of the simulated gravity is not actually the same as the field of true gravity. As we will see in what follows, the very essential difference concerns the geometry of spacetime itself, its curvature in the case of true gravity as opposed to lack of curvature, "flatness", in the case of simulated gravity. To describe the nature of curved spacetime mathematically, we take a brief detour into tensor calculus. This mathematics will enable us to formulate general relativity.

4.4 Basic tensor calculus–introduction

While we first encountered the particularized four-vectors and tensors in special relativity, the general description of such constructs is more complex. In general tensor calculus, while the situation with regard to scalars is the same as before, there are now two significantly different[b] kinds of vectors, "contravariant" vectors and "covariant" vectors. The prototype contravariant vector is the differential of the coordinates, dx^i. When we transform to a new coordinate frame x^{i^*}, by the chain rule of calculus, their differentials transform as

$$dx^{k^*} = \frac{\partial x^{k^*}}{\partial x^i} dx^i. \tag{4.3}$$

In analogy to the procedure in special relativity, we designate any set of four functions of the spacetime coordinates as the components of a general contravariant four-vector if under a transformation of coordinates x^i to x^{i^*}, the functions change to a new set in the same way that was the case with the coordinate differentials (4.3).

[b]Recall that there are also the two kinds of vectors in the special relativity described in Chapter 2 but they differ only in the signs of the spatial components while their time components are the same.

Similarly, we consider the prototype covariant vector, the gradient of a scalar function F, i.e. $\partial F/\partial x^i$. Again by the chain rule of calculus,

$$\frac{\partial F}{\partial x^{k^*}} = \frac{\partial F}{\partial x^i}\frac{\partial x^i}{\partial x^{k^*}}. \tag{4.4}$$

In both cases, we have transformed the quantities dx^i and $\frac{\partial F}{\partial x^i}$ from the "old" (unstarred) system of coordinates x^i to the "new" starred system x^{i^*}. They differ in that the coefficients for the transformation in the former, $\frac{\partial x^{k^*}}{\partial x^i}$, are the reciprocal forms relative to those of the latter. We now define any set of four functions A^i as the components of a contravariant vector if they transform in the same manner as the coordinate differentials (4.3) and any set of four functions B_i as the components of a covariant vector if they transform in the same manner as the gradient of a scalar function (4.4),

$$A^{k^*} = \frac{\partial x^{k^*}}{\partial x^i}A^i \tag{4.5}$$

and

$$B_k{}^* = B_i\frac{\partial x^i}{\partial x^{k^*}}. \tag{4.6}$$

Superscript indices denote contravariant and subscript indices denote covariant.

As before, we now have the prescription to define tensors of second and higher rank with each tensor index carrying a bank of transformation coefficients. Tensors can be wholly contravariant, for example A^{ik} as a second rank contravariant tensor, wholly covariant, for example B_{ik} as a second rank covariant tensor or "mixed", for example C_i^k for a mixed second rank tensor. The transformation law for the second rank contravariant tensor A^{ik} is

$$A^{lm^*} = A^{ik}\frac{\partial x^{l^*}}{\partial x^i}\frac{\partial x^{m*}}{\partial x^k}. \tag{4.7}$$

For the second rank covariant tensor B_{ik}, it is

$$B_{lm}{}^* = B_{ik}\frac{\partial x^i}{\partial x^{l*}}\frac{\partial x^k}{\partial x^{m*}}. \tag{4.8}$$

The mixed second rank tensor $C_i{}^k$ transforms with one of each type of partial derivative

$$C_l{}^{m*} = C_i{}^k\frac{\partial x^i}{\partial x^{l*}}\frac{\partial x^{m*}}{\partial x^k}. \tag{4.9}$$

The metric tensor in (4.2) is an example of a second rank covariant tensor, indicated by its two lower indicies. A vector is a tensor of rank 1. A scalar is a tensor of rank 0.

We introduce the contravariant tensor g^{ik} "conjugate" to the metric tensor g_{ik} as follows: we write g_{ik} as a matrix and produce its inverse in the usual way by forming the cofactor matrix of its transpose[c] and dividing by its determinant. As a result, g^{ik} is related to the metric tensor g_{ik} as

$$g_{il}g^{ik} = \delta_l^k \tag{4.10}$$

where δ_l^k is the Kronecker delta, i.e. the unit matrix, with 1 values on the diagonal and 0 everywhere else.

Related to a contravariant vector A^i is its "associate" covariant vector A_k defined by taking the inner product with the metric tensor as

$$A_k = g_{ik}A^i. \tag{4.11}$$

Similarly, the associate contravariant vector B^k of the covariant vector B_i is

$$B^k = B_i g^{ik}. \tag{4.12}$$

Similarly, this process of "raising and lowering indicies"[d] is applied to tensors of second and higher rank with the metric tensor used repeatedly for each index that is raised or lowered. It is often convenient to have expressions and equations expressed in covariant, contravariant or mixed form in different situations and it is the inner product with the metric tensor that accomplishes the transition.

It should be noted that in the flat space of special relativity in Cartesian coordinates as discussed in Chapter 2, we had the coordinates x^i forming vectors but in general tensor calculus, they no longer do so. In Chapter 2, a deviation from the precise formalism of general tensor calculus was permissible. There, we had x^i designated a contravariant vector and x_i designated a covariant vector with $x^0 = x_0$, $x^\alpha = -x_\alpha$.

One of the strengths of tensor calculus lies in the freedom to express equations in forms that are wholly general, applicable to any reference frame, a property frequently referred to as "general

[c] Taking the transpose is not actually required here because g_{ik} is symmetric.

[d] Some colleagues have remarked that all that general relativists do is raise and lower indices. Actually we do other things as well.

covariance". This is particularly appropriate for general relativity which deals with arbitrary systems of reference. For example, by the transformation rules discussed above, suppose in the frame x^i, there exists an equation of the form

$$A_{ik} = B_{ik}. \tag{4.13}$$

Then the difference of the elements in (4.13) is the tensor $C_{ik} = A_{ik} - B_{ik} = 0$. If we transform to a new frame x^{i^*} by the transformation rules discussed above, we see that in the new frame, $C^*_{ik} = A^*_{ik} - B^*_{ik}$ is also zero because every term multiplying a partial derivative of one frame with respect to the other is zero. Hence the equation is as in (4.13) with the quantities starred. Equations so-expressed are said to be "in covariant form".[e]

To probe the mathematics of curvature, we require the covariant generalization of the partial derivative of a vector or tensor. It can be shown that while $\partial F^i/\partial x^k$, or in the more compact $F^i_{,k}$ notation,[f] is a mixed second rank tensor in special relativity in Cartesian coordinates, it is not so otherwise [8]. The covariant generalization is

$$F^i_{;k} = F^i_{,k} + \Gamma^i_{kl} F^l \tag{4.14}$$

where

$$\Gamma^i_{kl} = \frac{g^{im}}{2}(g_{mk,l} + g_{ml,k} - g_{kl,m}) \tag{4.15}$$

is called the Christoffel[g] symbol of the second kind and a semicolon denotes "covariant differentiation". Note that in the flat space of special relativity in the reference frame where the metric tensor has the only non-zero components with the everywhere constant values $g_{00} = 1, g_{11} = g_{22} = g_{33} = -1$, their derivatives all vanish. Hence the Christoffel symbols all vanish and the covariant derivative reverts to the partial derivative. Moreover, it can be shown that $F^i_{;k}$ transforms like a mixed second rank tensor and is thus the required covariant

[e]The use of the word "covariant" in the two different senses is unfortunate but it is a universal dual usage.

[f]We will use the $(..),k$ notation frequently to denote the partial derivative of $(..)$ with respect to x^k.

[g]Note that in spite of its appearance, a Christoffel symbol is not a tensor since its (complicated) transformation law differs from that of tensors (see [8]). Students have been known to hyphenate the name after its first syllable to express their feelings about this symbol.

generalization for the partial derivative of a vector with respect to the coordinates.[h] Specifically,

$$F_{;l}{}^{m*} = F_{;i}{}^{k} \frac{\partial x^i}{\partial x^{l*}} \frac{\partial x^{m*}}{\partial x^k}. \tag{4.16}$$

Further, we require the covariant generalization of the derivative of a vector or tensor with respect to a scalar such as the spacetime interval. First, consider the derivative of the simplest tensor, the scalar function F. By the chain rule, this can be expressed as

$$\frac{dF}{ds} = \frac{\partial F}{\partial x^l} \frac{dx^l}{ds}. \tag{4.17}$$

Since the elements on the right hand side are tensors, this equation is tensorial as it stands. However, if we were to replace the scalar function F with a vector F^i or a tensor of higher rank, say F^{ik} in this equation, the derivative with respect to x^l of F^i and F^{ik} would no longer be a tensor and the equation would not be tensorial. Clearly the generalization required is the replacement of the partial derivative with respect to x^l by the covariant derivative. This yields the universally applicable "intrinsic derivative", denoted by D/ds, of the tensor to which it is applied, the covariant generalization of the ordinary derivative with respect to a scalar. As applied to F^{ik}, the equation is

$$\frac{DF^{ik}}{ds} = F^{ik}_{;l} \frac{dx^l}{ds}. \tag{4.18}$$

It can be applied in a similar manner to any tensor.

4.5 Parallel transport, spacetime curvature and the Riemann tensor

We now have the essential mathematical tools in position to explore the very important spacetime curvature. To do so, we begin with the process of translating a vector F^i parallel to itself along a curve

[h] Details of the tensor calculus for these results as well as for what follows can be found in [8] and with the physics woven in at [3]. This mathematical description is meant for physicists; for mathematical rigor, there are a variety of mathematics texts available.

that we parametrize by its arc length s, a process generally referred to as "parallel transport". In flat space in Cartesian coordinates, clearly the components of the vector do not change in the process, $dF^i/ds = 0$. If we were now to change to a general coordinate system to express this equation, we must express this parallel transport with the intrinsic derivative

$$\frac{DF^i}{ds} = 0. \tag{4.19}$$

It is natural to define parallel transport by this covariant equation (4.19) even when the spacetime is curved.

At this point, the question arises: how does one distinguish locally between a flat region and a curved region in a mathematically precise manner? The answer lies in having a vector F^i (or a more general tensor) undergo parallel transport around an infinitesimally closed loop and tracking the net change in the components over one circuit. In a flat space, since the components of a vector do not change when there is parallel transport, they will not change even when the loop is closed. However, when the space is curved, a lengthy calculation [3], [10] shows that in generality, the net change ΔF^i is

$$\Delta F^i = R^i_{klm} dx^k dx^l F^m \tag{4.20}$$

where the important Riemann tensor R^i_{klm} is given by

$$R^i_{klm} = \Gamma^i_{km,l} - \Gamma^i_{km,l} + \Gamma^i_{nl}\Gamma^n_{km} - \Gamma^i_{nm}\Gamma^n_{kl}. \tag{4.21}$$

This fourth rank tensor carries within itself the essence of curvature, hence the essence of gravity. It vanishes in all of its components if and only if the spacetime is flat, i.e. in the absence of curvature or, physically speaking, in the absence of true gravity. Ordinarily, given its fourth rank, it would be expected to have 256 (i.e. 4^4) components in four-dimensional spacetime. However, because of its various symmetries [8], this number is reduced to 20 independent components. The symmetries are

$$R_{iklm} = R_{lmik}, R_{iklm} = -R_{kilm}, R_{iklm} = -R_{ikml} \tag{4.22}$$

and

$$R_{iklm} + R_{imkl} + R_{ilmk} = 0. \tag{4.23}$$

As an exercise to familiarize oneself with the mathematics of curved space, it is useful to step down in dimensions and consider the parallel propagation of a vector in the simplest interesting curved two-dimensional surface, that of a sphere. It is developed in some detail in [8] where it is seen that in general, the vector that is parallely propagated, does not return to its original form on traversing a closed path. The exception is the very important case when the chosen curve is a "geodesic", the curve of shortest distance between two given points. For the sphere, these curves are the great circles.

An excellent visualization of the effect of curvature on the parallel propagation of a vector is provided by cutting thin bands of material around a selection of meridians of a sphere and laying them flat on a plane. (Bands carefully cut from a hollow rubber ball or the peel of a grapefruit would serve the purpose.) Note that the bands now lie in curves on the plane apart from the special band encircling the equator that unravels along a straight line. On the plane, parallel translation can be done visually. If a vector is now drawn at the start of these bands and repeatedly drawn parallel to the original until the other end is reached, the following results are seen: when the bands are repositioned on the sphere, the end vectors, now rejoined at the same point, do not match in general. However, they do match in the special case of the equatorial band, the geodesic. In performing this exercise, we are exploiting the fact that the local geometry of this curved surface is approximately flat.

4.6 Geodesics

Mathematically, the geodesic is characterized as the curve along which its tangent vector u^i is parallely propagated,

$$Du^i/ds = du^i/ds + \Gamma^i_{kl} u^k u^l = 0. \tag{4.24}$$

As we are familiar from elementary physics, what is the tangent vector for mathematics is the velocity for physics, in our case four-velocity. The equation of motion of a free particle in special relativity in Cartesian coordinates is $du^i/ds = 0$. In an arbitrary coordinate system, this is expressed in the generally covariant form $Du^i/ds = 0$, i.e. (4.24). The Equivalence Principle guides us to adopt (4.24) as well for the equation of motion of a free particle in a gravitational

field. *This is one of the key equations of general relativity.* It must be said that the Equivalence Principle has served as a very useful guide indeed, in spite of its being only an approximation to truth. To first order, the solutions of these equations recover the closed elliptic motions of the planets around the Sun. For the weak field of the solar system, the corrections due to general relativity are very small.[i] To first order, the second term with the Christoffel symbol in (4.24) plays the role of the gradient of the gravitational potential in Newtonian gravity. Examining to greater accuracy, here as in various instances in general relativity, the possibilities and subtleties as compared to Newtonian gravity, are richer and more interesting. In general, we see that there is a nonlinear interplay between the metric tensor and its partial derivatives that govern the motion of bodies in general relativity. This equation was used with great success in accounting for the residual 43 seconds of arc per century perihelion precession of the planet Mercury, a major source of confidence in the correctness of general relativity. We will discuss this further in Section 5.4.

4.7 Covariant conservation laws and the Einstein field equations

The next key issue concerns the determination of the metric tensor for a given physical situation: what are its sources and what are the field equations that will connect the field to source? From (4.1), we recall that in Newtonian gravity, the field ϕ has mass density ρ as its source. In the appropriate limit, general relativity must reduce to Newtonian gravity since the latter, for hundreds of years, has served physics well under the right conditions. So we seek the appropriate relativistic mathematical structure that incorporates the density ρ. Since we observed that it is the metric tensor that is playing the role of the gravitational potential ϕ and since the former is a second rank tensor, the natural choice is the energy-momentum tensor T^{ik}. From special relativity, this is a second rank tensor that has the energy density as its T^{00} component. (The $\frac{T^{0\alpha}}{c}$ components comprise the momentum density or $1/c^2$ times the energy flux density and the $T^{\alpha\beta}$

[i] See however in later chapters where the corrections become important for many-body systems with comparable mass constituents, even when the field is weak.

components represent the stresses and the momentum flux densities.) Thus, to maintain a covariant structure, we replace ρ with T^{ik} on the right hand side of the generalized equation that we seek.

For consistency, we must have a second rank tensor incorporating the gravitational field (now the metric tensor) with appropriate structure, on the left hand side. In this regard we have a constraint: energy-momentum conservation. Recall that conservation of a quantity is expressed mathematically by the vanishing of its divergence. For example, in electromagnetism, the conservation of charge is expressed as[j]

$$J^{\alpha}_{,\alpha} + \frac{\partial \rho}{\partial t} = 0 \tag{4.25}$$

where here, J^{α} is the three-current ρv^{α}, ρ is the charge density and v^{α} is the velocity of the charges. With ρc expressed as J^0 and ct as x^0, (4.25) can be written as the vanishing ordinary four-divergence

$$J^{k}_{,k} = 0. \tag{4.26}$$

Similarly, in standard special relativity, the conservation of energy-momentum is expressed by the vanishing of the ordinary four-divergence of the energy-momentum tensor

$$T^{ik}_{,k} = 0. \tag{4.27}$$

If we wished to work in special relativity in general coordinate systems, the conservation equations of (4.26) and (4.27) are employed with $(..),k$ replaced by $(..);k$, i.e. the partial derivative replaced by the covariant derivative. As well, to convert (4.26) and (4.27) to their corresponding forms applicable to general relativity, i.e. to incorporate conservation of charge or energy-momentum, we simply substitute the partial derivative with a covariant derivative, i.e. a semi-colon in place of a comma.

Thus, the latter becomes

$$T^{ik}_{;k} = 0. \tag{4.28}$$

Clearly this is also the expression for energy-momentum conservation in general relativity. For consistency, the tensor to incorporate the

[j] As before, a repeated Greek index denotes a sum over the spatial indices 1, 2, 3.

effects of gravity on the left hand side of the field equation that we seek must also have a vanishing covariant divergence. Moreover, for consistency with the differential structure of the Newtonian limit, there must be no higher than a second derivative of the metric tensor present in this tensor. After some effort, Einstein deduced that the required tensor is (now appropriately named) the Einstein tensor

$$G^{ik} = R^{ik} - \frac{1}{2} g^{ik} R \tag{4.29}$$

where the Ricci tensor R^{ik} is the contraction[k] of the Riemann tensor given by R^i_{kim} and the Ricci scalar R is the contraction of the Ricci tensor, R^i_i. The Einstein tensor has an identically vanishing covariant divergence as a consequence of the contracted Bianchi identities [3] [8].

These Bianchi identities are

$$R^i_{jkl;m} + R^i_{jmk;l} + R^i_{jlm;k} = 0. \tag{4.30}$$

After appropriate inner multiplication of this identity with the metric tensor and two contractions, the result is

$$R^i_{j;i} - \frac{1}{2} R_{;j} = 0. \tag{4.31}$$

Thus, we have the replacement for $\nabla^2 \phi$ of Newtonian gravity on the left hand side of the field equations in the form of the Einstein tensor G^{ik}. On the right hand side, we have the covariant source of gravity, the energy-momentum tensor T^{ik} replacing the mass density ρ of Newtonian gravity. Finally, incorporating all of these factors, we have the very important Einstein field equations, the essential equations of general relativity

$$G^{ik} = \frac{8\pi G}{c^4} T^{ik}. \tag{4.32}$$

These equations are 10 in number because of the symmetry in i and k. The numerical coefficient $\frac{8\pi G}{c^4}$ of the right hand side has been

[k] A contraction is the setting of a covariant index equal to a contravariant index and summing from 0 to 3. Each contraction reduces the rank of a tensor by 2. Thus the Ricci tensor is of rank $4 - 2 = 2$ and the Ricci scalar R is of rank $2 - 2 = 0$, named "scalar" appropriately.

chosen to have the $(i, k) = (00)$ term of the Einstein field equations reduce to (4.1) in the Newtonian limit. (It is to be noted that the derivation of the Einstein equations is achieved with the full power of the Action Principle in [3]. Note also that the standard symbol G in this equation is the constant of universal gravitation. It is not to be confused with the trace G^i_i of the Einstein tensor.)

Many physicists have come to view the Einstein equations (4.32) as the most beautiful or even the most important equations in all of physics. Beauty and importance are of course subjective terms. One could argue that the Maxwell equations of electromagnetism or the Schrödinger and Dirac equations of quantum mechanics are also beautiful and in the case of Maxwell, find far more direct and practical utility than do the Einstein equations. However the existence of such testaments to the significance of (4.32) is presented to convey the sense of impact that they have had on physics. One can only stand in awe of the power of Einstein's imagination to identify the curvature of spacetime geometry with the phenomenon that we know of as gravity. While pre-Einsteinian physics had come to see gravity as just another field, conceptually as the fields of electricity and magnetism, general relativity had cast the matter and fields other than gravity onto the right hand side of (4.32) and embodied gravity uniquely into the Einstein tensor of the left hand side as a feature of geometry. In his later years, Einstein's quest was for the unification of all fields, for the removal of the energy-momentum tensor on the right hand side with its replacement by a geometrical structure to appear on the left hand side as had been achieved with the geometrization of the gravitational field.

However, there is an alternative view that we favor. We see all particles and fields as existing *in* spacetime but that gravity *is* spacetime, i.e. its curvature. Gravity is not a field in the sense of other fields and it is not a force like other forces. We are comfortable in seeing gravity as unique and therefore the quest to unify gravity with the other forces in nature as uncompelling at best. (We will return to this issue in Chapter 12.)

4.8 Einstein–Maxwell equations and motion of a charged body in general relativity

Electromagnetic fields and charges exist within spacetime. They are governed by the Maxwell equations, the set of differential equations for the Maxwell tensor. The Maxwell tensor incorporates the components of the electric and magnetic fields [3]. The first set of Maxwell equations incorporating the four-current J^i as source for the Maxwell field are

$$F^{ik}_{,k} = \frac{4\pi}{c} J^i \tag{4.33}$$

in Cartesian coordinates in flat space (i.e. for standard special relativity). In arbitrary coordinate systems (as well as to incorporate gravity for general relativity), these become

$$F^{ik}_{;k} = \frac{4\pi}{c} J^i. \tag{4.34}$$

The second set of Maxwell equations for flat space in Cartesian coordinates are

$$F_{ik,l} + F_{li,k} + F_{kl,i} = 0. \tag{4.35}$$

For arbitrary coordinate systems and hence also to incorporate gravity, the commas are replaced by semi-colons in (4.35). However, it turns out that after the replacement, all of the extra terms cancel identically and hence the original equation with partial derivatives is correct as it stands in general relativity as well as in flat space. It is remarkable that the entire set of Maxwell equations are converted to the form applicable to general relativity merely by the replacement of a single comma with a semi-colon!

The Einstein equations (4.32) with the energy-momentum tensor for the electromagnetic field [3]

$$T^{ik} = \frac{1}{16\pi}(g^{ik} F_{ab} F^{ab} - 4F^{ij} F^k_j) \tag{4.36}$$

provide us with what are sometimes referred to as the "already unified field equations" for gravity and electromagnetism. The complete set of Einstein–Maxwell equations are (4.34), (4.35) and (4.32) with (4.36) in "electro-vacuum". They are suitably augmented within matter with additional contributions to T^{ik}. An example, frequently

applied, would be one of adding the energy-momentum tensor for a perfect fluid to the right hand side of (4.36).

In special relativity, a particle of charge e, mass m and four-velocity u^i moving in an electromagnetic field F^{ik} has its motion determined by the Lorentz-force equation[l]

$$mc\frac{du^i}{ds} = \frac{e}{c}F^{ik}u_k. \tag{4.37}$$

This becomes

$$mc\frac{Du^i}{ds} = \frac{e}{c}F^{ik}u_k \tag{4.38}$$

to incorporate gravity. This is the equation for the general relativistic dynamics of charges.[m]

4.9 Summary of the steps from Newtonian gravity to Einstein's general relativity

Let us summarize the sequence of steps leading to the basic equations of general relativity:

i. The understanding that effects in nature propagate with at most speed c led to the rejection of the Newtonian equation for gravity, (4.1), as the latter embodies infinite propagation speeds.

ii. Locally, the dynamical effects of gravitation are equivalent to accelerated reference frames (the Principle of Equivalence).

iii. The acceleration of a reference frame is revealed through its modification of the form of the metric tensor g_{ik}.

iv. Because of the *local* gravity-acceleration equivalence and the invariance of the spacetime interval of special relativity, the metric tensor is seen as the replacement for the Newtonian gravitational potential ϕ in a relativistic theory of gravity.

[l] This is with neglecting radiation reaction. The inclusion of radiation reaction is a rich and interesting subject in its own right (see [3] and [6]).

[m] Of course it is also the equation for the special relativistic dynamics (where gravity is neglected) of charged particles in arbitrary coordinate systems.

v. In general relativity, the replacement for the mass density ρ as the source of the gravitational potential of Newtonian gravity is the symmetric second-rank energy-momentum tensor T^{ik}. This tensor, already an important element in special relativity, is also naturally chosen for its generally covariant character and its embodiment of ρ in its T^{00} component.

vi. The conservation of energy and momentum are expressed by the vanishing of the ordinary divergence of T^{ik} in special relativity in Cartesian coordinates and by its vanishing covariant divergence in general coordinate systems. The latter generally covariant property is required for the incorporation of gravitation.

vii. In the quest for the appropriate replacement G^{ik} for $\nabla^2\phi$ of the left hand side of the Newtonian gravity field equation, the focus turns to a search for a second rank symmetric tensor with vanishing covariant divergence that is composed of the metric tensor and no higher than its second partial derivatives.

viii. The root for this tensor lies in the geometry of curved spaces and the essential fourth-rank Riemann tensor R_{iklm} (4.21) which is the key to spacetime curvature.

ix. That the Riemann tensor provides an invariant local measure of curvature is seen in its role in changing a vector under parallel propagation around an infinitesimal closed path, (4.20).

x. The covariant expression for conservation of the quantity expressed as a vector or tensor is the vanishing of the covariant divergence of the vector or tensor as in (4.28) for energy and momentum conservation.

xi. From the contracted Bianchi identities (4.31), the required Einstein tensor G^{ik} is found in the form (4.29). The Einstein field equations are thus determined as (4.32) with coefficient $8\pi G/c^4$ to reduce to the Newtonian gravity equation in the appropriate limit.

xii. The equation of motion of a freely moving body in spe-

cial relativity is expressed in an arbitrary system of coordinates as $Du^i/ds = 0$ (4.24). By the Equivalence Principle, this must also be the equation of motion of a free body in a gravitational field, i.e. in general relativity.

Chapter 5

Schwarzschild Solution and its Consequences

5.1 The metric

Over the years, thousands of papers have been devoted to the study of the Einstein equations. In contrast to the single linear Poisson equation (4.1) for Newtonian gravity with only the mass density as source of gravitational field, the Einstein equations are ten highly nonlinear partial differential equations for the metric tensor g_{ik} with many terms and having as source, the full richness of the energy-momentum tensor T^{ik}. By "nonlinear", we mean that there are products of the functions present in the equations. This makes the mathematics far more complicated. Researchers have found a variety of exact solutions and have studied their properties. Physicists have explored the application of certain solutions to physical problems.

Most prominent among these solutions is Schwarzschild's exact solution for the spherically symmetric field in vacuum. We imagine a spherical mass such as the Sun or other spherical body as source and concentrate upon the spacetime geometry in the vacuum outside of the source. Much of the early history of general relativity centers around the analysis of this solution, both in its exact and approximate forms. It can be shown [11] that the most general metric form for spherical symmetry using spherical polar coordinates can be expressed with only two arbitrary functions of r and t in the form

$$ds^2 = e^{\nu(r,t)}dt^2 - e^{\Lambda(r,t)}dr^2 - r^2(d\theta^2 + \sin^2\theta d\varphi^2). \qquad (5.1)$$

An important theorem by Birkhoff has shown that in vacuum, the solution is intrinsically static. This has the interesting implication that arbitrary spherically symmetric disturbances of the interior source cannot have any consequences for the vacuum region that surrounds it; the exterior field must remain unaltered.

The solution in simplest time-independent form is[a]

$$e^{\nu} = 1 - \frac{2m}{r}, \qquad e^{\Lambda} = \left(1 - \frac{2m}{r}\right)^{-1} \tag{5.2}$$

reflecting the fact that this field is intrinsically static.

5.2 The measurement of distance and time in general relativity

At this point, it is appropriate to develop different aspects of measurement in relativity. In the special relativity sections, we introduced the concepts of proper distance and proper time, distance and time as read, respectively, in the rest frames of a meter stick and a clock. There, the focus was on Cartesian coordinates where the coordinate differentials dx, dy, dz were physical distance and dt was physical time, quantities that were not merely of mathematical significance. These quantities were proper measure or non-proper measure depending upon whether the frame of reference was or was not the rest frame of the stick or clock in question. Regardless, they were of physical significance.

Now suppose that we are *not* using Cartesian coordinates. As a background to the new features to consider, we turn to the most familiar and simple mathematics. Consider, in elementary spatial geometry, the expression for an infinitesimal distance dl squared, first in the Cartesian coordinates (x, y, z)

$$dl^2 = dx^2 + dy^2 + dz^2. \tag{5.3}$$

It can also be expressed as

$$dl^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2 \tag{5.4}$$

[a]It should be noted that while one can express the solution in explicit time-dependent form (and we will do so in Section 5.3), the intrinsic character remains static. In what follows, we will generally use units in which $G = c = 1$. Thus, what would appear in the conventional form as $2Gm/c^2r$, becomes $2m/r$ in (5.2).

in spherical polar coordinates (r, θ, ϕ).

While the coordinate differential dr is distance in the radial direction, the coordinate differentials $d\theta$ and $d\phi$ are not the distances in the polar and azimuthal directions. Rather, these distances are $rd\theta$ and $r\sin\theta d\phi$ respectively. Note that these quantities are formed from the square roots of the *total* quantities present in the sum of three separate parts of the line element, not just of the coordinate differentials. Thus, if we were to write (5.4) as

$$dl^2 = g_{rr}dr^2 + g_{\theta\theta}d\theta^2 + g_{\phi\phi}d\phi^2, \tag{5.5}$$

we would write the actual distances $(dl_r, dl_\theta, dl_\phi)$ in the three directions as

$$(dl_r, dl_\theta, dl_\phi) = (\sqrt{g_{rr}}\, dr, \sqrt{g_{\theta\theta}}\, d\theta, \sqrt{g_{\phi\phi}}\, d\phi). \tag{5.6}$$

While $g_{rr} = 1$ in (5.4), it is *not* 1 in (5.1) (with(5.2)). In the latter, not even dr is an actual physical element of distance. Clearly, for the Schwarzschild metric, the true physical distance in the radial direction, what we call the "proper" radial distance dl_r(proper), is

$$dl_r(\text{proper}) = \sqrt{g_{rr}}dr = (1 - 2m/r)^{-1/2}dr. \tag{5.7}$$

Thus, with gravity present, physical distance takes on a different and interesting character. While at great distances from the central body, $2m/r \ll 1$ and dl_r(proper) is approximately equal to dr, for r close to $2m$, we have $dl_r(\text{proper}) \gg dr$. Thus, gravity has the effect of modifying the measure of physical distance in terms of the coordinates. The length intervals dr perceived by distant observers are much smaller than the corresponding length intervals dl_r(proper) perceived by a local observer at r when r is close to $2m$. They approach zero as r approaches $2m$.

Even more interesting is that gravity also modifies the measurement of time. Just as $\sqrt{g_{rr}}dr$ is the physical or proper measure of radial distance, clearly $\sqrt{g_{tt}}dt$ is the physical or proper measure of time. In general relativity, if g_{tt} is not equal to 1, t is not the real time. In general, t is often referred to as the "time-like coordinate" or the "time parameter". In the Schwarzschild spacetime, just as r gained its character as physical radial distance for $r \gg 2m$, so too g_{tt} approaches 1 for very large r and t gains its character as physical or proper time. However for r close to $2m$, the proper time measure differs greatly from the quantity t. As we expect from previous

arguments, we witness the reciprocal behavior for time. Noting that

$$dt(\text{proper}) = \sqrt{g_{tt}}dt = (1 - 2m/r)^{1/2}dt, \tag{5.8}$$

we see that $dt(\text{proper}) \ll dt$ for r close to $2m$. The time intervals dt perceived by distant observers are much greater than the corresponding time intervals $dt(\text{proper})$ perceived by a local observer at r when r is close to $2m$. They approach infinity as r approaches $2m$.

5.3 The event horizon, black holes and singularities

In the previous section, when we spoke of drastic changes in distance and time intervals that arise when r is close to $2m$, the question naturally arises: what if $r = 2m$ and indeed, what if $r < 2m$? These are interesting issues and they have been discussed and debated, at times with great intensity, from the early years of general relativity.

Normally one would say that with length intervals going to zero and time intervals going to infinity as perceived by the distant observers, there must be a singularity at $r = 2m$. A singularity represents a breakdown of physics and some new theoretical construct is then called for to remove it. Indeed Einstein and N. Rosen certainly felt this way.[b]

> Nathan Rosen (1909–1995) was arguably Einstein's most important collaborator. According to A. Pais, Rosen was the key contributor to the famous EPR (Einstein–Padolsky–Rosen) paradox paper bringing into question a key element in quantum mechanics regarding observables. Other famous works with Einstein included the papers on cylindrical gravitational waves and the Einstein–Rosen bridge. He was an inspiration to many.

In his early work, Einstein considered a spherically symmetric swarm of test particles and showed that as the swarm condensed towards a radius where $r = 2m$, the speed of the particles approached the speed of light. Thus he saw the $r = 2m$ surface as a physical barrier that cannot be realized in nature. Later, Einstein and Rosen

[b] In his later years, Rosen related to us that he remained of this opinion.

[12] constructed a solution of two Schwarzschild spacetimes meeting at their mutual $r = 2m$ surfaces (now referred to as the "Einstein–Rosen bridge"). The aim was to excise the region at and within the troublesome surface.

Since then, a variety of authors debated the issue (see [13] for a review). A major sticking point had been the issue as to whether a test particle would actually attain the speed of light if dropped to $r = 2m$. The answer is really very simple: relative to a sequence of observers adjacent to the particle, the observers being at rest relative to the central body, the velocity does indeed approach c. However, these adjacent observers have a harder and harder time trying to be rest observers as they are situated closer and closer to $r = 2m$. In fact, it would be impossible for them to remain at rest at $r = 2m$ because to do so would require an infinite amount of reverse thrust of rockets that they would have to carry to remain rest observers. Thus the answer is that no observer can make this measurement.

An important insight was provided by Synge [14].

> John L. Synge (1897–1995) was a highly distinguished mathematician and physicist. One of the most prolific authors in mathematical physics, classical physics and relativity, he was a man of great intellectual depth and imagination.

He showed that the spacetime trajectory of such a test particle lies not on the light cone (which would have made the $r = 2m$ surface undisputedly singular) but within the forward light cone at $r = 2m$ and therefore the particle would be following a physically allowed timelike trajectory. Proceeding inwards, as the particle gets closer to $r = 0$, the light cone at the particle's position folds into a narrower and narrower funnel until at $r = 0$, the light cone degenerates into a straight vertical line (see Figure 5.1). Then the particle trajectory merges with the light cone. At that point, everyone to our knowledge agrees that the particle at $r = 0$ is at a true singularity.

Another argument that is used to demonstrate that the $r = 2m$ surface is benign, concerns the Kretschman scalar K. This is the scalar product of the Riemann tensor with itself in the form

$$K = R_{iklm}R^{iklm}. \tag{5.9}$$

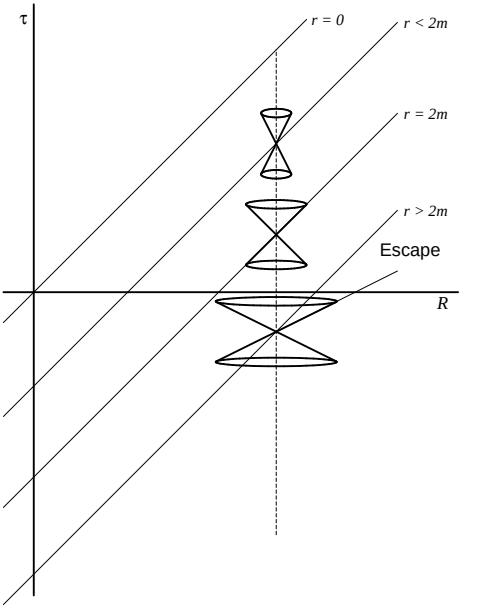

Figure 5.1: Free-falling particle, trajectories at fixed distances and the light cones.

Since K is a scalar, its value has an invariant significance. For the Schwarzschild spacetime,

$$K = 48m^2/r^6. \tag{5.10}$$

Its value is an innocuous $3/4m^4$ at $r = 2m$ but at $r = 0$, K is infinite. This bears out the view expressed earlier, that while the $r = 2m$ surface is non-singular, the $r = 0$ point is truly singular. But the story does not end here.

In the present coordinates it is troublesome to deal with the analysis since metric components g_{tt} and g_{rr} become 0 and ∞ respectively at $r = 2m$. To investigate the dynamics of a particle further, a new system of coordinates is called for. Various authors have provided these including Eddington, D. Finkelstein, G. Szekeres and M. Kruskal. We will focus on the system described in [3]. These authors change from the standard Schwarzschild coordinates (t, r) to (τ, R) by the transformation[c]

$$\tau = t + \int \frac{\sqrt{2m/r}}{1 - 2m/r} dr, \quad R = t + \int \frac{\sqrt{r/2m}}{1 - 2m/r} dr \tag{5.11}$$

with (θ, ϕ) left unchanged. Taking differentials of (5.11) and substituting into (5.1, 5.2) yields[d]

$$ds^2 = d\tau^2 - (2m/r)dR^2 - r^2 d\Omega^2, \quad d\Omega^2 = d\theta^2 + \sin^2\theta d\phi^2. \tag{5.12}$$

[c]Note that the choice of signs in the first of (5.11) is the one appropriate to collapse (see [3]).

[d]To achieve the new form of ds^2 in (5.12), a cancellation of a factor $(1-2m/r)$ is required. This is not problematic for all non-zero r except for the value $r = 2m$. All transformations of this kind which remove the infinity in the metric at $r = 2m$ are also required to invoke a cancellation from a fraction of the form $0/0$ or the equivalent at $r = 2m$, a point seldom brought up in the literature. When we have raised it, some have responded to this point with the suggestion that the transformation that brought the form of (5.12) into existence be ignored and that the new form be treated as the starting point of investigation. While this can be done, and there is merit to this stance as there is no longer any hint of unusual behavior in the new coordinate system, it does hide the fact that the spacetime is intrinsically static for $r > 2m$. The generally accepted present attitude is that there is nothing more sinister about the singular aspect of the metric in Schwarzschild coordinates than there is about the coordinate singularity at $r = 0$ in polar coordinates describing flat space. Thus, what used to be called the "Schwarzschild singularity" for $r = 2m$ is now called a coordinate singularity or event horizon. The early name is now accorded to the singularity at $r = 0$ whose singular character is unassailable.

This new form is a hybrid in the sense that while it is expressed in the new (τ, R) coordinates, the old radial coordinate r is still present in the metric coefficients. To render the expression entirely in the new coordinate system, we note from (5.11) that[e]

$$R - \tau = \int \sqrt{\frac{r}{2m}}\, dr = \frac{2r^{3/2}}{3\sqrt{2m}}. \tag{5.13}$$

Hence the r in (5.12) can be replaced by

$$r = \left[\frac{3}{2}(R-\tau)\right]^{2/3} (2m)^{1/3}. \tag{5.14}$$

Now the metric is expressed entirely in terms of the new coordinates (τ, R, θ, ϕ). With this replacement, we see that the metric has explicit time dependence through the presence of the timelike coordinate τ. While the advantage in having this form is that there is no explicit unusual behavior apart from where $R = \tau$ (the $r = 0$ singularity), what is masked in this form is the fact that the spacetime is intrinsically static for $r > 2m$ and it is intrinsically dynamic for $r < 2m$.

Its intrinsic static character for $r > 2m$ is revealed by the fact that there exists a coordinate system in which the metric has no t dependence in any of its components. This is the coordinate system at rest relative to the central body. By contrast, for $r < 2m$, the intrinsic dynamic character is revealed by the fact that no coordinate system can be found in which the metric components are independent of the time coordinate. This statement might appear to be contradicted as (5.1, 5.2) could be used for $r < 2m$ in addition to the usual application when $r > 2m$. However, the unusual twist is that for $r < 2m$, the r coordinate becomes the timelike coordinate and the t coordinate becomes the spacelike coordinate. This is because the metric coefficients of dt and dr in ds^2 change signs as the $r = 2m$ boundary is crossed. For $r < 2m$, the reader might feel more comfortable in replacing r by t^* and t by r^*. Thus, the point $r = 0$ is not really a "where" but rather a "when", more aptly expressed as $t^* = 0$. In the new (R, τ) coordinates, each particle participating in

[e] Again, this final simplified expression requires a cancellation in the form $\frac{1-2m/r}{1-2m/r}$.

the collapse has its particular R value for all time. The proper time τ which is the "when" for each particle to reach the singularity is given by the translation of the condition $t^* = 0$ to the new coordinates, namely $R = \tau$ (c being 1). Rosen [15] has expressed considerable concern about these and related aspects.

The metric in the new coordinate system has the structural form[f]

$$ds^2 = d\tau^2 + g_{\alpha\beta}dx^\alpha dx^\beta \tag{5.15}$$

(i.e. no space-time cross terms and $g_{00} = 1$) where

$$\begin{aligned} g_{RR} &= -(2m)^{2/3}\left[\frac{3}{2}(R-\tau)\right]^{-2/3}, \\ g_{\theta\theta} &= -(2m)^{2/3}\left[\frac{3}{2}(R-\tau)\right]^{4/3} = \frac{g_{\phi\phi}}{\sin^2\theta}. \end{aligned} \tag{5.16}$$

A spacetime with metric of the form (5.15) is said to be in "synchronous" form [3]. For such a case, an observer who has fixed spatial coordinates in this system is *physically* in free-fall.[g] It can be visualized by imagining the observer in free-fall dragging his spatial coordinate system with him as he falls. It is an easy exercise to prove this: setting (x^1, x^2, x^3) all constant, we have $u^\alpha = dx^\alpha/ds = 0$. Also, with $dx^\alpha = 0$, $ds^2 = d\tau^2$ which implies that $d\tau/ds = 1$. Thus the four-velocity for such an observer is $u^i = (1, 0, 0, 0)$. If the observer with this four-velocity is truly in a state of free-fall, his four-velocity must satisfy the geodesic equation that we discussed earlier,

$$\frac{du^i}{ds} + \Gamma^i_{kl}u^k u^l = 0. \tag{5.17}$$

Using the definition of the Christoffel symbol (4.15) with these components of u^i and the synchronous form of the metric, it is very easy to show that (5.17) is indeed satisfied and the result is proved.

We will concentrate upon radially moving observers and photons in what follows. Thus, θ and ϕ will be set to constants and will play

[f] Recall that Greek indices range over $(1, 2, 3)$.

[g] At this point, we can appreciate why the metric had explicit time dependence when expressed in the form (5.12) with r given by (5.14): From the vantage point of the free-fall observer, the central body appears to be approaching him in the course of the fall and his view of the world is a continually changing one.

no role in the dynamics. We now concentrate on one particular free-fall observer. Since free-fall observers have fixed spatial coordinate values in this coordinate system, this observer's spacetime trajectory is the vertical line (R = constant) in the (R, τ) plot in the Figure 5.1.

Having considered fixed R values, we now consider fixed r values. From (5.14), we see that a given fixed r value corresponds to the fixed value for $R - \tau$ given by (5.14), i.e. a straight line of slope 45 deg in the (R, τ) plane. In the Figure 5.1, we draw a series of such trajectories for different fixed r values. Physically, such trajectories correspond to different observers who are at different fixed physical distances from the body outside of which they are hovering. (They must blast retro-rockets with appropriate thrusts to maintain their fixed positions rather than to fall towards the central body.) In the Figure 5.1, we have drawn such trajectories for $r = 0, r < 2m, r = 2m$ and $r > 2m$.

We now consider radially moving light rays. Recall that light rays have zero spacetime intervals. Thus, setting $ds = 0$ as well as $d\theta = d\phi = 0$ in (5.12), we have

$$d\tau/dR = \pm\sqrt{2m/r}. \tag{5.18}$$

Thus we see that the slopes of the light cones at different points in the trajectory of the free-fall observer are:

a) slope < 1 for $r > 2m$

b) slope $= 1$ for $r = 2m$

c) slope > 1 for $r < 2m$.

Light cones have been drawn in the Figure 5.1 for these three cases.

Much interesting information can be gleaned from the figure. As the free-fall observer occupies steadily diminishing r values, his choices vary. When his r value is greater than $2m$, he can decide to reverse his motion by firing rockets and blast his way to larger r values. That he can do so is indicated by the existence of a wedge of space in the diagram between the 45 deg constant r line at his position and the positive sloped trajectory of the photon at his position. This is indicated as the "escape" wedge in the figure. The choice of escape is available to the free-fall observer because trajectories that proceed into the escape wedge are timelike and hence physically per-

missible trajectories. They are physically permissible because they lie within the forward light cone.

At this point, the free-fall observer can escape with the help of rockets (at which point he ceases to be in free-fall) or he can continue to move inwards towards the center, to smaller r values. Similarly, there are inward moving photon (negative slope) and outward moving photon (positive slope) trajectories available at this point. While the negative sloped photon trajectories proceed toward the source center, the positive sloped trajectories proceed to larger $r's$. These photons will escape.

Should the observer choose to continue in his fall, he will eventually reach the r value of $2m$. At that point, his choices are severely diminished. Any future-directed timelike trajectory (i.e. into the forward light cone) for him at that point is heading towards smaller r values: he can no longer escape and he cannot even stay at the $r = 2m$ position. He must continue to move towards the source center. Moreover, the photon trajectory with the positive slope is coincident with the constant r (at $2m$) trajectory so in a sense, this photon hovers at $r = 2m$. The reader might question this as photons always travel at speed c. They do so indeed, relative to local observers. For example, the free-fall observer, as he crosses the $r = 2m$ surface, would measure the speed of this photon to be c. However, from the vantage point of the distant observer who notes that the photon is always at $r = 2m$, he would say that the photon is at rest as far as he is concerned! Of course the frequency is infinitely red-shifted so this is a statement in principle.

This is an excellent example of the relativity of velocity in general relativity. It will be of great importance in later chapters.

If we examine the situation for $r < 2m$, we see that even the "outward" (positive sloped) photon trajectory actually proceeds to smaller r values. In other words, all physical trajectories, whether they be timelike or null (in the case of photons), must be inward-directed for $r < 2m$.

The surface $r = 2m$ is the divider between being able to escape and/or to send messages to the outside versus being trapped and being unable to communicate with the world beyond $r = 2m$. Once that surface is reached, all contact with the world outside is impossible. For this reason, the $r = 2m$ surface is referred to as the "event

horizon". A body, once having reached a point of collapse where its r value crosses the $2m$ threshold, is said to have formed a "black hole". It is referred to as black because not even light can emanate from it.

At $r = 0$, the light cone closes up into a vertical line. The free-fall observer merges with the light cone in a singularity, a total breakdown of physics. Some researchers accept such singularities as possible occurrences in nature. Probably most researchers do not regard them as a real part of physics. R. Penrose proposed a hypothesis that systems proceeding toward complete gravitational collapse will invariably develop event horizons, preventing the outside world with having to face the prospect of an unphysical singularity. It has become known as the "cosmic censorship hypothesis", the censor protecting us from unsavory sights. Outside observers would be shielded from this horror and the singularity within the horizon could form without any physical crisis. However, even if the hypothesis should prove to be true, what can be said for the observer who does cross the horizon? He is not shielded from the singularity. Therefore, in our view, even if the hypothesis should prove to be correct (at least in some sense), the Penrose hypothesis is not the shield for physics that is often believed. Moreover, we [16] showed that for a range of equations of state, spherical gravitational collapse will lead to the formation of a "naked singularity", a singularity that is not shielded by an event horizon.

There are a variety of views regarding the issue of singularities. Some take the stance that the singularity at $r = 0$ does not actually form, that it is illegitimate to accept classical general relativity at such extreme levels of spacetime curvature. Rather, it is argued, quantum mechanics somehow takes over and prevents the breakdown of physics, that matter does not actually crunch to zero volume as a result of gravitational collapse. If the matter did manage to reach the state of zero volume, there is a further issue to face.

If we assume that time must continue to advance, the question arises as to where that matter goes at that point. As we discussed previously, once we have $r < 2m$, r becomes a timelike coordinate and t becomes a spacelike coordinate. This is seen from the changes in sign of the metric components g_{tt} and g_{rr}. Thus, while r is seen as spacelike for *outside* observers, i.e. those for $r > 2m$, giving

a proper measure of circumference of value $2\pi r$ (noting that the angular part of the Schwarzschild metric is the same as that of flat-space geometry), for the interior the r coordinate is timelike. Thus, if time must continue to run, the matter having reached $r = 0$ must continue beyond $r = 0$ to negative r.

Some have proposed that the matter under complete gravitational collapse emerges as a "white hole" in another universe (the negative r referred to above). The white hole is the time reverse of the black hole. Instead of taking positive signs before the two terms on the right-hand-side of the first of (5.11), one can use negative signs. In that case, instead of having a picture of a particle inevitably drawn towards $r = 0$ once having crossed $r = 2m$, the picture is the time-reverse: a particle emerges from the interior to $r = 2m$ and beyond, i.e. the film run backwards.

Our own view is that while these studies are of interest for the exploration of general relativity taken to extreme limits, it is possible that nature resists these catastrophes. While it is widely believed that the existence of black holes in the centers of galaxies has been thoroughly substantiated, one might question whether the concentrated material had really reached the point where it had to enter complete gravitational collapse. It is well to ask how well we really understand the nature of matter at extreme levels of compactification. In the very early years, one would hardly have imagined the existence of a neutron star, a nucleus-like structure that is kilometers in size. What possibilities might exist with quarks and even beyond, physics of matter that we can hardly imagine at present? Might nature always intervene to prevent the formation of black holes and singularities? We know that supernovae are a regular occurrence of nature, blowing off immense quantities of matter during extreme conditions. Is this indicative of nature resisting event horizon and singularity formation? These are possibilities that are seldom discussed publicly.

While it is widely indicated by researchers that certain bizarre phenomena are truly inevitable *in nature*, i.e. beyond their inevitability purely within the confines of the theory under investigation, our view is that these researchers are being overly optimistic about the powers of the theories. The history of physics has been one of theories reaching limits of viability and new theories overtaking

these to improve our understanding of nature. However, it is well to recall the views of Einstein in resisting the extension of theorizing to concepts that are out of tune with our experience, going beyond the experimentation that is reasonably within our grasp. The key word is "reasonably". Of course each person will have his or her idea of what is reasonable.

At this point, general relativity is performing admirably in explaining the observable phenomena that are clearly indicated. This will be discussed further in later chapters to follow.

5.4 The tests of general relativity

The special theory of relativity is very well-tested. On a daily basis, special relativistic dynamics is applied to particle motions in the world's particle accelerators with great success. However, it must be said that general relativity enjoys a far more limited range of verification. We will confine ourselves to a general description of these verifications. It is customary to begin with the three "classical" tests of general relativity and the issues that arise with them. These are: a) The gravitational red-shift; b) The deflection of starlight by the Sun during solar eclipse; c) The precession of the perihelion of the planet Mercury.

- **Gravitational red-shift**

  Since clock rates differ at different positions in gravitational fields, the frequencies attributed to photons (and hence their energies) must differ in these different positions. In particular, photons from the relatively intense gravitational field of the Sun, should be seen to have lower energies, lower frequencies (hence, in terms of waves, longer wavelengths, i.e. shifted to the red end of the electromagnetic spectrum) as compared to what they would have if they had not come from the Sun's field. The comparison is possible and the identifications are achieved from the known spectral pattern from given elements as observed on the Earth.

  However, there are various issues to consider. In terms of the accuracy that has been achieved in measuring red-shift, there

is another derivation that achieves the same result without recourse to the field equations of general relativity. This is described in [11]. The photon is viewed as an element of mass-equivalent (not rest mass) through its energy divided by c^2 according to $E = mc^2$. Newtonian gravity is applied to note its change in potential energy for that amount of mass-equivalent when located at different levels of gravitational potential. This non-relativistic calculation gives a result that matches the lowest order result using general relativity The observations are only sensitive to this level and hence this cannot be viewed with present accuracy as a test of general relativity.

A further complication arises from the fact that spectral lines are broadened by the turbulent motions of the emitting atoms in the Sun, thus reducing the accuracy of the test. This effect has been minimized in a celebrated terrestrial experiment by Pound and Rebka [11]. They exploited the Mossbauer linewidth suppression effect to detect the frequency shift in gamma rays from radioactive iron to high accuracy at different levels of gravitational potential. While very successful to a level of 1% accuracy, it was nevertheless a lowest order test and hence could not be seen to verify the field equations of general relativity. Vessot and Levine achieved an accuracy of 0.02% in an experiment known as Gravity Probe A.

- **Deflection of starlight**

  If we were to view photons as little balls with effective mass given by their energies $h\nu$ (h, the Planck constant and ν, the frequency) divided by c^2 according to Einstein's $E = mc^2$, then by Newtonian gravity, these balls would be attracted by other bodies such as the Sun. This attraction becomes significant when the distance of separation is small. Thus, photons emanating from a distant star and grazing the Sun on their way to us on Earth, will be deflected from their straight paths. Such an admixture of Newtonian physics and special relativistic ideas can only be suggestive of a physical realization. The correct approach is one that follows general relativity.

  Using the null geodesics for the equivalent general relativistic calculation of their trajectories, the result is a different degree

of deflection from that based upon the Newtonian theory plus special relativity combination. Of course under normal circumstances, these photons cannot be detected since they are lost in the immense sea of light coming directly from the Sun. However, during a total solar eclipse, the solar disk is blackened for viewers on the Earth and the faint light from the distant stars skirting the disk become visible.

In the early years, A. S. Eddington was the most noteworthy advocate of Einstein's general relativity in the English-speaking countries of the world and in 1919, he undertook an expedition to Africa to observe the deflected starlight during the solar eclipse of that year. Headlines appeared in the world's newspapers highlighting the subsequent success of the mission and the triumph of Einstein's theory. However, in later years, there was some controversy regarding the accuracy of the measurements taken by the astronomers on the expedition.

An issue of importance concerns the appreciation of what constitutes a *bona fide* test of general relativity.

A true verification must invoke a solution to the field equations, not merely a test of the Equivalence Principle. Indeed Synge [17] has argued with great vehemence that the Equivalence Principle should be seen as nothing more than a kind of signpost indicating the way to the real theory of gravity, general relativity, with its field equations connecting real gravity to geometry. Real gravity, Synge argued, has as its essence, spacetime curvature and the Equivalence Principle is devoid of any connection to spacetime curvature. Nevertheless, it must be noted that the Equivalence Principle has served as a very useful guide toward the development of general relativity. It has been reported that modern tests have been successful, going beyond the Equivalence Principle, with an accuracy of 0.04% using very long baseline interferometry (VLBI) (synchronized radio telescopes) in the observation of the deflection from distant radio sources

- **The precession of the perihelion of Mercury**

  According to Newtonian gravity, the planets move in closed elliptical paths with the Sun at a focus. This is based upon the

treatment of the planets as mass points of sufficiently small size as to be ignorable in terms of interfering with each other's paths. However, Newton's theory provides for attraction between any two masses and therefore causing additional effects on their motions when their masses are taken into account. Astronomers had calculated these perturbative effects from the other planets on the motion of the most inner planet Mercury and found that the orbit should not be that of a closed ellipse. Rather, the perihelion shift for its orbit was calculated to be on the order of 500 seconds of arc per century. The precisely observed perihelion precession of Mercury's orbit fell short by 43 seconds of arc per century. Such was the confidence of the astronomers of the time in Newton's theory of gravity that they hypothesized the existence of a yet-unseen planet between the Sun and Mercury that would provide the residual precession. They even gave it a name, Vulcan, to match the high velocity that it would have, being so close to the Sun. However, diligent searches revealed no Vulcan.

The power of general relativity is manifest in the solving of this problem. General relativity does not lead to closed elliptic orbits of mass points in the field of a dominant mass as does Newton's gravity. Rather, the solution of the geodesic equations leads to precessing elliptical orbits, even before perturbative effects from other planets are taken into account. Moreover, for the specifics of Mercury, the general relativity contribution to that precession is 43 seconds of arc per century, a truly remarkable achievement for the theory.

We have outlined the three classical tests of the theory of general relativity. In terms of solid support, we see that it is only the precession of the perihelion of Mercury that stands out. In modern times, a more dramatic illustration of the precession effect has been evidenced in the orbits of the binary pulsar, PSR 1913+16. This system consisting of a neutron star plus companion exhibits precession on the order of degrees per year. Since then, at least eight other such systems have been discovered, all showing similar changes in periods. In more recent times, a double pulsar has been discovered, two pulsars in a tight orbit about each other, providing valuable timing data from each constituent. This allows for mutual cross-checking of

several general relativity effects at a claimed level of 0.05%.

Another test, the "fourth test", formulated and carried out by I. I. Shapiro [18], consists of sending pulsed radar signals from Earth to Venus during superior conjunction with the Earth and Venus on opposite sides of the Sun, and timing the returning echoes. Very accurate agreement between theory and observation has been reported and even better accuracy has been achieved by bouncing the microwaves off artificial satellites. The Shapiro test is regarded as the one of greatest accuracy. It has achieved the strongest constraint on the departures from the predictions of general relativity in the solar system at the level of 0.002% from ranging to the Cassini spacecraft near Saturn.

A test in progress consists of monitoring the spin axes of four gyroscopes that have been placed in low orbit around the Earth. This "Gravity Probe B" (GPB) experiment[h] [20] tracks the spin-axis directions of these four gyroscopes relative to a distant guide star in an effort to detect two predictions of general relativity: geodetic precession and "frame-dragging" (which is often referred to as the Lense–Thirring effect), the effect of a rotating mass to drag the local inertial frames around in the direction of the rotation.

The geodetic effect is clearly visible in the preliminary data at the 1% level, even before analysis. However unexpected systematics have cropped up and interfered with a clear detection of the latter (frame-dragging) effect, which is 100 times smaller than the geodetic precession. The GPB data analysis team is working hard on modelling those systematics and hope to clarify the frame-dragging situation at better than the 10% level in the near term (with geodetic precession at better than 0.1%).

[h]The theoretical basis for the axis shift was derived by A. Papapetrou [19].

Chapter 6

Gravitational Waves

6.1 Introduction

Just as there are water waves, sound waves and electromagnetic waves, it is natural to consider the existence of gravitational waves. As the oscillation of a charge generates electromagnetic waves, predicted and observed in accordance with Maxwell theory, the Einstein equations predict that an oscillating mass will produce gravitational waves. While various researchers through the years have questioned the existence of gravitational waves, in our view it would be difficult to reconcile their non-existence with the basic tenet of relativity, that information can propagate at a maximum speed c. Waving one's arms produces a change in the mass density distribution of the universe and one would expect the effect of that change to be propagated through the universe as an outgoing wave of information, a gravity wave. We have used the word "information" rather than "energy" to describe what is flowing outwards for reasons that will be developed in what follows.

6.2 Linearized field equations

We outline the steps that led Einstein to his equations for gravitational waves. He considered a perturbation of his field equations to describe weak gravity by having the metric tensor g_{ik} be nearly the Minkowski η_{ik} metric, $\eta_{ik} = \text{diagonal}(1, -1, -1, -1)$,

$$g_{ik} = \eta_{ik} + h_{ik} \tag{6.1}$$

where the components of the h_{ik} perturbation are all much smaller than 1 in magnitude. In such a weak gravity case, the Einstein equations are "linearized", i.e. higher order terms that are products of the perturbation with each other are discarded. After substituting this form of g_{ik} into the field equations, he found that the resulting equations linearized in h_{ik} could be expressed in the suggestive form

$$\left(\nabla^2 - 1/c^2 \frac{\partial^2}{\partial t^2}\right) \psi_i^k = \frac{16\pi G}{c^4} T_i^k \tag{6.2}$$

where the ψ_i^k are defined by[a]

$$\psi_k^i = h_k^i - \frac{1}{2} h_j^j \delta_k^i. \tag{6.3}$$

The equations (6.2) emerge in this particularly convenient form when the conditions

$$\psi_i{}^k{}_{,k} = 0 \tag{6.4}$$

are imposed. Since spacetime is four-dimensional, four infinitesimal coordinate transformations can be made which maintain the perturbative form (6.1). This "gauge freedom" is what permits the imposition of these "harmonic coordinate conditions" (6.4).

These equations are suggestive in that they parallel the electromagnetic field equations

$$\left(\nabla^2 - 1/c^2 \frac{\partial^2}{\partial t^2}\right) A^i = \frac{-4\pi}{c} J^i \tag{6.5}$$

for the four-potential A^i in the Lorentz gauge

$$A^i, i = 0. \tag{6.6}$$

These are the inhomogeneous wave equations with the four-current source J^i. Electromagnetic waves in the case of the latter suggest gravitational waves in the case of the former.[b] In electromagnetism,

[a] The symbol δ_k^i is the Kronecker delta as defined in Chapter 4., taking on the value 0 if the indices i and k are unequal and of value 1 if i and k are the same.

[b] See [21] for a review of the early history of gravitational wave research.

the lowest radiating mode is that of electric dipole radiation. However, because of linear momentum conservation, the lowest mode for the emission of gravitational waves is that due to time variation of the mass quadrupole moments of a source [3], [10], [22]. Partly because of the higher multipole mode and partly because of the weakness of the gravitational coupling constant G, the direct detection of gravitational waves remains to be achieved whereas the direct detection of electromagnetic waves is the daily visual experience of the billions of people on Earth.

6.3 The energy issue and the pseudotensor

We have carefully avoided the use of the word "radiation" in connection with gravity waves. This is because the word "radiation" connotes an energy flow and we have assembled reasons to question whether these waves, whose existence we do not doubt, actually convey energy [23]. To see this, we return to the equations describing conservation, (4.25–4.28). In electrodynamics and fluid dynamics, we expand the $i = 0$ component of (4.27) and integrate over a volume V enclosed by a surface S. Using the Gauss divergence theorem, this results in the global conservation equation

$$\frac{d}{dt}\int_V T^{00}dV = -\int_S T^{0\alpha}dS_\alpha. \tag{6.7}$$

The left side of (6.7) represents the rate of change of energy within the volume V and the right side gives the rate at which energy is flowing out of the bounding surface S.[c] Energy flowing out is matched by energy lost within, the essence of conservation. It is interesting to contemplate that this form of global conservation of energy which builds upon the covariant energy-momentum tensor, appears throughout physics *except* for the case of gravity.

To express global energy conservation, the traditional approach is to retain the quadratic terms in the field equations. Then, by a variety of methods that have been chosen by various researchers through the years, it is possible to express a vanishing ordinary divergence in

[c] dS_α is an infinitesimal three-vector with magnitude equal to that of dS and multiplied by a unit outward normal to the surface element.

place of (4.27) in the form[d]

$$\left[(-g)\left(T^{ik}+t^{ik}\right)\right]_{,k}=0 \tag{6.8}$$

where $(-g)$ is the negative of the determinant of the metric tensor and t^{ik} is the "energy-momentum pseudotensor", a complicated product of derivatives of the metric tensor. As the name implies, the t^{ik} is not a true tensor and in fact can be made to vanish at any pre-assigned point by the right choice of coordinate system, unlike the case for a true tensor such as T^{ik}. This lack of covariance in the case of gravity is a fundamental issue. While one can follow through the steps with (6.8) as was done above with (4.27) to get[e]

$$\frac{d}{dt}\int_V(-g)\left[T^{00}+t^{00}\right]dV=-\int_S(-g)t^{0\alpha}dS_\alpha, \tag{6.9}$$

with the suggested flux of gravitational field energy flow on the right hand side with the aid of a pseudo-energy Poynting vector $(-g)t^{0\alpha}$, the "pseudo" aspect cannot be ignored. Covariant structures are to be sought and when none is available, alternative ideas must be considered.

While some authors (e.g. [24], [25]) have developed a rationalization of the procedure using pseudotensors, our own preference has been to consider what might seem to be a radical departure from traditional thinking on the subject.

6.4 The energy localization hypothesis

Partly inspired by [26], we have brought forth an energy localization hypothesis [23] that resolves the non-covariant issue of the pseudotensor. *The hypothesis is that energy, including the contribution from gravity, resides in the regions where the energy-momentum tensor is non-zero.*[f]

[d]Different researchers have found a variety of forms for the pseudotensor, usually multiplying $\sqrt{-g}$ rather than $(-g)$.

[e]Note the absence of a $T^{0\alpha}$ term on the right hand side of (6.9). This is because the assumption is made that there is no flux of matter or non-gravitational fields across the very distant surface surrounding the distribution.

[f]While this might appear to be a somewhat radical proposal, to add some perspective, a very profound thinker in the person of Synge suggested to us that

Since the energy-momentum tensor is a covariant object, the hypothesis has a covariant basis and the issue of the non-covariance of the pseudotensor is removed. The implication of the localization hypothesis is striking: since the energy-momentum tensor vanishes in vacuum, gravity waves cannot convey energy through the vacuum.

Traditionally, we have come to view the meaning of the word "wave" as a disturbance that carries energy. If the localization hypothesis should prove to be correct, the word would have to carry a different meaning in the case of gravity. Gravity waves would be propagating spacetime curvature disturbances that do *not* convey energy.

Various researchers have dismissed out of hand the very idea of a gravity wave that does not carry energy. However, one such wave has already been well-documented. It is that of a propagating field of gravity, i.e. a wave, emanating from an asymmetrical collapse of dust in the Szekeres exact solutions of the Einstein equations. All the indications are that this wave does not convey energy (see [28]). Rather than grope for rationalizations for this supposed anomaly as others have done, we accept it as a pathway to a generalization in the form of our localization hypothesis.

The points to consider regarding the localization hypothesis are:

- **The nature of gravitational plane waves**

  The simplest electromagnetic wave that exists is a plane wave. It has a well-defined energy Poynting vector and energy density that are covariantly expressed in terms of the components of its energy-momentum tensor. The experimental confirmation of the energy that it conveys is realized routinely. However, the waves that should be the analogous gravitational plane waves belong to the class of spacetimes called "Kerr-Schild". These waves, when cast into Kerr-Schild form, have the interesting property that all components of their associated energy-momentum pseudotensor vanish *globally* [29]. We recall that the pseudotensor can always be eliminated locally by the correct choice of coordinate system. This aspect alone places in

consideration should be given to the possibility that the very concept of energy simply does not belong within general relativity. Very recently, an interesting paper by M. J. Dupre [27] has brought forth the Ricci tensor as the key element of energy localization, in support of our localization hypothesis.

doubt the reality of their energy content. However, the ability to eliminate the energy content at all points simultaneously with a single coordinate system is an even more severe constraint.

Thus, it would be difficult to rationalize any sense of energy content and conveyance when it is possible to totally erase any hint of energy to such waves merely by the judicious choice of coordinate system. It is to be emphasized that such is not the case for truly physical energy-carrying waves such as electromagnetic waves which are tensorially rather than pseudotensorially based. These waves can never be erased by coordinate transformations, not even locally.

- **The essential nature of gravity**

  As we stressed in previous chapters, all particles and fields exist *within* spacetime but gravity *is* spacetime, that is, its curvature. To use an analogy, the particles and fields of nature (apart from gravity) are like the actors on the stage whereas gravity is like the stage itself. The difference is striking. Once the difference is acknowledged, the perceived need to have gravity waves fall into line with other waves is removed.

- **The issue of quantization**

  It is generally believed that like other fields, the gravitational field must have a quantum aspect. The quantum of gravity, called the "graviton", is believed to have spin 2, i.e. an angular momentum of $2\hbar$ where $\hbar$ is Planck's constant h divided by 2π, and an energy $2h\nu$ where ν is the frequency of the underlying wave. However, there is no observational support for the existence of the graviton. At the turn of the 20th Century, there was a basis for believing that classical electromagnetism was inadequate. There was the unexplained photoelectric effect, the mysterious abrupt emission of electrons from a metallic surface when light of a precise frequency was shone upon it. There was the "ultraviolet catastrophe", the inexplicable drop in the high frequency emission intensity for blackbody radiation rather than the classically predicted steadily rising intensity with increased frequency. The quantization of the electromagnetic field with the introduction of the photon concept was

the answer to the phenomena exhibited in nature. However, without any experimental indicator for analogously anomalous behavior, one is left to question the need for the quantization of the gravitational field. If the gravitational field, i.e. space-time itself, is not quantized, then the quantum barrier to an energy-free vacuum gravitational field is removed.

- **The Feynman thought experiment**

  Often cited as a proof of the existence of an energy transport by gravitational waves through the vacuum is the following thought experiment originated by Feynman (see [23] for further discussion and references):

  A stick, on which looped rings are free to slide with friction, is oriented perpendicular to the direction of an oncoming gravity wave. The wave forces the the rings to slide on the stick, generating heat. Energy is said to be transferred from the wave to the stick. However, what was neglected by Feynman was the effect of the wave on the stick.

  We analyzed this aspect by considering an idealized elastic stress configuration suggested by P. J. Westervelt. An element of the elastic medium is simulated as follows: Two parallely aligned perfectly reflecting capacitor plates holding equal and opposite charges, are separated in a line. To counteract the Coulombic attraction, a high frequency electromagnetic wave is bounced between the plates, exerting radiation pressure. The idea behind this structure is that it constitutes a system with elastic properties yet has elements that can be readily analyzed in terms of its reaction to a gravitational wave. This is seen as follows: with the wave of properly chosen intensity, the resulting electromagnetic radiation pressure balances the Coulombic attraction and the system remains in equilibrium. It is shown that forcing the plates closer together increases the radiation pressure while having a lesser effect on the Coulombic attraction [30]. The plates are made to move back, overshooting their equilibrium position. At the increased separation, the radiation pressure is too low to sustain the Coulombic attraction and the plates move back together, again overshooting the equilibrium separation but now moving inwards. The process

repeats, constituting an oscillator. The configuration simulates the elastic property of an element of the Feynman bar. When the gravitational wave impinges on the system, it moves the plates harmonically. However, it also interferes with the high frequency electromagnetic wave, periodically increasing and decreasing the radiation pressure. The net effect is to make the system behave as an oscillator as described above. The plates are seen to move in synchronism with the periodically expanding and contracting simulated stress medium. We can think of the plates as playing a role equivalent to the rings in contact with the Feynman bar. The bar stretches and shrinks in harmony with the motion of the rings so there is no equivalent to the rubbing in the Feynman configuration. The missing element in the Feynman thought experiment is the effect of the wave on the bar itself. As a stress medium, it is not immune to the action of the gravity wave. Gravity acts universally.

- **Other considerations**

  In an interesting paper, Bondi and his collaborators [31] argued in favour of a mass loss for axially symmetric systems emitting gravitational waves. They connected the energy flux to a "news function" linked to the mass monopole moment of the source. However, Madore [32] (and we [33] later, independently, for a particular case) showed the equivalence of the news function to the energy-momentum pseudotensor. Thus the news function shares in its limitations. As well, the effort in [31] does not achieve what would constitute the truly convincing proof of a mass loss: the transition from initial to final stationarity with the final state exhibiting less energy than the initial state.

  It is often stated that the very slow period variation of the binary pulsar PSR1913+16 and the other known pulsars is direct evidence for energy loss by the emission of gravitational waves. This would be an attractive and natural deduction in analogy with electromagnetic radiation emission from accelerated charges in orbit were it not for the unique aspects concerning energy and its localization in general relativity. In this regard, it is well to note that general relativity, being a non-linear theory, would not be expected to yield purely periodic solutions.

Indeed, this was proved to be the case by A. Papapetrou [34]. In terms of energetics, we could understand the period diminution of the binary pulsars in terms of energy conservation without a radiative flow of energy. This would stem from an increase in kinetic energy accompanied by a re-distribution of the gravitational contribution to the internal energy of the composite system. In Newtonian physics, we are familiar with the interchange of kinetic and potential energies of the planets in their elliptical orbits with energy conservation and no radiative energy flow. In Newtonian theory, the question as to where that potential energy actually resides is not well-addressed. Here, we propose that the gravitational contribution has a natural home within the sources themselves and the twist that the general relativistic binary system lacks periodicity is attributed to the non-linearity of the theory.

While we have given reasons to question the generally-believed energy content of gravitational waves while accepting their existence, the question arises as to how these waves could actually reveal their existence. Experimental techniques hinge upon two basic mechanisms, Weber bars and laser interferometers. With the former mechanisms, gravity waves are supposed to deposit their energy to the bars. This energy is detected in the strain of piezoelectric crystals mounted on the bar. While Weber had claimed that he had detected gravity waves in this manner, other researchers were unable to duplicate his findings. Later studies suggested that the technology had not reached the level of sensitivity required to detect the waves. Research has continued over the years in attempts to improve the sensitivity to a sufficiently high level. If our energy localization hypothesis is correct and if it is the case that the bar detectors require an energy deposit from the waves to be functional, then the bars will never reveal the existence of gravity waves. Bel [35] has also concluded that bar detectors will never be viable, but by a totally different line of reasoning. However, it is unclear whether the energy deposit is an essential aspect. If the bar oscillations occur in a manner that is different from their effect on the strain gauges, then it is conceivable that such a mechanism could detect the waves without an energy deposit. There

are subtleties here that require further investigation.

With laser interferometry, electromagnetic waves are bounced between two sets of plates arranged at right angles. If the waves recombine out of phase, interference fringes are produced. A gravity wave impinges upon the array of reflecting plates, moving the plates and interfering with the waves. A careful analysis [36] reveals that the gravity waves will cause the electromagnetic waves to recombine out of phase, resulting in the production of interference fringes. Energy is not required to create this effect. Rather, it is simply a matter of timing, the issue of a coincidence or lack of coincidence of electromagnetic signal arrival times. Thus, we see laser interferometry as an unambiguous mechanism for the detection of gravity waves. This detection approach is presently a very active area of experimental research in different locations in the world.

Chapter 7

The Normal Scales of Physics and the Planck Scale

7.1 The hierarchy of scales

Scales are an essential part of physics. We examine the workings of nature in very different ways depending on the characteristic size of the elements in question. In the largest scale, cosmology, we speak of the Hubble scale, the extent by which the universe has grown since the Big Bang to the present era. The next largest scale, the scale of clusters of galaxies, is what we will consider in Chapter 10. Below this comes the scale of individual galaxies that we discuss in Chapter 9. Much of astronomy concentrates on the structure and evolution of the key elements within galaxies, the stars themselves of which our Sun is a member. Planetary astronomy comes next. It is interesting to contemplate how different are the problems and issues that arise in the physics of these different scales. (In modern times, essential contributions to these larger scale studies have come from the atomic and nuclear scales.)

In terms of human activity and involvement, the next scale of everyday macroscopic physics brings in the wide array of disciplines such as engineering which shape the everyday existence of the life of our planet. In earlier centuries, most of physics was concentrated at this scale where observations and experiments brought forth the

beauty and simplicity of classical physical laws, the three laws of motion of Newton, his law of universal gravitation and the simple elements of electricity and magnetism. It is well to remind ourselves just how remarkably far the classical laws of physics carry us in describing the everyday phenomena of our experience, from projectiles and gyroscopes to vibrating strings, boiling water and magnetized needles.

Below the macroscopic scale, we have the scale of the binding of atoms that leads to the cohesion of matter in bulk. This is at the scale of atomic physics. The scale of atoms, is commonly referred to as the Angstrom scale, 10^{-10} meters. For many years, the atom was regarded as a distribution of positive charge of this Angstrom size with minute electrons of negative charge $-e$ embedded within it as are raisins in a jelly roll. Rutherford's important scattering experiments revealed that the positive charge of the atom was actually concentrated as a nucleus of order 10^{-14} m in dimension. The atom was then seen (for a simplified visual picture) as it is to this day as a structure that is mostly empty space: the nuclear core comprising almost the entire mass of the atom with the minute electrons of charge $-e$ circling this dominant massive core as the planets circle the dominant Sun of the solar system. Within the nucleus are protons of charge e, the carriers of the positive charge and neutrons of zero charge, each of order 10^{-15} m in extent. The contemporary picture is that these particles, along with the mesons, although once regarded as elementary, are actually composites comprised of quarks, particles with fractional charges of magnitude $e/3$ and $2e/3$.

Many particle physicists view the electron as a point, a particle with zero size. The upper limit to the size of the electron as determined by high energy scattering experiments is of the order of 10^{-18} m. Our view, shared by others, is that a point is an element of the idealized world of mathematics rather than the real world of physics. Of great current interest is the theory that the elementary particles are actually all comprised of structures called "strings", minute vibrating segments whose diameters are of the order 10^{-35} m, the "Planck length". String theory was put forth in part by the perceived need to bring gravity into the quantum domain. Still more recent theories extend the string concept to higher dimensions in structures referred to as "branes". To this point, strings and

branes are at the purely theoretical stage, having no experimental support. It is interesting to note that following ideas initiated in his collaboration with Einstein [12], Rosen [37] launched a program with elements somewhat similar to the brane concept, building models of elementary particles as compactified bundles of fields which we would now designate as "solitons". The Rosen program was developed further with multiple scalar fields [38] and in a more sophisticated way in the form of solitons of Dirac–Maxwell theory [39]. Further plans include the investigation of Dirac–Yang–Mills solitons to incorporate the weak interaction. One of the aspects of similarity of Rosen's soliton program and string/brane theory is the notion of particle hierarchies arising from excited states.

7.2 The fundamental interactions of nature

When we mentioned string theory above, we spoke of the diameters of the strings being of the order of the Planck length, 10^{-35} m. This dimension arose because of the desire of the string theory originators to bring gravity into the fold of elementary particles. This in turn is actually a carry-over from Einstein's original vision of unifying gravity with electromagnetism into what is usually referred to as a unified field theory. In more recent times, this goal was expanded to demand the unification of gravity with the now three known interactions in nature, in order of descending strength, the strong interaction that binds the neutrons and protons in the nuclei of atoms, the electromagnetic interaction that couples all particles that carry electric charge (the original Einstein focus) and the weak interaction that is responsible for the weak decays such as when a muon decays into an electron, a neutrino and an anti-neutrino.

Understandably, particle physicists see gravity as just another field and the graviton, the conjectured quantum of the gravitational field, as just another particle, analogous to the photon as the quantum of the electromagnetic field. They divide physics into four forces, of which we have already mentioned the strongest three, the strong, electromagnetic and weak forces. To this, they add the gravitational "force" even though general relativity has taken us away from regarding gravity as a force. Rather, in general relativity, gravity is seen as a manifestation of the curvature of spacetime. It is a prop-

erty of the arena in which the particles of physics reside rather than just another member of the particle family within the arena.

7.3 The Planck scale and the issue of the quantization of gravity

However, it is natural to consider whether there may be a *quantum aspect* of some kind to gravity, regardless of whether a graviton actually exists or not. It is reasonable and interesting to pose the question: is there a dimension at which gravity might naturally mesh with the quantum world? To explore this possibility, physicists have constructed from the primary dimensional constants of nature, namely c, $\hbar$ and G, a combination with the dimension of length. This is the Planck length $l_{\rm p}$, referred to in Section 7.1,

$$l_{\rm p} = \sqrt{\frac{\hbar G}{c^3}}. \tag{7.1}$$

Similarly, combinations of the constants can be found with the dimension of mass, the "Planck mass" $m_{\rm p}$,

$$m_{\rm p} = \sqrt{\frac{c\hbar}{G}}. \tag{7.2}$$

This is approximately 2.2×10^{-8} kg, an extremely massive particle by elementary particle standards. It is the typical mass of a bacterium, which provides a sense of the extreme nature of the Planck scale.

Equally extreme is the value of the Planck length, approximately 1.6×10^{-35} m. It is to be noted that this length is 17 orders of magnitude below the current upper limit to the size of an electron. Probing smaller and smaller distances demands greater and greater energy and this level is likely many orders of magnitude beyond even the most visionary of experimental designers. Some physicists have gone so far as to dismiss any discussion of constructs with this Planckian dimension, such as strings, as being outside of the domain of true physics because it is beyond being subjected to experimental verification. However, our view is that while it may be forever beyond the scope of experimental physics, it is worth devoting some *limited* effort (keeping Einstein's justification of concepts in mind) to the explo-

ration of the theoretical side as far as it may take us. There may lurk hidden suggestions of deeper truths that are otherwise inaccessible.

A point to note is that by simply combining the fundamental constants to derive a Planck length, there is no distinction between the input of Newtonian gravity and that of general relativity, the coupling constant G being the same for each theory of gravity. A more meaningful approach would be to equate the Compton wavelength of a particle of mass m with the "gravitational radius", the r value at which we encounter the event horizon of the Schwarzschild spacetime, i.e.

$$\hbar/mc = 2Gm/c^2. \tag{7.3}$$

This sets the quantum aspect of the particle through its Compton wavelength to match the compactification scale at which the gravitational field becomes very strong where indisputably, gravity at the small scale must be treated with the theory of general relativity. From (7.3), we find the same mass magnitude as before using purely dimensional considerations apart from an insignificant factor of $1/\sqrt{2}$. Thus we see a convergence towards these scales from the two approaches.

7.4 Adding spin and charge to the Planck scale

While we have thus far a measure of mass for the Planck scale particle, there has been no reference to its spin angular momentum or to its charge, which are fundamental quantized aspects of particles in nature. To reflect the quantized union of gravity with matter, surely spin and charge should be incorporated. There is a natural route to doing so [40] in harmony with general relativity. The gravitational field of a charged body with spin is expressed in general relativity by the Kerr–Newman [41] metric. The gravitational radius corresponds to the upper sign in[a]

$$r_\pm = \frac{G}{c^2}\left(m \pm \sqrt{m^2 - \frac{q^2}{G} - \frac{c^2}{G^2}\,a^2}\,\right) \tag{7.4}$$

[a]Here, "a" is the angular momentum per unit mass.

and the second radius with the lower sign is referred to as the "null surface" radius. At this point, we turn to the earliest ideas in the development of quantum mechanics and consider the *ad hoc* approach of N. Bohr in his quantization of the hydrogen atom.

Max Planck (1858–1947) and Niels Bohr (1885–1962) were two of the most influential physicists of all times. Their work spanned various fields such as thermodynamics but they are best known as two of the key founders of quantum mechanics. They were both Nobel Laureates.

Here, we proceed in an analogous manner and firstly impose a quantization of the charge of the Planck particle in units of the known quantum of charge in nature, the charge of magnitude e of the electron and proton, labeling the quantum number N. Secondly, we impose a quantization of the spin angular momentum in units of the fundamental quantum of angular momentum $\hbar$, with quantum number s. Thus we have the total charge and spin angular momentum

$$q = N\,e\,, \qquad a = s\,\frac{\hbar}{m}\,. \tag{7.5}$$

(Note that the m appears again through the spin.) Setting the Kerr–Newman event horizon and null surface (7.4) radii of the particles equal to their Compton wavelengths, and substituting the quantized charge and spin from (7.5), we have

$$\frac{\hbar}{mc} = \frac{G}{c^2}\left(m \pm \sqrt{m^2 - \frac{N^2e^2}{G} - \frac{c^2\hbar^2\,s^2}{G^2m^2}}\,\right)\,. \tag{7.6}$$

Solving for m, we find that this mass, which we now designate as $m_{\rm plex}$ is

$$m_{\rm plex} = m_{\rm pl}\sqrt{\frac{2(1+s^2)}{2-\alpha N^2}} \tag{7.7}$$

for both cases, where $\alpha \equiv e^2/\hbar c \simeq 1/137$ is the fine structure constant and $m_{\rm pl}$ designates the standard Planck mass $\sqrt{c\hbar/2G}$. The symbol $m_{\rm plex}$, that we now refer to as the "extended Planck mass", expresses the larger role that the Planck quantities now play with charge and spin contributing to the mass value.

From the extended Planck mass, the new Planck length and "Planck time",[b] i.e. the complete new Planck scale is found in the usual manner.

From (7.7), we see that the presence of either spin or charge leads to an increase in the value of $m_{\rm plex}$ as compared to the traditional $m_{\rm pl}$. It is also interesting to find that these two fundamental quantities of physics, the (now extended) Planck scale and the fine structure constant, are actually linked. Moreover, the presence of the fine structure constant in (7.7) provides an additional source of interest, given the recent focus upon its apparent slow variation in time according to some authors [42]–[43]. Following their recent claims that the value of the fine structure constant underwent changes during the last half of the history of the universe, we focus on the possibility that α could have had a considerably different value in the still more distant past. If α undergoes significant variations, then $m_{\rm plex}$ does as well. Although rather unorthodox in the low-energy regime, this idea appears quite naturally in the context of renormalization, in which the coupling "constants" are actually running couplings. In the standard model, the early universe expands and cools precipitously in its very first instants when it emerges from the big bang, and the energy scale drops substantially, allowing for significant variations in the values of the "running" couplings, the couplings that vary with the evolution, as opposed to the older notion of couplings fixed for all time.

If the fine structure "constant" changes at all, a change in either c or e could be responsible (see [44] for a debate on the two possibilities). A time-varying α can be accomodated in the context of varying speed of light cosmologies, of which many proposals have appeared recently [45]–[46]. While the reported variation of α over the last 10^{10} years is minute (of the order of 10^{-5} [42]–[43]) and the variation of fundamental constants is restricted by primordial nucleosynthesis, it is quite conceivable that more radical changes could have occurred earlier in the history of the universe. Earlier evidence had pointed to a small increase in α going forward in time but this is now in question. Different limits on the variation on α are seen to be inconsistent. For example, constraints from the natural reactor

[b]The Planck time is the amount of time that light would take to move the distance of a Planck length.

in Gabon are incompatible with those from QSO's (quasi-stellar objects) unless the variation in α differs from era to era in an unnatural manner.

However, for the sake of argument, let us consider a potential variation in α. To fix our ideas, suppose that $N = 5$ and s is of order unity. Then, if at sometime in the past, α assumed a value close to 8×10^{-2} (approximately one order of magnitude larger than its present value), the value of the extended Planck mass $m_{\rm plex}$ would have been many orders of magnitude larger than its present-day value, regardless of the value of the quantum number s (larger values of N lead to large effects for smaller variations of α).

7.5 Quantum limits, spectra, the value of α

A very intriguing result arises when we consider how large a value the quantum number N can attain. Extremal values are generally useful to gain insight. In fact earlier, in Chapter 2 we discussed how the fundamental laws of physics arise from the extrema of the action integral. To determine the extremum here, we note that the critical upper-limit N value in (7.7) is $N = 16$ for the present α value of 1/137.036. If N is 17 or larger, the denominator in the square root of (7.7) becomes negative and the value of $m_{\rm plex}$ would be imaginary. We continue to examine the results that may be extracted by seeking extremal values. With this extremal N value of 16, the extended Planck mass becomes infinite for an α value of 1/128. Is there anything significant about an α value of 1/128? Interestingly, the α value governing high-energy radiation in Z-boson production and decay has been measured to be 1/127.934. The Z-boson is electrically neutral. Could it be that the connection to the *extremal* value reflects this neutrality?

We recall that in the early years of development of atomic physics, there was much interest in the value of α which was believed to be precisely 1/137, i.e. that the denominator was precisely 137. Speculations were entertained as to the possible importance of the integer 137, that it might have some deep significance. As the accuracy of measurements were increased and the α value was determined to be 1/137.036 and not precisely 1/137, the interest in the idea faded. However, now that we see that this extremal value for N gives an α

value that is so close to a measured value of significance for another area of particle physics, it is well to re-visit the notion that there may exist some connection between the fundamental constants of nature and pure integers after all.

We see that the scope for extending the Planck scale is severely limited if we were to be restricted to the choice of the event horizon radius r_+ as opposed to the null surface radius r_-. From (7.6) with the positive sign in front of the square root, we find the inequality

$$\frac{\hbar}{mc} - \frac{Gm}{c^2} \geq 0. \tag{7.8}$$

Therefore, with (7.7)

$$m_{\rm pl} \leq m_{\rm plex} \leq \sqrt{2}\, m_{\rm pl}. \tag{7.9}$$

These conditions in conjunction with (7.7) place the following restrictions on the allowed spin and charge quanta:

$$s^2 + N^2\alpha \leq 1\ , \qquad N^2\alpha < 2, \tag{7.10}$$

(the latter condition already contained in the former) and they lead to a spectrum of spin/charge values. The allowed values of s and N for $\alpha = 1/137$ are

a) for $s = 0$, $N \leq 11$
b) for $s = 1/2$, $N \leq 10$
c) for $s = 1$, $N = 0$.

Spin-two is not allowed in this case which might evoke some surprise as the graviton is seen as a spin-two boson. However the extended Planck mass, as the traditional Planck mass, is very large whereas the graviton mass is zero to a very high level of accuracy ($m_{\rm graviton} < 10^{-59}$ g). These are very different concepts.

Given the new extended approach, it is natural to introduce an extended *Planck charge* and a *Planck spin*. These quantities could be defined by assuming that the "Planck particle" considered is an extremal black hole, i.e. one defined by

$$m^2 = \frac{q^2}{G} + \frac{c^2}{G^2}\, a^2 \tag{7.11}$$

(corresponding to the equality in (7.10)) that is maximally charged [$s = 0$, $q = q_{max}$] or maximally rotating ($q = 0$, $s = s_{max}$). These

requirements yield the extended Planck quantities

$$q_{\rm plex} = \frac{e}{\sqrt{\alpha}} \simeq 11.7\, e, \quad s_{\rm plex} = 1 \tag{7.12}$$

(corresponding to the Planck angular momentum $L_{\rm plex} = \hbar$ and now allowing for non-integral N). While $q_{\rm plex}$ is large but not extraordinarily so, $L_{\rm plex}$ is rather ordinary on the scale of particles familiar at an energy much lower than the Planck scale.

According to the third law of black hole thermodynamics, an extremal black hole corresponds to zero absolute temperature, and is an unattainable state. If the third law survives in the Planck regime, the values of N and s are even further restricted, and the first of (7.10) should read as a strict inequality.

If we consider instead the null surface of radius r_- defined by (7.4) and with (7.7), the inequalities

$$s^2 + N^2\alpha \geq 1, \quad N^2\alpha < 2 \tag{7.13}$$

follow.

In this case, the allowed values of s and N for $\alpha = 1/137$ are:

a) for $s = 0$, $12 \leq N \leq 16$
b) for $s = 1/2$, $11 \leq N \leq 16$
c) for $s = 1$, $0 \leq N \leq 16$
d) for $s = 2$, $0 \leq N \leq 16$

In this case, spin two is readily allowed and with the extremal $N = 16$ value, the α value of $1/128$ gives an infinite $m_{\rm plex}$

Finally, a note is in order regarding the frequently mentioned observation that the natural length scale of "grand unification", the merging of the strong and the electroweak interactions, is only a few orders of magnitude larger than the standard Planck length scale. Thus, the suggestion arises that ultimately, gravitation may hold the key to a final "super-grand unification", the unification of all the interactions. It must be remarked that any spin or α modification in the new Planck scale can only increase the mass scale and hence lower the length scale. This favors the view that gravity is not to be regarded as just another field that must follow in the pattern of quantization that is obeyed by the other particles and fields of physics.

Chapter 8

General Relativistic Cosmology

8.1 Sizes of astronomical elements

In the previous chapter, we focused upon the smallest scales. We now proceed in steps towards the largest scale of all, the scale of the universe itself. In this process, it is useful to build an overall picture of the astronomical dimensions by the use of relative scales and comparisons with dimensions of our sphere of familiarity.

It is useful to begin with the Earth, the sphere with circumference of approximately 40,000 km. We have a picture of a long trip within any of the continents of 4,000 km so we can picture the Earth as a whole being one order of magnitude beyond this for a circumnavigation. The Moon orbits the Earth at a radius of approximately 10 times the size of the Earth's circumference. We can get a good picture of the size of the Sun by noting that the Earth with the Moon in orbit around the Earth would easily fit as a package inside our Sun. While these sizes are within our mental grasp, it becomes more challenging to go beyond this in real distance terms. Instead, we gain perspective by resorting to scaling.

Let us consider the Sun reduced to the size of the head of a pin. At this scale, we can readily picture our solar system with its array of planets by noting that at the outer edge, the now renamed "dwarf planet" Pluto would be a microscopic speck 20 m from the pinhead-sized Sun. This picture shows us clearly that there are

vast spaces between the planets and we have no reason to expect too many collisions.

At the pinhead scale for the Sun, the nearest star to our Sun is 50 km in distance so again, we can appreciate that a collision between two stars would be a rather exceptional event. We complete the picture on this scale by considering the basic building block of the cosmos, the galaxy, with its billions of stars. On this scale, the diameter of the galaxy is approximately 300,000 km in diameter, about 3/4 of the Earth-Moon separation. It is indeed a vast array of pinheads but the visualization is within our grasp, having built up a picture earlier of the distance from the Earth to the Moon. We now compress the scale further. We imagine the galaxy reduced to the size of a dime, i.e. the centimeter scale. Like a dime, a galaxy typically has a radius of approximately ten times its thickness. Having discussed the tremendous spaces between the planets of the solar system and of the spaces between the stars within the galaxy, it is somewhat surprising to note that on the dime scale, the typical neighbor galaxy is located a mere one meter away. Such is the power of current astronomical observations that quasars are observed at about 6 km in distance on the scale of dime-sized galaxies.

Typically, a galaxy rotates with a period of 10^8 years. Galaxies tend to cluster in widely varying numbers, from a very small number and up to several thousand in number. Overall, the galaxies are seen to be fairly isotropically distributed and this fits in well with the nearly isotropic microwave background radiation that is observed at 2.7 K. This is generally interpreted as the cool remnant of the original fireball from the "Big Bang" of the early universe that occurred approximately 14 billion years ago.

8.2 Early ideas about cosmology

The prevailing overall picture of the universe in the early years (and as late as the first three decades of the 20th Century) was one of a static structure. To this day, we remain surprised that even the greatest of thinkers such as Einstein entertained this view. After all, gravity tends to draw unsupported masses together and it would have seemed more reasonable from the outset to believe in a dynamical universe. Moreover, it was shown that even if the universe were

static, it would not be so for long: it was shown that even a small perturbation of the static universe would continue to grow and the perturbations that we see around us are unremitting. We will have more to say about this issue in Section 8.4.

In 1823, H. W. Olbers reasoned that if one were observing the night sky in the event that the universe were static and infinite, even a small amount of radiation intensity from the stars in the universe would cause the night sky to be a blaze of light. The darkness of the night sky had been seen as a paradox. However, later work by E. R. Harrison [47] and P. S. Wesson et al [48] showed that the darkness of the night sky is due almost entirely to the relatively short age of the universe which limits the amount of light that galaxies have produced. If there were no expansion or red-shift at all, the intensity of the background light from distant galaxies would increase by only a factor of approximately two (see [49] for a review of this issue).

The observation of the red-shifts of the light from distant galaxies was a key indicator that the universe was actually in a state of expansion. This was the important contribution of E. Hubble who observed that the red-shifts were proportional to the distances of the galaxies from us.[a]

Since red-shift is proportional to velocity by the Doppler relation, the combination of the Hubble observations gives us the important result that the velocities of recession of the distant galaxies are proportional to the distances they are from us, the "Hubble law"

$$v = HD \tag{8.1}$$

where v is the velocity of recession, D is the distance and H is the "Hubble constant". Actually, H is a time-varying parameter, changing as the universe evolves. When we speak of H as a constant, we refer to its value at the present epoch. One of the key issues in cosmology has been the question of the present value of H as it ties in with the basic structure of the universe. This will emerge in what follows.

Under the assumption that we do not occupy a particularly privileged position in the universe, we adopt what is called the "Cosmological Principle": *At a given time, the universe has the same*

[a]It has also been reported in [50] that seven years before Hubble, K. Wirtz in 1922 and L. Silberstein a year later remarked on the systematic red-shifts and discussed the possibility of a Hubble-like law.

averaged appearance for all observers. Earlier, we noted the high degree of isotropy that we observe on the cosmic scale. Therefore, by the Cosmological Principle, we assume that this will be the case for all observers. It can be shown that this demand is readily expressed in "co-moving"[b] polar coordinates in the metric form

$$ds^2 = dt^2 - R(t)^2 \left[\frac{du^2 + u^2(d\theta^2 + \sin^2\theta d\phi^2)}{\left(1 + \frac{ku^2}{4}\right)^2} \right] \tag{8.2}$$

where k can assume the values 0, 1 or -1.

Clearly, if $R(t)$ is a constant and $k = 0$, we can set $Ru = r$ and we retrieve the standard metric of flat spacetime in polar coordinates. If $R(t)$ varies with the proper time t and k still being 0, the spatial slices are still flat but there is an overall expansion for $dR/dt > 0$ and a contraction for $dR/dt < 0$. (There is still spacetime curvature as the Riemann tensor will be different from 0.) For $k = 0$, we have an open unbounded infinite universe. This is also the case for $k = -1$ however the spatial slices are not flat but rather are said to have negative curvature. For $k = 1$, we have a closed finite universe but still unbounded.

These three cases are best visualized by analogy with two space dimensional surfaces. For $k = 0$, we can imagine an infinite flat sheet of metal with Bunsen burners distributed uniformly under the sheet in all directions. Spots are painted uniformly over the surface. As time advances, the sheet expands homogeneously. For a circle drawn in the sheet, the ratio of its circumference to its diameter is π. The spaces between the spots grow with time and the view is the same from the vantage point of any chosen spot.

For $k = -1$, we consider the central region of the surface of a saddle. If we were to draw a circle there, we would find that the ratio of its circumference to its diameter is greater than π, indicating spatial curvature. If we were to try to flatten this section of the saddle onto a plane, a section would have to be brought together and folded to make a flap. Imagining the central saddle section in all directions

[b] In this coordinate system, the elements of the matter are anchored to the coordinates themselves for all time. Their physically changing separation arises from the time-dependent $R(t)$ factor.

of this analogous two-dimensional space gives a picture of the actual three (spatial) dimensional case.

For $k = 1$, the model to use is the surface of a balloon with spots painted uniformly over its surface. Clearly for a circle drawn on its surface, the ratio of its circumference to radius is less than π. To attempt to flatten the drawn section onto a plane without stretching, we find that there is an inadequate amount of material. It must be slit to allow wedges to open, in contrast to the excess of material in the $k = -1$ case. As we blow up the balloon, the spots separate from each other in a totally uniform manner. Any spot chosen as the center sees exactly the same picture of its neighbors retreating from it radially and uniformly in all directions. Moreover, for this case as well, there are no boundaries: a journey of any length on this surface will never encounter a barrier. Of particular interest is that this space is finite in size, namely $4\pi r^2$ in the value of its proper surface area.

For the $k = 1$ finite universe case, the $R(t)$ factor plays the role of the physical radius of the universe. However, for the infinite $k = -1, 0$ universe cases, the $R(t)$ is no longer the radius, which is infinite for an infinite universe. Rather, the $R(t)$ plays the role of a "scale factor", providing the actual physically expanding distances between galaxies which are anchored to the comoving coordinate grid (see [51], [52], [53] for further discussions).

From the metric (8.2) and from our earlier discussion regarding physical distance, we see that the physical radial distance $D(t)$ from us (that we take to be positioned at the origin of coordinates) to a galaxy at radial position u is

$$D(t) = R(t)u. \tag{8.3}$$

Therefore the velocity v at time t is

$$v = \frac{dD(t)}{dt} = u\frac{dR}{dt} = \frac{\dot{R}}{R}D \tag{8.4}$$

where a dot denotes d/dt and (8.3) has been used. Comparing (8.4) and (8.1), we see that the current Hubble parameter H is

$$H = \frac{\dot{R}(t_0)}{R(t_0)} \tag{8.5}$$

where t_0 is the present time, the present age of the universe.

8.3 Friedmann universes

In 1922, A. Friedmann found the non-static homogeneous isotropic universe solutions of the Einstein field equations. They were subsequently enhanced by Robertson and Walker and are now most often referred to as the FRW (for Friedmann–Robertson–Walker) universes. Some claim that Lemaitre deserves credit as well and with those authors, an L is wedged in after the F. The solutions assume a perfect fluid of galactic clusters with averaged density ρ and pressure P. With the isotropy, the energy-momentum tensor in mixed form has non-vanishing components only along the diagonal in the simple form

$$T_i^k = \text{diagonal}(\rho, -P, -P, -P). \tag{8.6}$$

The Einstein field equations in conjunction with the energy-momentum tensor (8.6) yield

$$8\pi\rho = \frac{3k}{R^2} + \frac{3\dot{R}^2}{R^2} \tag{8.7}$$

and

$$-8\pi P = \frac{k}{R^2} + \frac{\dot{R}^2}{R^2} + 2\frac{\ddot{R}}{R}. \tag{8.8}$$

An equation of state linking the pressure and the density provides the necessary third equation to determine the solution.

While there are numerous works that consider equations of state with significant pressure (see, for example [3], [16]), we will focus on the state of the universe at the present era where the density is negligible in comparison to the pressure. With P set to 0, (8.7) and (8.8) combine to yield

$$2R\ddot{R} + \dot{R}^2 + k = 0. \tag{8.9}$$

The first integral of (8.9) is

$$\dot{R}^2 + k = \frac{C}{R} \tag{8.10}$$

where C is a constant of integration.

For $k = 0$, a shift of the time origin yields the simple form of the solution as

$$R = At^{2/3} \tag{8.11}$$

where A is another constant of integration. For $k = -1$, the solution of (8.10) is most conveniently expressed in the parametric form with parameter τ as

$$t = B(\sinh 2\tau - 2\tau), \quad R = B(\cosh 2\tau - 1) \tag{8.12}$$

where B is a constant of integration. For both of these solutions, R increases monotonically from 0 to ∞. The 0 part of the range is not significant because it is totally unrealistic to use the zero pressure equation of state for the early universe of small R. However the fact that R goes to infinity at the upper end indicates that these cases entail an expansion of the universe without end.

For $k = 1$, the solution is again most conveniently expressed in parametric form as

$$t = E(2\tau - \sin 2\tau), \quad R = E(1 - \cos 2\tau) \tag{8.13}$$

where E is a constant of integration. This solution describes a cycloid[c] with R starting from 0, reaching a maximum and then retracing its steps to 0 value of R (although again, the start and end regions of this solution are not significant with the $P = 0$ equation of state). Thus, the finite universe analogous to the expanding balloon of two spatial dimensions does not expand forever but will reach a maximum size and then collapse. All of the red-shifts of the previous phase will turn to blue-shifts for the observers as the collapse unfolds.

To see which of the cases would prevail, we re-express (8.7) in the form

$$\rho - \frac{3H^2}{8\pi} = \frac{3k}{8\pi R^2} \tag{8.14}$$

using (8.5) in reference to the present era. Thus, if the average density ρ is less than $\frac{3H^2}{8\pi}$ in the present era, the value of k will be -1, the universe is of negative curvature analogous to the earlier discussed saddle surface, and it will expand forever. If ρ has the value $\frac{3H^2}{8\pi}$ at the present value, it is of zero spatial curvature and it will also expand forever. However, if ρ is greater than $\frac{3H^2}{8\pi}$ at the

[c] A cycloid can be drawn by attaching a pen to a point on the circumference of a disc and having the disc roll without slipping on the floor and flush with the wall while the pen traces out the curve on the wall.

present era, the universe is of positive curvature analogous to the surface of a balloon of two spatial dimensions, finite in size, and it will ultimately recollapse. The value of $\frac{3H^2}{8\pi}$ at the present era is of the order of 10^{-26} kilograms per cubic meter or approximately 7 protons per cubic meter. The present density of the visible matter is two orders of magnitude below this, indicating that the value of k is -1 and that at least as judged by the visible matter content, our universe is infinite and will expand forever.

There was also speculation through the years regarding the very early universe. With the evidence for a Big Bang, there were inevitable links to religious creationism. This would fit well with open universes that expand forever. But if the universe were closed and finite, the Einstein equations imply that the universe will reach a maximum size and then contract, reversing its earlier steps of expansion. The question naturally arises: will the universe end in a Big Crunch or will it re-emerge as a new Big Bang, perhaps recycling *ad infinitum*. Friedmann considered such an oscillating universe as did R. C. Tolman. In this vein, M. Israelit and N. Rosen [54] introduced a fundamentally new early universe model in which a singularity was avoided. Upon contraction, the universe reached a minimum size from which it re-emerged into a Big Bang. We [52] refined this model as a field theory.[d]

In very simplified form, this was the picture of the state of relativistic cosmology up to the 1980s.

8.4 The cosmological term

As we discussed earlier, when Einstein turned his attention to describing the universe as a whole, he did so with his preconception that in spite of the dynamical elements in play at the smaller scales, the universe as a whole is static. Unfortunately, his theory of gravity did not cooperate in this regard as it demanded that the universe evolve with time. This is entirely natural as the galaxies are not supported by struts to keep them in place and gravity attracts. In our view, it remains a mystery as to why Einstein retained his static

[d] Very recently, Penrose has revived the oscillating universe concept, suggesting that satellite microwave observations may reveal evidence to support the oscillating universe model.

universe bias although it should be said that it was a widely held view during his time and well before then. Far more reasonable, it would seem, is an evolving universe with time.

Einstein's creative imagination was not lacking. He searched for a means to modify his field equations in such a way as to achieve a static cosmological solution and this was found with the "cosmological constant" Λ. To the Einstein tensor in (4.32), he added a term Λg^{ik} where Λ is a constant,

$$G^{ik} + \Lambda g^{ik} = KT^{ik}. \tag{8.15}$$

This does not violate the covariant conservation laws as the covariant divergence of the metric tensor, $g^{ik}_{;k}$, is identically zero.[e] By choosing the value of Λ properly, it provided for Einstein the required effective tension to offset the attraction of the gravitationally attracting galaxies, thus providing him with his desired static universe model.

E. Hubble's important discovery of the red-shift of the light observed reaching the Mount Wilson telescope from distant galaxies changed Einstein's attitude. The observations indicated that the observed red-shifts were proportional to the distances of the galaxies from the telescope, codified into Hubble's famous law that we discussed earlier

$$V = HD \tag{8.16}$$

where V is the recession velocity, D is the distance to the observed galaxy and H is the Hubble constant. (Recall that H is best termed the Hubble "parameter" as it evolves with time. Its value at the present epoch is what is meant by the Hubble "constant".) Einstein reversed himself at that point, accepting that the universe is actually dynamic. It is often reported that he referred to the introduction of the Λ term as the biggest blunder of his life. There is a widely circulated photo of Einstein peering into the telescope with Hubble looking on, no doubt with great satisfaction. Another factor that helped convince Einstein of the correctness of the dynamic aspect of the universe, as we discussed earlier, was the discovery by Friedmann of exact dynamic homogeneous isotropic solutions of the Einstein equations. These solutions presented the interesting possibilities of

[e]It is an easy exercise to prove this result. An even stronger result is that the metric tensor has an identically vanishing covariant derivative.

infinite universe models of negative or zero spatial curvature as well as finite positively curved models depending upon the mean density of the universe at some given era. The measure of this mean density has been an important issue of modern cosmology, given its scope for determining the most basic nature of the universe.

With these events behind him, Einstein reversed himself regarding the introduction of Λ and forcefully advocated a return to the original form of the field equations, (4.32). However, once released, Λ persisted in occupying the attention of many researchers in spite of Einstein's disclaimers. In recent times, it has grown greatly in popularity as a mechanism to account for the supposed current epoch of acceleration in the expansion of the universe. Supernovae data in recent times are said to provide solid evidence that the universe is in a state of accelerating expansion. While Λ can be a tool to enforce statics, with appropriate magnitude it can also be an instrument to induce such an acceleration, hence the current focus of many researchers on Λ.

What has helped its acceptance was the argument by various researchers that the Λ term belongs quite naturally as a companion to the Einstein tensor as an additional "geometrical" term in the field equations. (Recall in Chapter 4 that the left hand side of the field equations expresses gravity geometrically in the form of spacetime curvature.) Moreover, some have argued, that to follow Einstein in rejecting the Λ term is tantamount to fine-tuning, that the choice of zero for Λ is such an incredibly special choice as to be absurdly improbable. At first glance this seems reasonable as Λ is multiplied by the metric tensor which in turn describes spacetime. However, consider the expression of the Einstein equations in "mixed" form[f]

$$G_i^k + \Lambda\delta_i^k = KT_i^k. \tag{8.17}$$

With g^{ik} now absent in the Λ term, there is nothing "geometrical" to be seen about the Λ term in this form, yet the Einstein tensor term G_i^k continues to describe the geometrical aspect just as well as it did in (8.15) in wholly contravariant form. *We would argue that if it is to be included, the Λ term belongs on the right hand side of the*

[f]Note that the mixed form of the metric tensor g_i^k is equal to δ_i^k. Thus, its elements are simply the constants 0 and 1, regardless of the complexity of g^{ik}.

field equations as an adjunct of the energy-momentum tensor:

$$G_i^k = KT_i^k - \Lambda\delta_i^k \tag{8.18}$$

(or more simply combined with T_i^k into a newly defined generalized energy-momentum tensor ${}_{\text{new}}T_i^k$). While this might appear to be a trivial change, there is reason to feel that this change is of some importance in shaping our attitude towards the significance of Λ. Once it is no longer seen as a part of geometry but rather of matter, we are in more familiar territory. We can evaluate it in relation to the kinds of matter with which we are familiar, pre-dating the introduction of general relativity. From the structure of the Λ term in (8.18), we see that it entails stress in the same magnitude as the density, the kind of matter that, as Bondi used to say, "is not what we can buy in the shops". Currently, the most often expressed view is that the matter constituted by Λ and referred to as "dark energy", makes up approximately 75% of the mass of the universe. This matter is indeed highly exotic. Even an ultra-relativistic fluid has stress only 1/3 the magnitude of the density. We are not saying that we should therefore discount the possibility of the physical existence of the Λ term in nature.[g] Rather, the argument is to bring some perspective into the discussion. The foregoing approach immediately removes the fine-tuning argument: to take Λ to be precisely zero is seen in the new context as the decision that none of this very exotic material exists in nature. While this possibility would be viewed as sheer heresy by many, others have felt it to be entirely plausible. For example, E. W. ("Rocky") Kolb has characterized dark energy as "the ether of the 21st Century". More recently, D. Wiltshire has proposed a different solution to the issue of resolving the evidence for an accelerated expansion of the universe [55] (see also [56] for a comprehensive discussion). It is from the debates, from the clash of ideas, that science progresses, but always with the experiments and the observations to guide the way.

Given that general relativity has been the key theoretical element in describing the universe, one might naturally regard it as rather bizarre that in the scales just below the cosmological scale, namely at the galactic and galactic cluster scales, that general relativity would be discarded and Newtonian gravity used in its place.

[g]Indeed we have added our own part to the study of cosmological models with a variable Λ term [51].

When surprisingly large velocities were discovered at these scales, surprising on the basis of Newtonian gravity, researchers turned to *ad hoc* modifications of Newtonian gravity rather than to general relativity, the preferred theory of gravity. The latter was our course of action and we turn to it in the chapters that follow.

Chapter 9

Motion of the Stars in the Galaxy

9.1 Introduction

The essential building blocks of the universe consist of galaxies, vast conglomerations of billions of stars bound together through their gravitational interactions. They come in various shapes and sizes but our focus will be on the magnificent spiral galaxies and the particularly noteworthy organized circular motions of the stars within them.

What researchers had come to expect is that as they look further from the axis of rotation of such a galaxy, the velocities of the stars would be seen to diminish as $1/\sqrt{r}$ where r is the distance from the axis of rotation. This is based upon our experience with the solar system: for weak gravitational fields such as in the solar field and for non-relativistic motion that is the case for the planets, Newton's theory of gravity is appropriately applied with the force law

$$GmM/r^2 = ma = mv^2/r \tag{9.1}$$

for (nearly) circular planetary motion, where M is the mass of the Sun, m is the mass of a planet, a is the acceleration and v is the velocity. In applying this analysis, we are using the fact that the Sun's gravity dominates and the planets, large as they are with respect to ourselves, are nevertheless treatable with quite reasonable accuracy as test masses relative to the enormous mass of the Sun. They react

to the gravitational field of the Sun but their own gravity is neglected as being negligible in comparison.[a]

Precisely this kind of picture is what the astronomers had in mind when it came to studying the motions of the stars in the galaxy. While they certainly understood that the masses under study are now stars rather than planets, it was felt that with the great bulk of stars within a given radius tugging at an individual star, the picture would be more or less what was described above for the solar system and the velocities of the stars would fall off with distance as $1/\sqrt{r}$. To their surprise, it was found that the velocities of the stars remained fairly constant as they were tracked with increasing distance from the axis of rotation, quantified and illustrated in the so-called "flat galactic rotation curves". To account for this anomaly, it was conjectured that there must be a great deal of matter in the form of vast halos beyond the visible disks of these galaxies to serve as tugs to speed up what would ordinarily be diminishing velocity with distance. This matter, 5 to 10 times the mass of the visible galactic contents, was given the name "dark matter" as it shed no light and apparently displayed its presence only through gravitational interaction. The latter property has led researchers to characterize the dark matter as "exotic". The current widely held view is that dark matter comprises approximately 23% of the mass of the universe as compared to approximately 3% for the known baryonic matter with the remainder, the dominant mass constituent, comprised of the even stranger "dark energy" that we discussed in Chapter 8. A great deal of effort and resources have been dedicated to finding verification for the reality of dark matter at the level of elementary particles.

Even earlier than the discovery of the flat rotation curves of single galaxies, such an anomalous phenomenon was realized at the still-larger scale, that of a cluster of many galaxies. Zwicky had noted that in the Coma cluster of galaxies, the velocities do not fall off from the center with the expected $1\sqrt{r}$ form and had conjectured that there had to exist a large reservoir of dark matter beyond the visible contents to realize the indicated motions. In this, as with the internal motions of stars within a galaxy itself, it was taken for

[a]For more refined studies, the perturbations in the motions due to the masses of the planets gravitationally interacting with each other and affecting the total gravitational field, are taken into account.

granted that Newtonian gravity was adequate to the task for the analysis of these motions.

A variety of researchers through the years were skeptical about the existence of this exotic dark matter and starting with A. Finzi, began to modify the Newtonian force law to match the requirement for the large velocities. M. Milgrom, and later in collaboration with J. D. Bekenstein [57] [58] [59], was very active in this approach as was J. W. Moffat and collaborators [61].

What was overlooked by these and the many other researchers through the years was the possibility that the very best theory of gravity, Einstein's general relativity, could play a role in resolving the dilemma of the large velocities. The common wisdom that prevailed was that when the gravitational fields are weak and the velocities are non-relativistic, it was taken for granted that general relativity was never required to describe the system and that Newtonian gravity theory would suffice. This is certainly an overly simplistic view. As a simple counter-example, consider two masses on a spring in oscillation or a spinning rod, systems studied in the very early years of general relativity. In these, the fields are weak and the motions non-relativistic yet Newtonian gravity predicts that these sources will produce no waves of gravity propagating outwards from the masses with finite speeds, as does general relativity.[b]

Also of interest is the early work of Eddington [62]. He pointed out that when the bodies that are the source of gravity in a problem are in "free-fall" or "gravitationally bound", i.e. moving solely under the influence of their mutual gravitational fields, that even for weak field non-relativistic motion, there is the potential for nonlinear terms from the Einstein equations to compete with the linear terms for significance of effect. Thus one must proceed with caution in rejecting the need to consider possible effects of general relativity.

9.2 General relativistic effects on the stellar motions in galaxies

With then-graduate student, (now Dr.) Tieu, we set out to explore the possible effects of bringing general relativity into the analysis

[b] At the present stage of technological development, this example is one of principle rather than experimental confirmation.

of the rotational velocities of the stars in spiral galaxies [63] [64] [65] [66]. Much earlier, we had thought about the curious situation that while general relativity was deemed necessary to deal with the dynamics of the largest scale in nature, the scale of the universe, it was thought adequate to switch to Newtonian gravity for the second and third largest scales in nature, that of the galactic clusters and the galaxies themselves. We had wished to apply general relativity to the galactic dynamics but never had we imagined it technically feasible until we happened to accidentally come across the paper [67] of W. B. Bonnor. His paper provided the impetus for us to proceed. For mathematical tractability, some simplifications are necessary. It would not be feasible to deal individually with billions of stars so we modeled the galaxy as a pressureless (hence elements in free-fall) fluid in stationary motion with axial symmetry. A "stationary" system is one in which there is no explicit time dependence but which allows for a reversal of motion when the direction of the flow of time is reversed. An example would be a perfectly homogeneous spherical ball rotating with a constant angular velocity about a fixed axis. A reversal of the direction of the flow of time would have the ball rotate in the opposite direction. A "static system" is without explicit time dependence and is also unchanged under a time-reversal. An example would be the non-rotating ball.

The most general structure for such a stationary axially symmetric metric is given by

$$ds^2 = -e^{\nu-w}(udz^2 + dr^2) - r^2e^{-w}d\phi^2 + e^w(cdt - Nd\phi)^2 \qquad (9.2)$$

where u, ν, w and N are functions of cylindrical polar coordinates r, z. However, u can be set to 1 for the weak field situations that we will consider because the field equations show that retaining higher orders for u induces corrections beyond the lowest order solution that we seek.

There is the choice available as to which axially symmetric system of coordinates to use. The two most natural choices available are to have either the coordinate system at rest relative to the distant galactic field of the expanding universe or to have the system of coordinates co-rotating with the fluid ("co-moving coordinates"). We follow van Stockum [68] in choosing the latter. In this case, the four-

velocity U^i is

$$U^i = \delta^i_0 \tag{9.3}$$

where the delta is the Kronecker delta, defined earlier in Chapter 4. This form places the fluid at rest in this coordinate system. We make a purely local transformation with (r, z) held fixed at each point when taking differentials [69] [67]

$$\bar{\phi} = \phi + \omega(r, z)\, t \tag{9.4}$$

in a special way to locally diagonalize the metric, i.e. to locally remove the space-time cross term $dtd\phi$. In so doing, we go from "stationary" to "static" at each point since it is this cross-term that makes the two directions of time flow un-equivalent. This enables us to deduce the local angular velocity ω and the tangential velocity V as (see also [70])

$$\omega = \frac{Nce^w}{r^2 e^{-w} - N^2 e^w} \approx \frac{Nc}{r^2} \tag{9.5}$$

$$V = \omega r \tag{9.6}$$

where the approximate value has been chosen because we are considering the weak fields in the galaxy and we are expanding in powers of the gravitational constant G, here taken as a smallness parameter. To first order, the field equations (4.32) are

$$\begin{aligned} 2r\nu_r + N_r^2 - N_z^2 &= 0, \\ r\nu_z + N_r N_z &= 0, \\ N_r^2 + N_z^2 + 2r^2(\nu_{rr} + \nu_{zz}) &= 0, \\ N_{rr} + N_{zz} - \frac{N_r}{r} &= 0, \end{aligned} \tag{9.7}$$

$$\begin{aligned} &\left(w_{rr} + w_{zz} + \frac{w_r}{r}\right) + \frac{3}{4} r^{-2}(N_r^2 + N_z^2) \\ &+ \frac{N}{r^2}\left(N_{rr} + N_{zz} - \frac{N_r}{r}\right) - \frac{1}{2}(\nu_{rr} + \nu_{zz}) = 8\pi G\rho/c^2 \end{aligned} \tag{9.8}$$

where ρ is the mass density.[c] Combining the equations gives

$$\nabla^2 w + \frac{N_r^2 + N_z^2}{r^2} = \frac{8\pi G\rho}{c^2} \tag{9.9}$$

[c]Here we are simplifying the notation using subscripts r and z without commas to indicate partial differentiation with respect to r and z respectively.

where the first term is the flat-space Laplacian operator in cylindrical polar coordinates

$$\nabla^2 w \equiv w_{rr} + w_{zz} + \frac{w_r}{r} \tag{9.10}$$

and ν would be determined by solving the relatively simple differential equations.

It is easily shown that the choice of co-moving coordinates and the fact that the system is gravitationally bound give $w = 0$ [64] [66]. With these simplifications, the field equations for N and ρ become

$$N_{rr} + N_{zz} - \frac{N_r}{r} = 0 \tag{9.11}$$

$$\frac{N_r^2 + N_z^2}{r^2} = \frac{8\pi G\rho}{c^2}. \tag{9.12}$$

It is noteworthy that if the minus sign were changed to plus in (9.11), N would satisfy the flat-space Laplace equation in cylindrical polar coordinates. Other important points are the following: from both the field equation for ρ (9.12) and the expression for ω (9.6), we see that N is of order $G^{1/2}$. Also, N cannot be eliminated without removing the rotation of the source. That the galactic dynamical problem is a non-linear one is now clearly seen from these equations, with a quadratic term in N determining the mass density.[d] By contrast, in Newtonian gravity, the relation between source (density) and field (ϕ) is a linear one through the Laplace equation.

It is interesting to observe that (9.11) can be expressed as

$$\nabla^2 \Phi = 0 \tag{9.13}$$

where

$$\Phi \equiv \int \frac{N}{r} dr. \tag{9.14}$$

Therefore flat-space harmonic functions Φ are ultimately connected with the axially symmetric stationary pressure-free weak fields that we seek.[e] These harmonic functions are the generating potentials referred to earlier. It is to be noted that these generating potentials play a different role in general relativity than do the potentials

[d] Some of our critics have now acknowledged this to be the case.

[e] The role of harmonic functions is even stronger than this: Winicour [71] has shown that all such sources, even when the fields are strong, are generated by such flat-space harmonic functions.

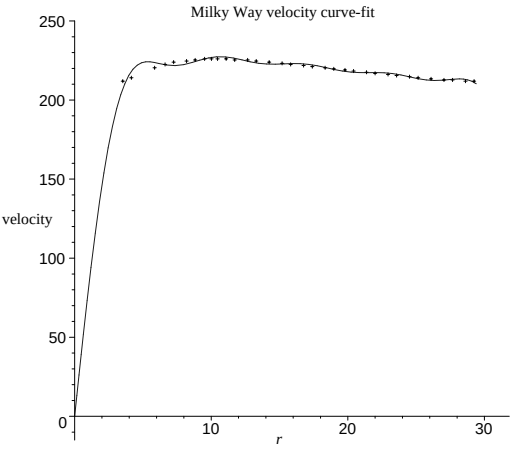

Figure 9.1: Velocity curve-fit for the Milky Way in units of m/s vs Kpc.

of Newtonian gravitational theory even though both functions are harmonic. This is seen as follows:

Using (9.6), (9.5) and (9.14), we have the expression for the tangential velocity of the distribution

$$V = c\,\frac{N}{r} = c\,\frac{\partial \Phi}{\partial r}. \tag{9.15}$$

Thus we see that in our general relativistic case, the gradient of the harmonic generating function is linked to velocity whereas in Newtonian gravity, the gradient of the harmonic potential function is linked to acceleration. The distinction is quite striking.

9.3 Modeling the observed galactic rotation curves

Non-linearity in differential equations is usually challenging and the field equation for ρ is indeed non-linear. However, in the galactic problem, we are fortunate in having N itself satisfy a linear equation (9.11) and N is related to a harmonic function Φ, the generating potential. Therefore, in galactic modeling, we exploit this by finding

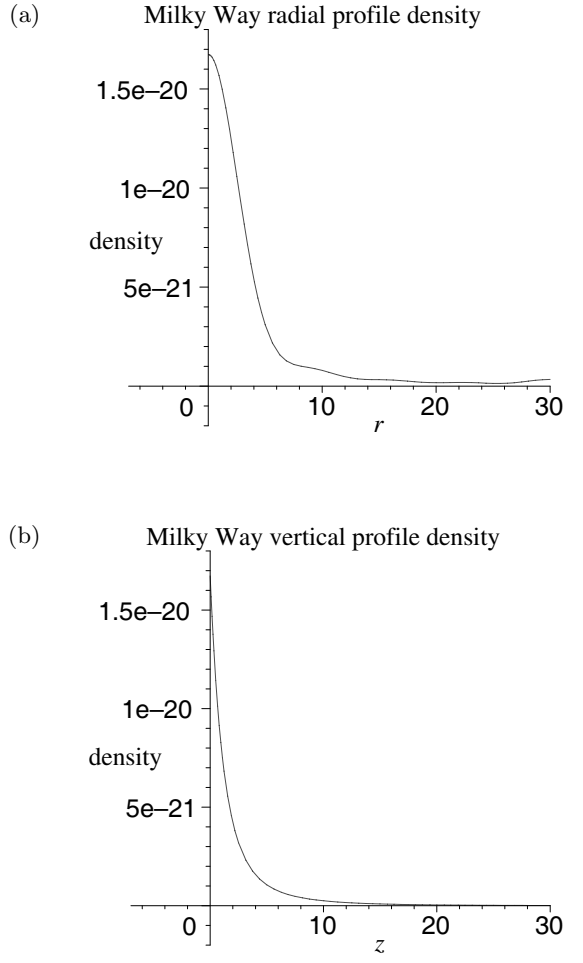

Figure 9.2: Derived density profiles in units of kg/m^3 for the Milky Way at (a) $z = 0$ and (b) $r = 0.001$ Kpc.

first the generating potential whose r gradient matches the observed tangential velocity distribution of the stars in the galaxy under study. This leads to the appropriate N function. With N found, (9.12) gives the density distribution directly. If this leads to a distribution of mass that is largely concentrated in the region of the visible symmetry plane of the galaxy with mass as expected from the visible contents, then general relativity will have resolved the issue of the observed flat rotation curves without exotic dark matter. If, however, these rotation curves cannot be realized without massive spherical halos surrounding the flattened visible disks, then general relativity will have given the same answer as had Newtonian gravity, i.e. that the visible contents of galaxies are but a small fraction of the total galactic mass. In that case, those of our critics who have asserted that Newtonian gravity is adequate for the galactic problem would have been indirectly vindicated, but their method of achieving this conclusion would still have been faulty. This is because they would not have used general relativity which we have seen to be necessary for logical consistency. Which of these alternatives is correct? In what follows, we will consider this issue carefully.

At this point, we consider the mathematical aspects of building the required solution. For a given galaxy, the challenge is to determine the composing elements that make up the generating potential for its rotation curve. Fortunately, with the generating potential satisfying a linear differential equation, we are able to mold a solution to satisfy a given shape through linear superposition. Further simplifications accrue with the choice of separation of variables form of solution. This yields the following base solution:

$$\Phi = Ce^{-k|z|}J_0(kr) \tag{9.16}$$

in cylindrical polar coordinates, where J_0 is the Bessel function $m = 0$ of Bessel $J_m(kr)$ and C is an arbitrary constant.[f] To provide reflection symmetry of the distribution for negative z, this form of solution requires that the absolute value of z be used. As a result, this produces a discontinuity in N_z at $z = 0$. While this has led to some considerable discomfort in some quarters, we point out that in the problem at hand, this discontinuity is consistent with the general case

[f] See for example [72].

of having a density z-gradient discontinuity at the plane of reflection symmetry. We will elaborate on this in what follows.

Using the linearity of (9.13), we are able to write the general solution of this form as a linear superposition

$$\Phi = \sum_n C_n e^{-k_n|z|} J_0(k_n r) \tag{9.17}$$

with the number n of terms in the series chosen appropriately for the level of accuracy that we wish to achieve. This procedure is reminiscent of standard Fourier analysis. From (9.17) and (9.15), the tangential velocity is

$$V = -c \sum_n k_n C_n e^{-k_n|z|} J_1(k_n r) \tag{9.18}$$

using $dJ_0(x)/dx = -J_1(x)$ [73].

In our galactic modeling, we chose the k_n to make the $J_0(k_n r)$ terms orthogonal. The Bessel functions $J_0(kr)$ satisfy the orthogonality relation $\int_0^1 J_0(k_n r) J_0(k_m r) r dr \propto \delta_{mn}$ where k_n are the zeros of J_0 at the r limits of integration. An excellent fit to the rotation curve for the Milky Way was achieved using only 10 functions with parameters C_n, $n \in \{1 \dots 10\}$ [63], [66]. It should be noted that such curve fits are constrained by the demand that they be created from derivatives of harmonic functions. The curve fit is shown in Figure 9.1. The $J_1(x)$ Bessel functions have the correct basic properties for the problem at hand, being 0 at $x = 0$ and falling as $1/\sqrt{x}$ asymptotically. However, this feature alone does not assure a realistic fall-off of matter. We will discuss this aspect in what follows. Also, the present curves drop as r approaches 0. This is in contrast to alternative proposed means of accounting for the rotation curves given by L. Mestel [74] and MOND authors [57], [58], [59] that lead to flat plots even up to $r = 0$. From (9.15) and (9.18), the N function is determined in detail and from (9.12), the density distribution follows. This is shown in Figure 9.2 as a function of r at $z = 0$ as well as a function of z at $r = 0.001$ Kpc. We see that the distribution is an essentially flattened disk with good correlation with the observed overall averaged density data for the Milky Way (see Figure 9.3). Moreover, we see that in the cross-sectional density plot, the equidensity contours are approximately elliptical around the visible galactic region. With

this model, we compute the integrated mass as $21 \times 10^{10} M_{\odot}$. This is at the lower end of the estimated mass range of $20 \times 10^{10} M_{\odot}$ to $60 \times 10^{10} M_{\odot}$ as established by various researchers. It is to be noted that the approximation scheme would break down in the region of the galactic core should the core harbor a black hole or even a naked singularity (see e.g. [16]). *The essential point is this: the matching of the flat velocity curve is achieved in general relativity with confined mass in the disk up to an order of magnitude smaller than the envisaged halo mass of exotic dark matter.*[g] It should be emphasized that in our general relativistic models, we are dealing with a generic continuous mass density ρ. The reality is a very lumpy distribution of luminous stars as well as planets, comets, burnt out stars, neutron stars, dust, etc., the non-luminous but normal baryonic matter that is well-understood in physics. We are aware that in nature there exists both light-emitting and "dark"(i.e. non-light-emitting) matter and both kinds of matter are part of this density ρ of which we speak. The expression "dark matter" by contrast, has come to signify the enormous quantities of matter in the vast extended halos surrounding the galaxies, matter with no known connection to the normal matter that is a part of current physics, "exotic" matter that displays its proposed existence only through its gravitational interaction. From the density distribution function, each term within the series has z-dependence of the form $e^{-k_n|z|}$ which causes the steep density fall-off profile as shown in Figure 9.2 (b). This is consistent with the picture of a standard galactic essentially flattened disk-like shape rather than a halo sphere.

From the rotation curve data that we have available to this point, there is no support for the widely accepted notion of the *necessity* for massive halos of exotic dark matter surrounding visible galactic disks. We see that when our premier theory of gravity, general relativity, is brought into the analysis, the observed flat galactic rotation curves linked to essentially flattened disks can be realized with no evident need for exotic dark matter, given the data that we have at present.

We have also performed curve fits for the galaxies NGC 3031, NGC 3198 and NGC 7331. The data are given in [66]. The remarkably precise velocity curve fits are shown in the figures in [66] and here, we display only that for the galaxy NGC7331, Figure 9.4. The

[g]See e.g. [60] for proposed values of extended halo masses.

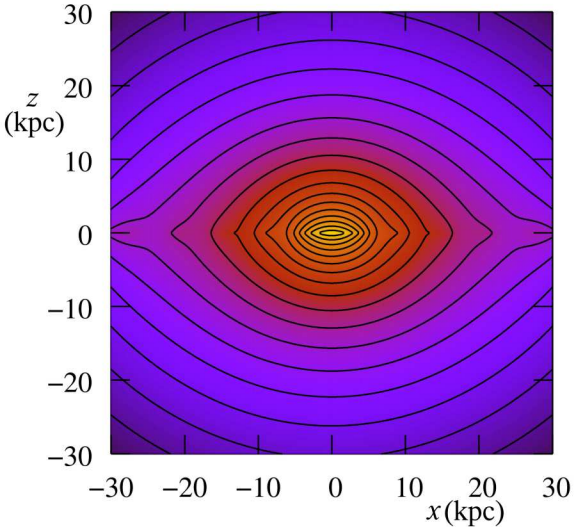

Figure 9.3: Cross-sectional density contour plot for the Milky Way model.

density profile is given for r at $z = 0$. Again the picture is consistent with the observations and the mass is found to be $10.1 \times 10^{10} M_{\odot}$ for NGC 3198. This can be compared to the result from Milgrom's [57], [58], [59] modified Newtonian dynamics of $4.9 \times 10^{10} M_{\odot}$ and the value given through observations (with Newtonian dynamics) by S.M. Kent [75] of $15.1 \times 10^{10} M_{\odot}$. For NGC 7331, we calculate a mass of $26.0 \times 10^{10} M_{\odot}$. Kent [75] finds a value of $43.3 \times 10^{10} M_{\odot}$. For NGC 3031, the mass is calculated to be $10.9 \times 10^{10} M_{\odot}$ as compared to Kent's value of $13.3 \times 10^{10} M_{\odot}$. Our masses are consistently lower than the masses projected by models invoking exotic dark matter halos and our distributions roughly tend to follow the contours of the optical disks.

From the figures provided by Kent [75] for optical intensity curves and our log density profiles for NGC 3031, NGC 3198 and NGC 7331, an interesting result emerges: we find that the threshold density for the onset of visible galactic light as we investigate the data in the radial direction is at $10^{-21.75}$ kg·m^{-3} (Figure 9.5 and Figure 9.6). We have hypothesized that this density is the universal optical luminosity threshold for galaxies as tracked in the radial direction. Should this

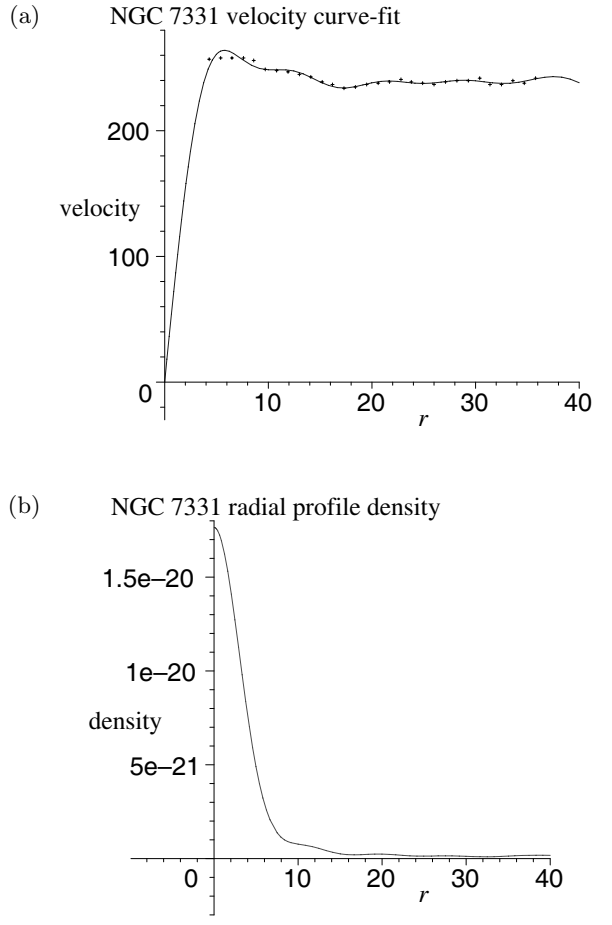

Figure 9.4: Velocity curve-fit and derived density for NGC 7331.

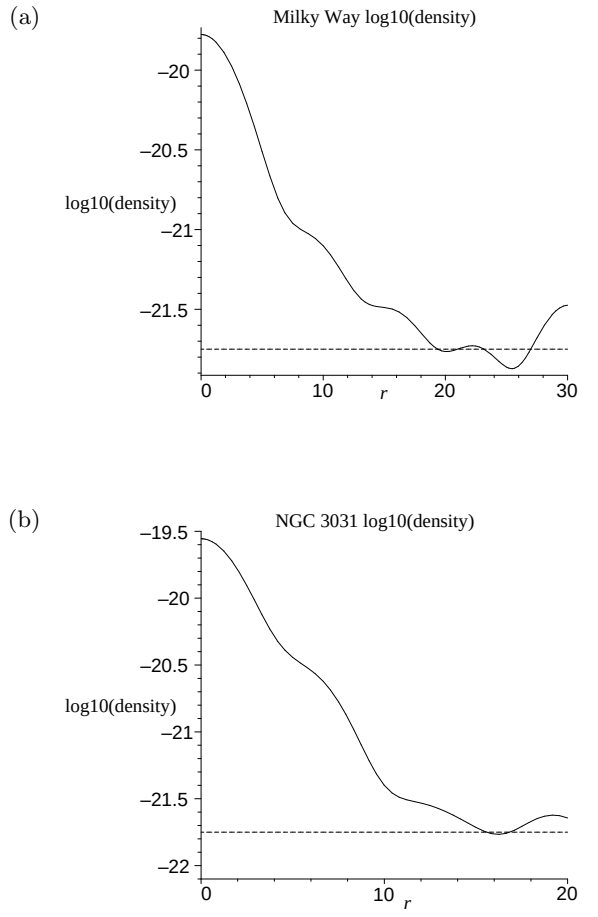

Figure 9.5: Log graphs of density for (a) the Milky Way and (b) NGC 3031 showing the density fall-off. The dashed line at the -21.75 logarithmic density level provides a tool to predict the outer limits of visible matter. The fluctuations at the end are the result of limited curve-fitting terms.

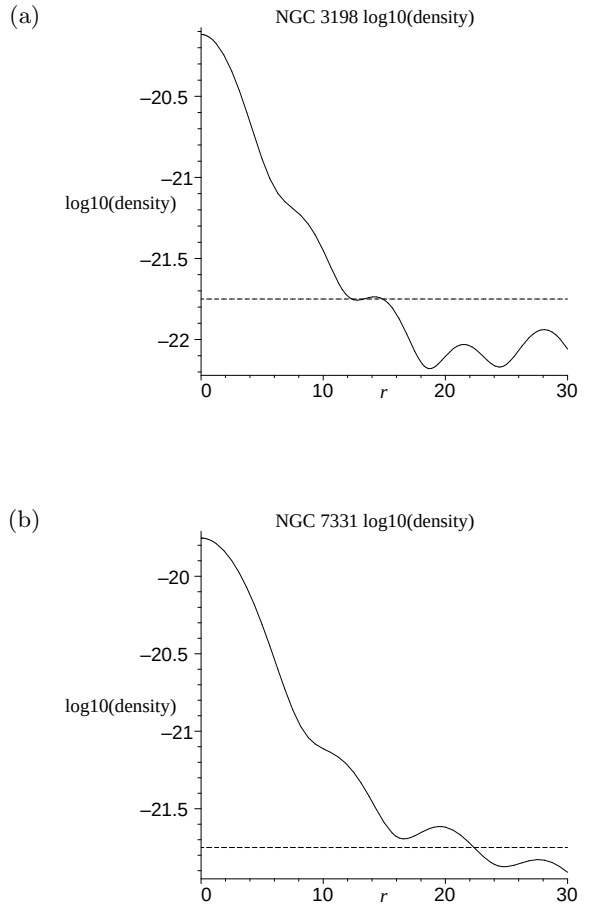

Figure 9.6: Log graphs of density for (a) the NGC 3198 and (b) NGC 7331 showing the density fall-off. The -21.75 dashed line provides a tool to predict the limits of luminous matter. As before, there are fluctuations near the border.

hypothesis be further substantiated, the radius at which the optical luminosity fall-off occurs can be predicted for other sources using this special density parameter. The predicted optical luminosity fall-off for the Milky Way is at a radius of 19–21 Kpc based upon the density threshold indicator that we have determined.

It was interesting to witness the strong response to our first posted paper [63] as well as the continued interest indicated by many researchers as we proceeded with our work. The frequently expressed message to us was that of a shared skepticism regarding the reality of exotic dark matter. With no evidence for exotic dark matter dominating ordinary matter in the universe by a factor of 5–6 other than an apparent extraordinary gravitational tug,[h] more conservative physicists were less inclined to rush to such radical conclusions. However, such sentiments do not generally lend themselves to research papers. Rather, it is usually the critics who express themselves in papers. As a result, many have come to view this essentially one-sided expression in print as evidence for a disbelief in our work. Therefore it has been incumbent upon us to bring forth the range of views and arguments and to counter our critics. It must be stressed that the large volume of criticism is understandable given the deeply ingrained belief that Newtonian gravity should suffice for galactic dynamics. We have welcomed the scrutiny as a spur to greater levels of investigation on our part.

In Appendix A, we have included a review of the challenges to our work and our replies.

9.4 A velocity dispersion test for the presence of extra matter

To this point, our goal was to explore whether the premier theory of gravity, Einstein's general theory of relativity, could account for the observed flat galactic rotation curves without the requirement for vast stores of mysterious dark matter. We showed that this was possible. At the same time, it is important to realize that systems could conceivably exist *with* large halos of matter of whatever form and if so, it would be incumbent upon us to deduce the criterion for

[h] But see later, regarding evidence from other channels.

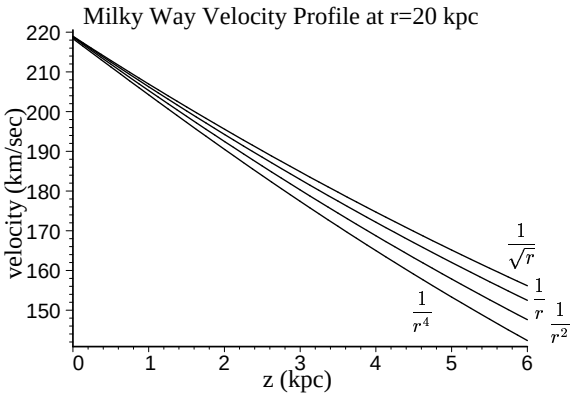

Figure 9.7: Velocity dispersion at $r = 20$ kpc for the Milky Way.

their existence. To achieve this, we have developed a test in principle that relies upon data in the *visible/HI* regime which makes it particularly useful. When we examine Figure A.4, we see that different constructed velocity fall-off profiles beyond the HI region imply different mass accumulations in those external regions. Carrying these back with continuity into the visible/HI region, we find that the extent of the velocity dispersion as we track curves at different non-zero z values depends on the assumed external velocity profile fall-off. (See, for example, Figure 9.7.)

Therefore, it is hoped that the astronomers will focus on gathering data for rotation curves in planes of different z values. With sufficient data, it should be possible, at least in principle, to provide limits on the extent of extra matter that might lie outside of the visible/HI region. To this point, we have only the data provided in [76], [77], [78] but far more data will be required to provide an adequate discriminating test.

9.5 Summary comments on rotation velocities of galaxies

It is natural to question how such a large departure from the Newtonian picture of galactic rotation curves could have arisen using general relativity since the planetary motion problem is also a gravitationally

bound system and the corrections there using general relativity are so small. The reason is that the two problems are very different: in the planetary problem, the source of gravity is the Sun and the planets are treated as test particles in this field (apart from contributing minor perturbations when necessary). They respond to the field of the Sun but they do *not* contribute to the field. By contrast, in the galaxy problem, the source of the field is the combined rotating mass of all of the freely-gravitating elements themselves that compose the galaxy.

We have seen that the non-linearity for the computation of density inherent in the Einstein field equations for a stationary axially-symmetric pressure-free mass distribution, even in the case of weak fields, leads to the correct galactic velocity curves as opposed to the incorrect curves that had been derived on the basis of Newtonian gravitational theory. Indeed the results were consistent with the observations of velocity as a function of radius plotted as a rise followed by an essentially flat extended region and no halo of exotic dark matter with multiples of the normally computed galactic mass was required to achieve them. The density distribution that is revealed thereby is one of an essentially flattened disk without an accompanying overwhelmingly massive vastly extended dark matter halo. With the "dark" matter being associated with the disk which is itself visible, it is natural to regard the non-luminous material as normal baryonic matter.

It is unknown how far the galactic disks extend. More data points beyond those provided thus far by observational astronomers would enable the extension of the velocity curves further. We have made simplifying assumptions for various velocity fall-off scenarios and we have seen that these can readily yield a picture of galactic structure devoid of huge extended very massive halos of exotic dark matter.

Of particular interest is that we have within our grasp a criterion for determining the extent, if of any significance, of extra matter beyond the visible and HI regions of a galaxy. We emphasize that it is possible in principle to determine this with data solely *within* the visible/HI region by plotting the velocity dispersion of rotation curves for various z values. This is an attractive area for future research. In particular, it expands the demands upon not only our galactic model but also upon any other proposed model by other researchers. It

asks for consistency between observation and theoretical prediction for the overall averaged picture of stellar motions within the galaxy.

Nature is merciful in providing one linear equation that enables us by superposition to model disks of variable density distributions. This opens the way to studies of other sources and with further refinements. It is to be emphasized that what we have taken is a first step, a general relativistic as opposed to a Newtonian analysis at the galactic scale. It is noteworthy that others have now come to recognize that the galactic problem is a nonlinear general relativistic problem even given the conditions of weak fields and non-relativistic velocities. It will be of interest to extend this general relativistic approach, with the hitherto neglected consideration of non-linearities, to the other relevant areas of astrophysics with the aim of determining whether there is any scope remaining for the presence of any exotic dark matter in the universe.

For example, at the scale of clusters of galaxies, the virial theorem of Newtonian physics is used. However, such a system, albeit now chaotic, can again be viewed as a continuum of free-fall matter as was the case for the galactic scale. Indeed at the scale of individual galaxies as units within the cluster, the motions comprise a multitude of randomly oriented free-fall rotations. While the chaotic nature of these rotations within a cluster might have the effect of minimizing or even erasing the kind of phenomenon that we have witnessed in the systematic rotation of stars within an individual galaxy, it might be otherwise. Since general relativity was seen to make such a difference in the case of the galactic scale, clearly it is necessary to analyze the scale of the clusters anew.

The first step in this direction is described in the next chapter with an idealized spherically symmetric model of a cluster of galaxies. The results are both interesting and surprising. Moreover, they provide further evidence to counter the claims of some of our critics that it is the presence of singularities that led to our having general relativity account for the flat galactic rotation curves.

We emphasize that our focus to this point on single galaxy dynamics constitutes a first step. Apart from the need to fully extend the study to galactic clusters of a chaotic nature, the issue of gravitational lensing studies and the currently favoured structure formation models that appear to support the presence of non-baryonic dark

matter have yet to be considered. With regard to the lensing data, what is now to be taken into account is the non-linearity in the gravitational field solution for a rotating dust distribution that we have found. Up to this point, the mathematics of lensing had been treated as if the gravitational field were simply a Schwarzschild field. The current evidence for dark matter is discussed more fully in the next chapter.

The scientific method has been most successful when directed by "Occam's razor", that new elements should not be introduced into a theory unless absolutely necessary. If it should turn out to be the case that the observations of astronomy can ultimately be explained without the addition of new exotic dark matter, this would be of considerable significance. After all, a reduction of the mass of the universe by about one quarter of its assumed value, would have to be seen as worthy of note.

Chapter 10

Clusters of Galaxies

10.1 Preliminary comments

To this point, our focus has been on the dynamics of a single galaxy. We have seen that the observations of higher-than-expected velocities of stars within a galaxy could be rationalized with essentially the visible mass alone, within the context of Einstein's theory rather than Newton's theory. The latter requires vast stores of dark matter to accomplish the task. While it is almost everyone's preferred theory of gravity, Einstein's theory is far more complex and mathematically demanding. Many, if not the majority of astronomers (and a sizable fraction of physicists) have not studied general relativity to any depth, if at all. Therefore it is understandable that there should be such a high resistance on their part to consider that the cherished Newtonian gravity could be inadequate. However, almost everyone would acknowledge that there would be nothing wrong in using general relativity for galactic dynamics; only that it would be a waste of effort if Newtonian theory would be of sufficient accuracy and this sufficiency is what has been generally believed. Many have challenged our work to date using general relativity, and we have patiently replied to every challenge. Future challenges lie ahead in coming to grips with our central hypothesis that the exotic dark matter advanced by many researchers does not actually exist.

In this chapter, we consider the next astrophysical scale, the scale of a cluster of galaxies. Rather than the essentially organized rotational pattern of stars within a typical spiral galaxy, we are gener-

ally dealing with disorganized motions of a multitude of individual galaxies within a cluster. Ideally we should be turning to the general relativistic analysis of such a disorganized group of many mutually interacting gravitating elements. We would hope to achieve the general relativistic equivalent of the virial theorem of Newtonian physics, averaging over a chaotic system of very many elements. This is for the future. However for the present, we have succeeded in shedding light on the problem from quite a different direction.

10.2 Spherical dust collapse

Rather than dealing with the extreme of disorganization, we analyze the ultimate in organization. We focus on an idealized cluster of galaxies, a spherically symmetric swarm of particles with each particle in purely radial motion under the influence of gravity alone [79]. This system could be referred to as "spherical dust collapse"—"dust", because it is pressure-free motion and "collapse", because the particles of the structure are driven inwards by gravity. It should be noted that many researchers have come to view the use of the word "collapse" to signify the total crunching of matter to a singularity at the center. In this book, we are using the word to describe the radially infalling motion in generality.

We will see how considerable insight into the galactic cluster dynamics can be gained from the study of such a pure system. There are features that make it particularly desirable. Firstly, unlike our previous single galaxy study with stationarity, this system is dynamic with an explicit evolution in time. Secondly, the solution is simplified because of the spherical symmetry. No gravitational waves are produced for purely spherical systems. Thirdly, we benefit from the fact that the *exact* mathematical solution for this system is known [3]. The general solution can be particularized to a system of mass distribution and epoch of value for our study. We choose it to be one for which a ball structure of mass is evident with vacuum outside. In addition, we choose the stage of collapse where the field is weak everywhere at the epoch being analyzed, long before the crunching to a singularity has occurred.

To our knowledge, this state has never been considered before within the context of general relativity. We would suspect that such

a configuration would have been dismissed out of hand as one that could be analyzed effectively with Newtonian gravity. In the past, general relativistic collapse studies have focused on the extremes of strong gravity where the singularity is approached. Indeed we have participated in such a study with colleagues from the Tata Institute [16] where we studied the collapse with pressure. The great bulk of the research efforts with strong fields previous to our work had dwelt upon dust collapse. But this is totally unrealistic physically: for the stage at which the collapse is nearing a singularity, the densities mount to tremendous levels as the size shrinks towards zero volume and the pressures become enormous. However in simulating an idealized spherically symmetric cluster of galaxies such as we observe at present, it is globally weak gravity with relatively low density that exists and pressure can be ignored. The galaxies are seen as individually freely moving elements under their mutual gravity without collisions. This is ideal for our present purposes as there can be no question of any singularity issues arising at this stage of collapse, issues that have fixated our critics in the past when we studied the single rotating galaxy model.

We will consider this spherically symmetric collapsing system to exhibit high (but "non-relativistic", i.e. $v \ll c$) velocities for its elements, considerably higher than would be expected on the basis of its total mass according to Newtonian gravity. In doing so, we are retracing the situation confronted by Zwicky in the 1930s who was led on the basis of high galactic velocities within clusters to propose the existence of dark matter to provide the means to propel these galaxies. That he did so on the basis of observations of galaxies in the *non-symmetric* Coma cluster system rather than the presently considered idealized spherically symmetric model system is not important for our purposes; it is the principle that is important. We will consider whether general relativity rather than Newtonian gravity can explain the high velocities without any extra dark matter.

We reiterate for emphasis: by "high velocity", we mean velocity high compared to the expected Newtonian velocities but still much smaller than the speed of light, c. Later, we will be referring to "general relativistic velocity" which is velocity based on the theory of general relativity and which can take on the entire range of values. It should not be confused with the commonly used expression

"relativistic velocity" derived from special relativity which connotes velocity approaching c.

10.3 Velocity of particles falling in vacuum toward a spherical concentration of mass

To familiarize ourselves with the concepts and to develop a valuable basis for later comparison, we first develop the particulars of the motions of radially falling particles in the vacuum Schwarzschild spacetime. This has been expounded with particular clarity in [3] which we will now refer to as "LL". The spacetime geometry for the spherically symmetric mass m was given in (5.1) and (5.2).

We recall some of the development in the case of spherically symmetric vacuum spacetime that we described in Chapter 5. Because of the issues surrounding the metric functions when $r = 2m$, LL transform the coordinates r, t to a new system of coordinates R, τ (θ and ϕ are left the same) in which radially freely falling particles are at rest relative to the new coordinate system. As well, the $0-0$ component of the metric tensor is 1 and there are no space-time cross terms in the metric. LL refer to these coordinates as "synchronous".

The transformation is

$$\begin{aligned} \tau &= t + \int \frac{f(r)}{1 - \frac{2m}{r}}\, dr \\ R &= t + \int \frac{1}{f(r)\left(1 - \frac{2m}{r}\right)}\, dr \end{aligned} \tag{10.1}$$

where $f(r)$ is taken as

$$f(r) = \sqrt{\frac{2m}{r}}. \tag{10.2}$$

The result can be expressed as a simple relationship between the coordinates

$$r = \left(\frac{3}{2}(R - \tau)\right)^{2/3} (2m)^{1/3} \tag{10.3}$$

in the two reference systems.

The expression of the metric form is

$$\begin{aligned} ds^2 =& d\tau^2 - \frac{dR^2}{\left(\frac{3}{2(2m)}(R-\tau)\right)^{2/3}} \\ & - (2m)^{2/3}(d\theta^2 + \sin^2\theta d\varphi^2)\left(\frac{3}{2}(R-\tau)\right)^{4/3} \end{aligned} \tag{10.4}$$

in these new "comoving" (R, τ) coordinates. We recall that the metric has explicit time dependence in terms of the new coordinate system in spite of the fact that the spacetime is intrinsically static for $r > 2m$. This is understandable as particles at rest relative to these coordinates are physically in a state of free-fall, a feature of all synchronous systems [3]. As a result, from the vantage point of these particles, the view of the system is continually changing as the central mass draws continuously closer to the free-fall observer (who is at a given fixed R) as time advances. For a particle at a given R, the field at its position grows in strength as the proper time τ increases and becomes very strong as τ approaches R. In the process, this particle at R approaches the singularity and the corresponding r value approaches the singularity at $r = 0$. From the vantage point of the free-falling observer, the appearance of the spacetime is changing dramatically even though the intrinsic character of the spacetime is static.

In our later work on galaxy cluster simulation, we will concentrate on the weak gravity regime where $R \gg \tau$ for all R. In terms of the original system, this implies that $r \gg 2m$ for all r in the (r, t) frame. The distinction in time measurement to keep in mind is this: the time coordinate τ measures proper time, the time read by the clock of an observer in free-fall. The time coordinate t measures time read by the observer who is very distant from the central mass. (For the latter, whose distance from the source is very large, the distinction between free-fall and being at rest relative to the source becomes negligible.)

In the majority of papers on general relativity, the assumption is made that the gravitational field is strong. This is because of the ingrained bias that general relativity is only of real significance for such fields. An important goal of this book is to show that general relativity has consequences for weak field situations that are of

considerable physical significance. However, it is worthwhile for the sake of contrast, to work through a case where the general relativity treatment in the weak field limit does not produce anything new. We do so for the study of velocity.

Suppose we were to inquire as to the evolution of the radial velocity of a freely falling particle moving in the radial direction in the Schwarzschild field. Velocity is a relative quantity. Velocity relative to what? Velocity observed by whom? These are essential questions that come to mind immediately. As to the first question, of most interest is the velocity relative to the central mass. Even in Newtonian physics, when we study the solar system, it is the Sun as reference anchor that is the choice of interest. As to the second question, there are two observers of interest: first, the observer who is at rest relative to the central body adjacent to the falling particle as it passes him by and second, the observer very far from the central mass, the "asymptotic" observer, who is also at rest relative to the central body. In relativity, the perspective of the observer is also an issue. To find these particle velocities, it would be futile to work in the (R, τ) coordinate system because the particle is always at rest in this system.

Instead, we calculate in the familiar (r, t) system which is ideal for our purposes because the asymptotic observer reckons radial distance and time intervals as dr and dt respectively. He does so because for him, the metric is

$$ds^2 = dt^2 - dr^2 - r^2(d\theta^2 + \sin^2\theta d\varphi^2) \tag{10.5}$$

i.e. (5.1) with $\nu = \Lambda = 0$ which is the value approached by these functions as r approaches infinity.

The radial motion of the particle that is released from rest at infinity is described by the first integral of the geodesic equation (4.24), which can be expressed in the form [3]

$$\frac{dr}{dt} = -\left(1 - \frac{2m}{r}\right)\sqrt{\frac{2m}{r}}. \tag{10.6}$$

This is the radial velocity of the freely falling particle at any given $r > 2m$ as reckoned by the asymptotic observer. It is explicitly seen to be 0 at infinite r and of particular interest, it is also 0 at $r = 2m$ where the gravity is very intense. An integration of (10.6)

reveals that r approaches $2m$ as t approaches infinity. Thus, from the vantage point of the distant observer who measures the time t, the particle never actually reaches $r = 2m$ but only approaches it asymptotically, as suggested by the zero velocity value.

We now consider how an observer who is adjacent to the falling particle but is also at rest relative to the central body would reckon the velocity. With his different distance and time measures, he would use the more complicated proper measure ratio of radial distance to time in the form

$$v = -\sqrt{\frac{-g_{11}}{g_{00}}}\frac{dr}{dt} \tag{10.7}$$

where we adopt the naming of the coordinates as

$$(x^0, x^1, x^2, x^3) = (t, r, \theta, \phi). \tag{10.8}$$

When the metric coefficients and (10.6) are substituted, this proper local velocity is seen to be

$$v = -\sqrt{2m/r} \tag{10.9}$$

for the case where the particle is released from rest at infinite r. Consistently, when r is infinite, this velocity is again 0. However, at $r = 2m$, this velocity is 1 in magnitude, the speed of light in our system of units where we have taken $c = 1$.[a] This is in the ultimate sharp contrast to the value 0 as gauged by the asymptotic observer; it is an example of "relativity" in the extreme.

For our purposes it is of particular interest to contrast this relative observer issue with the case where the gravity is weak, $r \gg 2m$. Then, the $(1 - 2m/r)$ factor in (10.6) is approximately 1 and the local proper and asymptotic measures of velocity are approximately equal in the value $-\sqrt{2m/r}$. It is for this reason that we did not have to ask the "relative to whom?" question in the case of weak gravity. For particles falling at r values much greater than $2m$, we see that the general relativity measurements revert to the Newtonian measurements.

Given the blending of results in the case of weak gravity, it is perhaps understandable that prior to our study of radially falling

[a]Recall that no physical observer can be at rest at $r = 2m$ to make this measurement.

dust, it would have been assumed that the situation would follow in step with what was just described for the problem of a falling particle in empty space. However, we shall see that just as in the case of a single rotating galaxy that we considered in Chapter 9, the nonlinearities of general relativity bring interesting new elements into the analysis of falling dust. The interacting elements of the conglomerate do not contribute in a linear fashion.

10.4 The velocity of dust in collapse

It is particularly useful to begin with comoving coordinates for the analysis of dust collapse [3]. The metric is expressed as

$$ds^2 = d\tau^2 - e^{\lambda(\tau,R)}dR^2 - r^2(\tau,R)(d\theta^2 + \sin^2\theta d\varphi^2) \tag{10.10}$$

where a freely falling dust particle maintains constant space coordinate values for all time. In terms of these coordinates, the four non-trivial Einstein field equations are [3]

$$-e^{-\lambda}(r')^2 + 2r\ddot{r} + \dot{r}^2 + 1 = 0, \tag{10.11}$$

$$-\frac{e^{-\lambda}}{r}\left(2r'' - r'\lambda'\right) + \frac{\dot{r}\dot{\lambda}}{r} + \ddot{\lambda} + \frac{\dot{\lambda}^2}{2} + \frac{2\ddot{r}}{r} = 0 \tag{10.12}$$

$$-\frac{e^{-\lambda}}{r^2}\left(2rr'' + (r')^2 - rr'\lambda'\right) + \frac{1}{r^2}\left(r\dot{r}\dot{\lambda} + \dot{r}^2 + 1\right) = 8\pi\rho \tag{10.13}$$

$$2(\dot{r})' - \dot{\lambda}r' = 0 \tag{10.14}$$

where a dot denotes the partial derivative with respect to τ and a prime denotes the partial derivative with respect to R.

The *exact* solution of this complicated set of very nonlinear partial differential equations assumes a surprisingly simple form:[b]

$$e^{\lambda} = \frac{(r')^2}{1 + E(R)} \tag{10.15}$$

$$\dot{r}^2 = E(R) + \frac{F(R)}{r}. \tag{10.16}$$

[b] However, it should be noted that the form is simple only when the coordinates from both systems, i.e. r and R, are employed together.

where $E(R)$ and $F(R)$ are functions of integration. We will concentrate on the special case where $E(R) = 0$ corresponding to the configuration where the particles have been released from rest at spatial infinity in the infinitely distant past. The solution expressions for positive and negative $E(R)$ add extra complications which would detract from the essentials of the problem but could be useful for more detailed analysis. The three possible cases are analogous to the familiar classical mechanics problem of the toss of a ball in the vertical direction: if the imparted velocity is too small (the normal case for human pitchers), the ball will eventually stop and fall back to the ground. This corresponds to $E(R) < 0$. However, if the toss is of super-human strength, the ball will never return, corresponding to positive $E(R)$. The critical case corresponds to our $E(R) = 0$ case for dust; the ball just makes it to infinity, asymptotically approaching zero velocity in the process.

For $E(R) = 0$, the r value of the collapsing dust solution can be expressed in terms of the comoving coordinates as

$$r = \left(\frac{9F}{4}\right)^{1/3} (\tau_0(R) - \tau)^{2/3} \tag{10.17}$$

with $\tau_0(R)$ being an arbitrary function of integration. For our purposes, we choose $\tau_0(R) = R$ so that this solution interior to the ball of dust joins smoothly with the exterior vacuum Schwarzschild metric (10.4) describing the geometry of the empty space surrounding the collapsing ball of dust.

In general relativity, the entire energy-momentum tensor drives the evolution of the gravitational field. However, with dust, there is no pressure present, and therefore, the field is produced by the density alone. For all three cases, positive, negative or zero E, the density ρ couples to the field by the T^{00} field equation

$$8\pi\rho = \frac{F'}{r'r^2}. \tag{10.18}$$

By a simple integration of (10.18) (see [3]), we find that the mass $M(R)$ within the region of the ball defined by the radial coordinate R is

$$M(R) = F(R)/2. \tag{10.19}$$

Therefore the entire mass M is given by $M(R_0)$ where R_0 is the outer comoving radial coordinate of the entire dust ball.

The matching of the interior dust solution to the exterior vacuum Schwarzschild solution is easily achieved. With this spherically symmetric geometry, there is no scope for confusing the presence of a boundary surface layer of mass. The matchings of spherically symmetric interior matter distributions with the Schwarzschild exterior vacuum metric are present in the literature. Probably best known is the Schwarzschild static constant density ball with pressure matched to the exterior Schwarzschild vacuum metric. One of the motivating factors for our study of spherical dust collapse was to show that the importance of general relativity that we witnessed for the stationary rotating galaxy problem re-appears here where there is no issue of singular mass layer that can be raised.

Our astronomers on Earth are "asymptotic" observers, very distant from the galaxies being observed. Therefore the radial dust velocity that we require is that measured by these observers, the dr/dt that we discussed previously in the case of test particle motions in vacuum. The local velocities, those measured by observers adjacent to the matter, are not relevant for the distant observers here on Earth.

The field equations and solutions have been conveniently set up in terms of the comoving frame of reference. However, to determine the velocity recorded by the asymptotic observers, we must again change to a non-comoving frame since velocities are zero relative to the comoving frame. For this purpose, we continue to follow the approach taken by LL for vacuum and evaluate the radial velocity dr/dt in the Schwarzschild-like (r, t) coordinate frame. However, for our dust collapse case, there are some interesting differences. In the vacuum case, there was no material present whose motion we could otherwise track. Instead we injected a test particle to probe the geometry's effect on matter that would be driven by this field were it to come into the environment. Being a test particle means that it is affected by the field but it does not alter the field. The test particle's motion is derived from the geodesic equations that we discussed before in Chapter 4.

However, in the case of dust, we already have the motion description implicit in the solution to the field equations but in an inconvenient reference frame, the comoving frame, relative to which the particles are all at rest. Therefore the approach here is somewhat

different in that we only need to transform the known solution to the convenient (r, t) coordinate frame to render the motion explicit. It should be noted that any infinitesimal element of dust also satisfies the geodesic equations because it is a test particle in having negligible mass and is moving freely under gravity alone, as there is no pressure. However, it should be kept in mind that it moves under the influence of the conglomeration of all of the dust elements in the distribution. While as individual elements, each can be treated as a test particle, as a collective their character is subsumed into the form of a nonlinearly interacting mass.

Thus, while the motion could have been derived for the dust using the geodesic equations applied to the dust metric, the mathematics required would have been daunting. While still fairly challenging, the mathematics required to change the existing dust solution to the (r, t) frame is considerably easier.

There is another useful aspect to the (r, t) frame: since the coefficient of the angular part of the metric in this frame is r^2, the circumference of a ring of particles at r assumes the familiar flat-space value $2\pi r$ for both the proper measure and for the measure as judged by distant observers. Thus, a flat space aspect is retained which helps in the visualization of the structure.

We now outline the procedure for deriving the velocity of the radially falling elements of the distribution as gauged by the distant observers. For consistency with the solution form of (10.3) for later blending with the vacuum solution at the boundary of the dust ball, we choose $\tau_0(R) = R$. For maximum generality, we express the general form of transformation with initially arbitrary functions $p(r, t)$ and $q(r, t)$ in the form

$$\sqrt{F}R = p(r, t), \quad \sqrt{F}\tau = q(r, t) \tag{10.20}$$

with the constraint

$$p(r, t) - q(r, t) = (2/3)r^{3/2} \tag{10.21}$$

for consistency with the solution. From (10.21), we see that

$$p'(r, t) - q'(r, t) = r^{1/2}, \quad \dot{p}(r, t) = \dot{q}(r, t) \tag{10.22}$$

where a dot over these functions denotes the partial derivative with respect to t and a prime on these functions denotes the partial derivative with respect to r.

To assist those readers who might wish to delve into the details of the rather lengthy and complicated derivation of the velocity, we outline the steps with some detail in Appendix B. The essential quantity for consideration is the final result of this derivation, the velocity as viewed by distant observers in (10.23):

$$\frac{dr}{dt} = -\frac{(\alpha+\beta)(1-\beta^2)}{8\pi r^2\rho^2}\left[\frac{\alpha}{F} + \beta\left(\frac{F''}{(F')^2} - \frac{1}{2F}\right)\right]^{-1}\frac{\partial\rho}{\partial t} \qquad (10.23)$$

where, from (B.2) and (B.3),

$$\alpha = \frac{rF'}{3F}, \quad \beta = \sqrt{\frac{F}{r}}. \qquad (10.24)$$

It stands in sharp contrast to the very simple Newtonian-like expression

$$v_{\text{local}} = -\beta = -\sqrt{\frac{F}{r}} \qquad (10.25)$$

for the velocity v_{local} as gauged by observers at rest relative to the center of the distribution who are in the vicinity of the radially falling material. While it would greatly simplify computations if one were to use (10.25) to refer to radial velocity for distant observers as a Newtonian would wish to do, one who regards general relativity as the correct theory of gravity would have to use (10.23). While one might lament the complexity of the latter, alternatively one could rejoice in its richness. Regardless of the emotions, this is the result.

For a Newtonian observer, it is simply the mass interior to the sphere at the radius of the matter being observed, that drives the velocity. For the general relativity observer, it is this value as well *but only if he is doing the measuring in the locality of the material.* However, for the distant observers, the velocity expression (10.23) is the relevant expression for velocity and it is far more complicated: it depends not only on the usual interior quantity of mass but also the reciprocal of the local density squared and its time rate of change (which can also be expressed as the time rate of change of reciprocal density $1/\rho$) and using (10.19), the gradient of the mass within the radius in question, $M'(R)$ as well as its gradient, $M''(R)$.

Using general relativity, it is interesting to compare the velocities of the dust elements in a distribution with the velocities of test particles in vacuum for both very strong gravity and for weak gravity.

From (10.23), in the limit of very strong gravity, with β approaching 1, the observations parallel what was the case in vacuum. The local observers see the velocity approach 1 in magnitude while the external observers see the velocity approach 0.

However these equations show that for weak gravity with $\beta \ll 1$, the vacuum and dust comparison is very different. While from (10.9) and (10.6) the local and asymptotic velocity measures for observers plotting freely falling test particles in vacuum in the field of a concentrated mass are approximately the same, namely $-\sqrt{2m/r}$, the corresponding velocities for local and asymptotic measure in the case of dust are very different in general: the velocity is simply β for the local measure whereas the asymptotic measure is given by the various factors in (10.23) with $1 - \beta^2$ approximated by 1. Indeed, given the complexity of the form of dr/dt in (10.23), it would be a very special occurrence for dr/dt to have the value β. Therefore, had the astronomers in the 1930s been using general relativity rather than Newtonian gravity, they would not have had any reason to expect to see β values for velocities and the need for additional mass in the form of some mysterious dark matter would not have arisen.

10.5 Observing an idealized galactic cluster

Thus far in this chapter, we have considered a purely radially moving conglomeration of dust within the context of general relativity and have suggested linkages to the motions of galaxies within a cluster. Now we will develop the linkage further with fit to known data.

It is interesting to consider that while most of the gravity in the universe is weak gravity, most of the interest in general relativity concerns situations where the gravity is very strong. This is understandable as general relativity had been thought to be of real consequence primarily in those situations where the gravity is very strong. Newton's gravity had been considered perfectly adequate for most situations, a notable exception being cosmology where the global structure of possibly infinite matter cannot be adequately dealt with in Newtonian gravity. Thus books have been written on astrophysics in which general relativity has been given only lip service, even while acknowledging it to be the premier theory of gravity.

At the scale just below cosmology is the scale of clusters of galax-

ies. When the gravity was deduced to be weak within these clusters, astronomers naturally turned to Newtonian gravity to correlate the seemingly anomalously large galactic velocities that they measured with the masses that they believed to be present. In this manner they initially deduced that there must be unseen "dark matter" in the order of 100 times as much as the visible matter to make the mass totals accord with the velocities. However, with the later discovery of very large quantities of gaseous matter, this figure was reduced dramatically but there still remained a large quantity of matter yet to be accounted for. This apparent need is still promoted vigorously by researchers throughout the world. It has spawned a plethora of papers advocating new particles that would conceivably play the role of this exotic missing material. However, we have seen that insofar as high rotational velocities of stars in galaxies as the basis for the need for dark matter is concerned, the replacement of Newtonian gravity by general relativity gravity removes this requirement. An essential point is that the nonlinearities of general relativity play an important role in systems of freely gravitating masses, leading to expressly non-Newtonian behavior, even when the gravitational field of such systems is weak.

Insofar as gravitational field strength is concerned, we focus on the Coma cluster of galaxies for which we have some directly useable, albeit very limited, data provided by J. P. Hughes [80]. For this cluster, the ratio $2M(R_0)/r_0$ is of order 10^{-4} if we assume as would a Newtonian, that there exists dark matter present to account for the observed velocities and of the order 10^{-5} if we accept only the existence of the matter that we see. In either case, with this ratio being much smaller than 1, the gravity is very weak.

In this ratio, the subscript 0 indicates the outermost radius of the cluster. However, this outer radius is expressed in terms of the two coordinate systems. As well, in its most convenient form, the radial velocity (10.23) is also expressed in terms of r and R. The r coordinate has a direct measurable connection to the source in that $4\pi r_0^2$ is the surface area of the system as a whole. However, R is a comoving radial coordinate, an abstract labeling of the radial positions of the elements of the system for all time. The R coordinate has no *a priori* geometrical connection. Consequently there is a great deal of arbitrariness attached to the choice of the numerical value of

R_0. From the transformation equations, we see that for a given value assigned to r_0, different settings of the zero value for the clock will change the number attached to R_0. Clearly a logical form of normalization is called for. For the sake of coordinated measurement, we attach the same geometrical scale to the R coordinate as we attach to the r coordinate and we can achieve this by the proper zero setting of the clock. By the correct choice, we make $R_0 = r_0$ so that the average density of the system calculated using r_0 will equal that using R_0 and the surface area measure of the system will be equal in the two systems using the same numerical value for the radius.

A typical cluster of galaxies would consist of elements whose motions are randomized and in future research, it will be valuable to develop a framework for virialized motion within general relativity. What we do have at present is the required structure to analyze an idealized system of purely radially moving elements. We apply this as a test model for the Coma Cluster of galaxies. As reported in [80], at a radius of 1 Mpc, the total cluster mass, including dark matter, is given as $6.2 \times 10^{14} M_\odot$, with the 13%–17% portion being normal baryonic matter. Within a radius of 3 Mpc, the total mass is reported to be $1.3 \times 10^{15} M_\odot$, with the normal luminous matter portion within the wide range of 20%–40%.

We can fit these data with an accumulated mass function

$$F(R) = k_1 R^{k_2}, \tag{10.26}$$

(k_1, k_2 constants) as shown in Figure 10.1. By choosing the function in this form, we have $F(0) = 0$ from (10.26). Thus, there is no mass present in the vanishing volume of $R = 0$. As a consequence, there is no singularity at the origin [3]. Using (10.26) in conjunction with (10.18), we are able to plot the density profile for the distribution. The graph of the densities for the two extremes of the uncertainty range and the average is shown in Figure 10.2.

From the viewpoint of the distant observer, the velocity associated with each $F(R)$ is given by (10.23). Clearly, it would be a unique unusual circumstance to have this velocity be the same as the normally attributed velocity β that a Newtonian observer would deduce. We are interested in determining what conditions are required to have this velocity match the actual velocity as viewed by astronomers from the Doppler shifts. It is convenient to introduce a

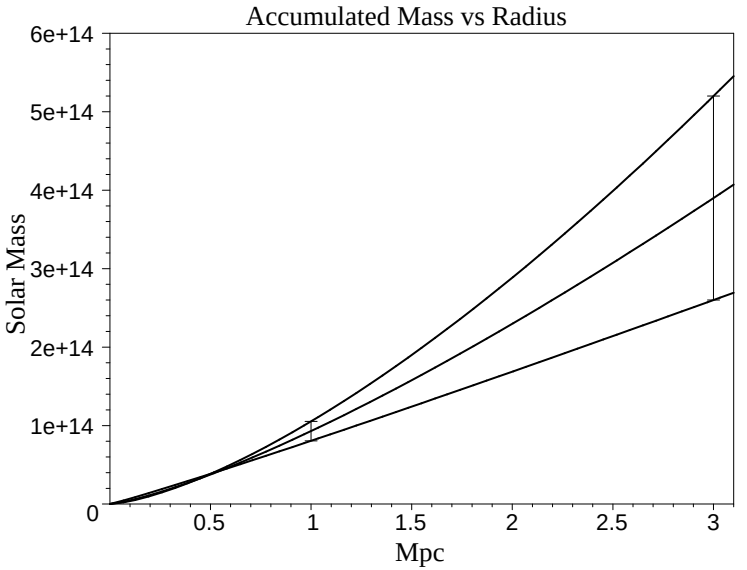

Figure 10.1: The upper, middle and lower limits of mass accumulation vs radius are described by the functions, $F = 6.641 \times 10^{-16} R^{1.453}$, $F = 1.244 \times 10^{-12} R^{1.305}$ and $F = 2.531 \times 10^{-7} R^{1.066}$ respectively.

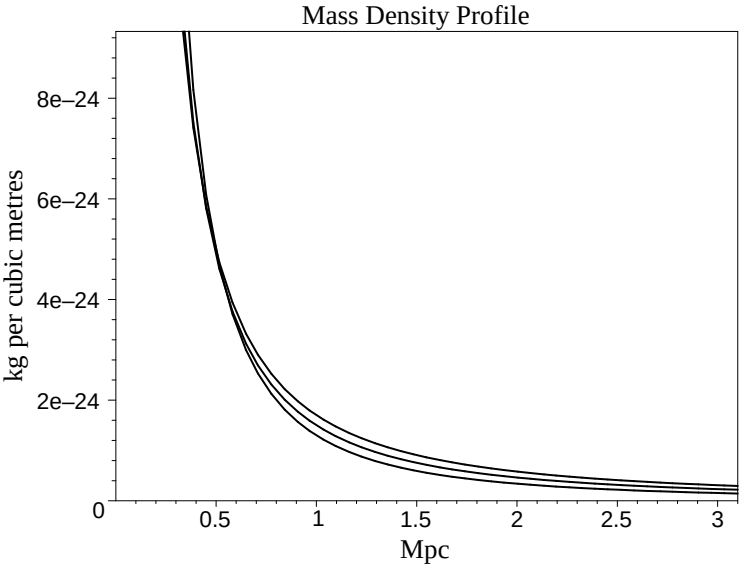

Figure 10.2: From the three functions, $F = 6.641 \times 10^{-16} R^{1.453}$, $F = 1.244 \times 10^{-12} R^{1.305}$ and $F = 2.531 \times 10^{-7} R^{1.066}$, we can derive the mass density profile as shown in the graph.

numerical "boost" factor n with the "boosted" velocity defined by

$$dr/dt = -n\beta. \tag{10.27}$$

Here β, as throughout the book, is composed from the baryonic mass alone and n is the booster number to bring dr/dt to the observed level of velocity. In general, n could be less than 1 as well as greater than 1. However, in the present context, we are interested in the conditions that make n greater than 1 to accord with the observations, hence the description "booster".

Assuming the baryonic mass is 20%, 30% and 40% of $1.3 \times 10^{15} M_\odot$, we find that the boost factors n are 2.23, 1.82 and 1.58, respectively to bring $n\beta$ to the observed value of velocity. With the aid of the accumulated mass function (10.26), we can determine all of the elements of the RHS of (10.23) apart from the time rate of change of density, $\partial\rho/\partial t$. We equate the RHS of (10.23) to $n\beta$ and solve for $\partial\rho/\partial t$. The results are: $2.13 \times 10^{-41}\text{kg/m}^3/\text{sec}$, $2.62 \times 10^{-41}\text{kg/m}^3/\text{sec}$ and $3.02 \times 10^{-41}\text{kg/m}^3/\text{sec}$, respectively. Rates of density change of the order of magnitude $10^{-41}\text{kg/m}^3/\text{sec}$ are quite reasonable as over a period of one billion years, the density would grow by the order of 10^{-25} kg/m^3, roughly doubling the value of the present density.

While this is only one example, and a very rough one at that, it can be said that we have been able to account for the observed velocities of galaxies within a cluster. This has been accomplished with the simplification of spherical symmetry, simply using the reasonable ingredients furnished to us *and solely within the framework of general relativity and without any extraneous dark matter.* While Newtonian gravity required only the accumulated mass within a given radius to determine the velocity of the material at that radius, using general relativity, we now require the new elements of local density, its time rate of change, the gradient of the mass interior to the observation point as well as its gradient as additional factors. It is interesting to reflect that the usual Newtonian factor would come into play were the observer viewing the galaxy motion in the neighborhood of the galaxy rather than at a distant point. We should not be surprised by this: general relativity, just as we are familiar from special relativity, brings in the state of the observer as a particularly important element in the perception of physical phenomena.

An immediate objection one might raise is that we have considered here an idealized case of perfect spherical symmetry. However, it would seem reasonable to expect comparable effects for non-spherical accumulations of freely-gravitating collections of bodies as we have in clusters of galaxies. Had Zwicky made this calculation 70 years ago with general relativity in mind, he might have come to very different conclusions regarding the requirement for vast stores of exotic dark matter.

In this chapter, we have focused on the simplest case, $E(R) = 0$, of an idealized cluster, with perfect spherical symmetry. Firstly, even within the confines of this perfect symmetry, it will be useful to explore the effects of both positive and negative E values. Secondly, it will be of great interest to work at the extreme opposite to perfect spherical symmetry, namely chaotic motion. In Newtonian gravity, it is the virial theorem that plays this role. The challenge will be to achieve the equivalent within general relativity. Also to be considered is the issue of the interpretation of lensing as a mechanism for the deduction of mass. For consistency, lensing must show the same mass as the observation of Doppler shifts. Therefore the subtleties of general relativistic weak gravity that we have discussed in this chapter must now be directed to the consideration of lensing.

As we saw in Chapter 9 in the description of the rotational velocities of stars in spiral galaxies and in [65], [66], we see in this modeling of an idealized cluster of galaxies the effect of the nonlinearities inherent in general relativity in the context of weak gravity to effect very significant changes relative to the results expected on the basis of Newtonian theory.

With these new results at hand, it is natural to consider whether there exist other problems in astrophysics that could lend themselves to reinterpretation with the application of general relativity where Newtonian gravity had been used to this point. As well, we suggest that an experimental design could be considered to formulate a direct controllable test of what we have developed here. This could ultimately lead to a new independent test of general relativity.

10.6 Current evidence for dark matter

It is generally acknowledged that the flat galactic rotation curves of stars in spiral galaxies and the observations of the high velocities of galaxies in galactic clusters have provided the most compelling long-established sources of evidence for the existence of vast stores of dark matter. The 1975 deduction by Rubin that about half the mass of a galaxy was contained in a mostly dark galactic halo was eventually revised drastically upward. This came about from measurements of diffuse interstellar gas at the edge of galaxies with application of the Newtonian virial theorem. At the present time, dark matter is typically assigned to constitute up to approximately 95% of a galaxy's mass on this basis. However, given our results that general relativity provides a more complex picture of the dynamics than that from Newtonian gravity, such a Newtonian-based deduction cannot be relied upon. The required general relativistic equivalent of the Newtonian virial theorem which would be definitive, is yet to be realized.

We have shown that general relativity could accommodate typical galactic rotation curves with relatively small amounts of dark matter and hence not of the exotic variety that the description "dark matter" generally connotes. This was achieved *with given velocity data over a plane through the visible disk*. Our work brought into question the existence of the large spherical halos of dark matter that are currently believed by probably most astronomers to surround the visible contents of galaxies. The question naturally arises: could general relativity also accommodate the presence of these massive external halos with the given flat rotation curves over a single plane? The answer is "yes". General relativity demonstrates its greater richness than Newtonian gravity in this way. While Newtonian gravity demands the extra matter from the limited data, general relativity declares that this data is insufficient to determine the full extent of the matter. We then discussed a velocity dispersion test with data *beyond* the plane. This would determine the actual extent if any, of matter beyond the visible disk with general relativity being the guiding theory. After all, if nature should actually present us with systems of mass distributions as presently generally believed to exist, general relativity as the premier theory of gravity should be brought to bear as the best indicator of such distributions. With such sparse

dispersion data presently available, the dispersion test is yet to be applied. Hopefully adequate data will eventually be available for the dispersion test.

Working against the presence of dark matter, however, is the dark matter theory of galaxy formation. The theory predicts that 10 to 100 times the number of small galaxies than that which are observed are permitted. In our view, this fact alone should be an issue of concern for those who accept the dark matter premise without reservation. Other challenges to the existence of dark matter that have been brought forward by researchers involve density cusps and the distribution of angular momentum. However, many see these issues as having been resolved (see for example [81]).

Earlier, we mentioned gravitational lensing which has emerged as an important tool in astronomical research. Foreground masses act as lenses for the light coming from background masses that gets bent by the foreground mass in its journey to us on Earth. In so-called "strong lensing", the background galaxies appear distorted into arc shapes as a result of the gravitational lens. While the claim has been made that the interpretation of the arcs translates into the presence of large reservoirs of dark matter within clusters, it is unclear whether the subtleties of general relativity have been fully implemented in the analysis. In the previous sections as well as in Chapter 9, we have seen the richness and the complexities of general relativity in action in rendering unsuspected interpretations and results. Insofar as lensing by galaxies is concerned, an immediate consideration is the role of general relativity in a careful consideration of rotating mass metrics. The same could be said of "weak lensing" results, in which the large number of very small distortions due to foreground masses of light from background galaxies are observed and statistically analyzed. This is an area for future research.

Turning to clusters of galaxies, we have seen in our simple model that the connection between the observed velocities of the individual components and the mass of the system driving the components is far more complicated on the basis of general relativity than is the case when simple Newtonian gravitation is applied. We have seen that on the basis of the simple model, the observed Coma Cluster velocities can be rationalized with general relativity applied to only the visible contents of the system. While the model is simplified

as an idealized perfectly spherical system, and a general relativistic formalism for chaotic systems is yet to be realized, this result is further cause to question the pronouncements that have been made concerning supposed vast quantities of dark matter within clusters.

A key cited indicator for the reality of dark matter is derived from the "Bullet Cluster" [82]. The claim is that in this system, there was a collision between two clusters of galaxies which resulted in a separation between the dark matter and the normal visible baryonic matter contents, the latter being concentrated in the middle of this complex system. According to the researchers, weak lensing data suggest that much of the system's mass, which is dark, lies outside the central region. On this basis, a scenario is formulated as follows: In the process of the collision, electromagnetic interactions between passing gas particles broke their speed, concentrating them near their region of collision at the center whereas the dark matter which does not interact electromagnetically, sailed through and is located outside the central region.

While this might at first glance appear to be solid evidence for the existence of dark matter, a study of the cluster Abell 520 [83] indicates a more complicated picture. In the authors' words: "The rich cluster Abell 520 ($z = 0.201$) exhibits truly extreme and puzzling multi-wavelength characteristics. It may best be described as a 'cosmic train wreck.' It is a major merger showing abundant evidence for ram pressure stripping, with a clear offset in the gas distribution compared to the galaxies (as in the bullet cluster 1E 0657-558). However, the most striking feature is a massive dark core ($721 h_{70} M_{\odot}/L_{\odot B}$) in our weak lensing mass reconstruction. The core coincides with the central X-ray emission peak, but is largely devoid of galaxies. An unusually low mass to light ratio region lies 500 kpc to the East, and coincides with a shock feature visible in radio observations of the cluster. Although a displacement between the X-ray gas and the galaxy/dark matter distributions may be expected in a merger, *a mass peak without galaxies cannot be easily explained within the current collisionless dark matter paradigm.*" (Italics are our own.)

What this indicates to us is that the state of the dark matter theory is extremely fragile. There is much more to observe and consider before one can reasonably accept what has become the prevailing dark matter picture.

Currently, various researchers cite the evidence from the density fluctuations in the cosmic microwave background radiation (CMB) as the very best evidence for the existence of both dark matter and dark energy. The spectrum of density fluctuations from the WMAP data is analyzed and the primary peak is said to be a direct measure of the total energy density of the universe. The view is that the last scattering surface at decoupling, the baryons falling into overdense regions and bouncing back in response to radiation pressure, is being recorded. The analysis is said to involve few uncertainties and leads, according to the proponents, to a deduction that the universe is of the spatially flat variety, $k = 0$ or at least very close to this value. Thus, if this result is truly so reliable, and with visible matter making up only a few percent of what would be required to produce a $k = 0$ universe, the great bulk of the universe mass must come from dark matter and the mass equivalent! of the dark energy. The split between the two types could be debated but the majority favor the percentages discussed earlier. The early universe studies are the glamorous focus of current interest by the scientific community. This is understandable as the microwaves reaching us have come to us from such an enormous distance and given the finite speed of propagation, they reflect the state of the matter of the universe from the very distant past. This is archaeology on the cosmic scale. Some have even waxed poetic, referring to the CMB maps as the face of God. However, while progress in early universe study has certainly been impressive, it strikes us as somewhat presumptuous to rush to firm pronouncements as to what the true implications of the CMB data entail. Further alternative scenarios should be, and surely will be explored in the years to come.

Returning to the envisaged unknown material that is said to dominate the matter content of the universe, the potential exotic non-baryonic dark matter has in the past been divided into three categories:

1) Ultra-relativistic particles called “hot dark matter”.

2) Non-relativistic particles called “cold dark matter”.

3) Relativistic particles called “warm dark matter”.

For many readers, this might begin to have similarities to the tale of Goldilocks and the three bears. We now proceed to do a taste test from the dark matter menu.

While the first variety is generally regarded as least favored if not totally rejected, it is ironic that in this class, we have dark matter particles that in fact are well-established elements of particle physics, namely neutrinos. However, being well-established leptons, we should not really regard them as exotic. While their existence is secure, current bounds on ordinary neutrinos indicate that they contribute only a small amount of dark matter. As well, hot dark matter does not fit into the currently favored scenario of galaxy formation after the Big Bang.

Cold dark matter is generally regarded by the great majority of dark matter proponents as the most likely form of dark matter. As a result of their becoming non-relativistic at the very early stages, they are hardly diffused and hence are said to provide the required clumping mechanism for structure formation in the early universe.[c] The COBE and WMAP satellites have measured the very small deviations from perfect isotropy of the early universe and these are interpreted as small-scale clumping, providing the seeds of galaxy formation. The problem, however, is one of identifying the source for this clumping. A lack of inventiveness has never ailed the particle physics community. While big bang nucleosynthesis has been deemed to rule out regular baryonic matter in the form of MACHOS (massive compact halo objects) as an adequate source, new never-seen particles beyond the Standard Model of particle physics have been proposed for this purpose. Candidates are WIMPS (weakly-interacting massive particles), some varieties derived from a proposed theory extension to the Standard Model called Supersymmetry. In this theory, every known lepton has a supersymmetric bosonic partner and every known boson has a supersymmetric leptonic partner. A variety of other kinds of WIMPS have been proposed such as axions (see [84] for a brief review).

[c]In building a scenario for dark matter in the early universe to provide a clumping mechanism, it is easy to lose sight of the fact that speculation is inherent in the process. CMB maps detect the clumping but there is no definite knowledge as to what came before this earliest snapshot of the universe. There is only speculation that the earlier picture had no clumping and that dark matter was required for the realization of the revealed picture.

WIMP enthusiasts face a delicate problem however, in that if they exist, trillions of these particles must pass through the Earth every second. While a great effort has been launched in laboratories around the world (and yet further searches are being planned), the awkward fact is that not a single confirmed WIMP detection has ever been made. Part of the interest in the current LHC (Large Hadron Collider) experiments at CERN entails the search for WIMPs.

A very small contingent of researchers aims to avoid such difficulties by favoring warm dark matter (but often as a small admixture to cold dark matter), consisting of particles more massive than neutrinos. Candidates for this class are sought once more in the supersymmetry zoo, this time in the form of partners to the photon and the graviton, called photinos and gravitinos repectively.

The reader will be excused for wondering whether this kind of theorizing may have gone a bit too far. From our perspective, we have to wonder how different physics might have evolved had the early astronomers made themselves more knowledgeable about general relativity and had made the kinds of calculations for galactic velocities that we have described in this book. Indeed for the most part, astronomers continue to ignore general relativity in making deductions from their observations. Thus, an industry has arisen of massive computer simulations with billions of conjectured dark matter particles. The claim has been made that these simulations confirm that the CDM (cold dark matter) model of structure formation is in accord with observed structures in galaxy surveys such as the Sloan Digital Sky Survey. However, the basis for these simulations is Newtonian gravity. The lesson from our work is that the best theory of gravity, general relativity, is capable of providing surprises.

This page intentionally left blank

Chapter 11

Closed Timelike Curves and Time Machines

11.1 The background

A recurring theme in science fiction and the popular media centers on the notion of time travel. Typically the hero travels into the past and performs some miraculous deeds to the delight of everyone, most of all the promoters of the tale. The hope of the latter is that the consumers of such inventions will suspend thought and happily digest the product. However, some alert consumers might feel uncomfortable with the idea that the hero could have just as well been a villain and could have murdered his great-great-grandmother who was a child at the time of his travel into the past. Then the disturbing question arises: how could he have lived to take this journey in the first place?

Good question. Surprisingly, some physicists to this day have maintained the possibility of such time travel, skirting around the causality issues by proposing the following: while the traveler could indeed co-exist with his distant forebear, he would not be able to perform the evil deed. Hmm.

It is well to ask how the notion of travel to the past entered the realm of physics. To our knowledge it began with the distinguished academic K. Gödel.

Kurt Gödel (1906–1978) was one of the most highly renowned logicians and mathematicians of the 20th Century. His contributions to the fields were varied but his best known works were his two "incompleteness" theorems. He developed a close friendship with Einstein. Sadly, in later life he became obsessed with fear of being poisoned and he starved himself to death.

In an interesting paper, I. Ovsvath and E. Schucking [85] describe Gödel's excitement in 1949 upon learning that in his new cosmological solution of the Einstein equations, one has the ability to "travel into the past". They discuss a lecture on the subject that he gave at the renowned Institute for Advanced Studies with the attendance of such luminaries as Einstein, Oppenheimer and Chandrasekhar. We might suppose that the lecture was possibly preceded by a hearty lunch for the trio followed by intervals of slumber during Gödel's presentation as there is no record of their having raised any objections at the time.

In 1956, W. Kundt investigated the Gödel solution and calculated the geodesics for this spacetime [86]. Chandrasekhar returned to the issues raised by Gödel as he and J. P. Wright [87] independently re-calculated the geodesics. From the geodesics, they found no evidence for travel into the past. Technically, the capacity for such travel translates mathematically into the presence of "closed timelike curves" (CTC's) in the spacetime. The curves being timelike renders them capable of being traversed physically by a time-traveler, and being closed means that the traveler returns to earlier spacetime events. (Unstated in the wording of the definition is that the curves must always be future-directed, that they always proceed into the future part of the light cone.) As a result, they rejected the claim of Gödel.

However, in 1970, H. Stein [88] noted that Gödel had never claimed that his CTC's were geodesics, suggesting that time travel was still a possibility. If it were travel with a spaceship, the difference would be in the necessity of having the rockets turned on rather than off (i.e. free-fall) for the voyage.

The interest in the subject as a part of physics as expressed by publications and citations was rather limited for many years. However, in their well-known book, S. Hawking and G. F. R. Ellis [89]

included the spacelike, null and timelike curves of the Gödel spacetime in a diagram. It is possible that this inclusion was the factor that led many authors to take the subject more seriously. Subsequently, quite a variety of papers on CTC's, time machines and related exotica (see, e.g. [90], [91], [92]) appeared. Claims were made that CTC's were necessarily present in a variety of other spacetimes. Surprisingly, there are researchers who have even taken the CTC notion so seriously as to propose experiments to look for the presence of CTC's in nature. In addition to the issue of causality violation, the acceptance of CTC's as an element of physics brings in the issue of entropy flow as one would have to face the violation of the Second Law of Thermodynamics, one of the holy grails of physics.

11.2 Creating closed timelike curves and Gödel's spacetime

It is very simple to create a CTC in flat space. Suppose one were to follow a circular path starting at the angular position $\theta = 0$ at some time, let us say 1 PM and completing the journey, reaching $\theta = 2\pi$ at 2 PM. This is a very mundane journey and we raise no alarms when we identify the angular positions $\theta = 0$ and $\theta = 2\pi$. However, suppose we were to identify the time 1 PM with the time 2 PM creating a CTC. Mathematically, there is nothing to prevent such a identification. However, we would reject this time identification but not the space identification because of our experience in the physical world: physically we understand that we can return to the same place but only at a later time, never the same time. This naturally raises the question as to whether the identifications in the Gödel spacetime follow the same pattern or whether for this case, there does exist some underlying physical basis for the identification of the time coordinates as well as the space coordinates to actually produce a physical as opposed to a purely mathematical CTC.

To answer this question, we probe the Gödel spacetime. The Gödel metric is a member of the class given in [93]:

$$ds^2 = -f^{-1}[e^{\nu}(dz^2 + dr^2) + r^2 d\phi^2] + f(d\bar{t} - w d\phi)^2 \qquad (11.1)$$

where f, ν and w are functions of r and z with coordinate ranges

$$-\infty < z < \infty, \quad 0 \leq r, \quad 0 \leq \phi \leq 2\pi, \quad -\infty < \bar{t} < \infty \qquad (11.2)$$

and $\phi = 0$ and $\phi = 2\pi$ are identified as usual. There is an interesting issue that emerges with this metric insofar as time is concerned. It arises because of the change of sign of the $g_{\phi\phi}$ component of the metric. The sign flips when

$$f^2 w^2 = r^2 \tag{11.3}$$

in the metric component

$$g_{\phi\phi} = -f^{-1}(r^2 - f^2 w^2). \tag{11.4}$$

As a consequence, the normally spacelike coordinate ϕ becomes a timelike coordinate for

$$f^2 w^2 > r^2. \tag{11.5}$$

In this case, the spacetime curve

$$\bar{t} = \bar{t}_0, \quad r = r_0, \quad \phi = \phi, \quad z = z_0 \tag{11.6}$$

with z_0, r_0, $\bar{t}_0$ being constants has been created as a CTC as a result of the now-timelike coordinate ϕ having $\phi = 0$ and $\phi = 2\pi$ *still being identified* as we had when ϕ was a spacelike coordinate. When

$$f^2 w^2 < r^2 \tag{11.7}$$

the situation was quite different. The same identification in ϕ led to the curve as again being closed but now spacelike at a given value of time rather than being timelike. Note that in (11.5), the metric has *two* timelike coordinates $\bar{t}$ and ϕ. For the curve under consideration, one coordinate $\bar{t}$ is held fixed while the other coordinate ϕ advances. This is certainly unusual but in actuality there is nothing mathematically wrong in coordinatizing a spacetime with more than one timelike coordinate.

Synge [94] has given an extreme example where a spacetime is described with four timelike coordinates. Having four timelike coordinates refers to having all of the diagonal terms of $[g_{ij}]$ positive. If it is a physical spacetime, the signature, derived from its eigenvalues, will still have the signs $(+ - - -)$. However, it is confusing to have a timelike coordinate held fixed in the description of a timelike curve. But of essential concern to us is the issue of physical reality. Is the usual interpretation that has been attributed to the CTC's of the Gödel spacetime necessarily the physical necessity that has been believed prior to our work?

11.3 Re-examining the standard closed timelike curve interpretation

There are two points that should have raised concerns as to the viability of the standard CTC interpretation:

Firstly we would argue that the interpretation becomes suspect when a timelike coordinate does not advance in the description of a timelike curve for which the physical proper time must necessarily advance.

Secondly we would question the continuation of identifying the ϕ values of 0 and 2π when ϕ becomes a timelike coordinate. The ϕ values identification is logical when ϕ is spacelike because this is our understanding of the azimuthal spatial symmetry that is our experience in nature. We understand the act of returning to the same spatial position as a common experience of perception. Moreover it is supported by our physical participation as an anchor of our very existence. However, our experience with time is that it is non-periodic. While some might argue that in spite of our lack of experience with a true time, general relativity is surprising us here with a distinctly new phenomenon. They could claim that continuity demands the identification when ϕ becomes timelike. However there is in fact a *discontinuity* in the process of the transition, the abrupt change from spacelike to timelike. Hence this continuity rationale is not at all compelling.

A very simple example illustrates how the CTC phenomenon can be produced. Consider flat spacetime in cylindrical polar coordinates

$$ds^2 = dt^2 - dr^2 - r^2 d\phi^2 - dz^2 \tag{11.8}$$

with the standard coordinate ranges and where $\phi = 0$ and $\phi = 2\pi$ are identified in the usual manner,

$$(t, r, 0, z) = (t, r, 2\pi, z). \tag{11.9}$$

We effect a coordinate transformation as follows:

$$\bar{t} = t + a\phi, \quad \bar{\phi} = \phi, \quad \bar{r} = r, \quad \bar{z} = z \tag{11.10}$$

where a is a constant, while we maintain the identification in ϕ for 0 and 2π as we did in (11.9).

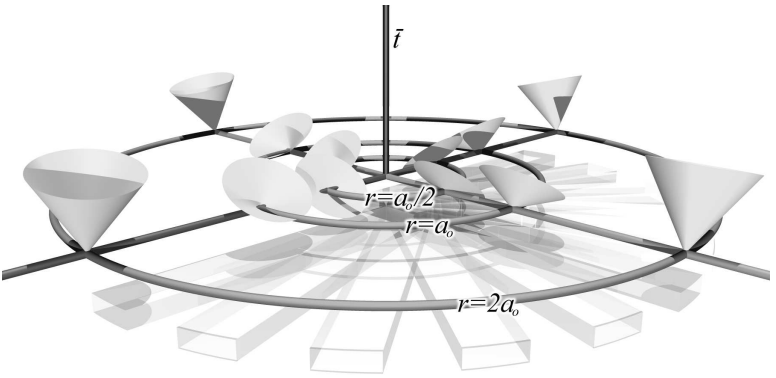

Figure 11.1: Tipping light cones produce a CTC for $r < a$ in the $(\bar{t}, \phi)$ coordinates. Boxes at the bottom follow the curves for constant $\bar{t}$.

After the transformation, the metric becomes

$$ds^2 = d\bar{t}^2 - dr^2 - 2ad\bar{t}d\phi - (r^2 - a^2)d\phi^2 - dz^2 \tag{11.11}$$

which is precisely of the type (11.1) with f, w and ν taking on constant values. If we were to follow the Gödel approach with (11.11), we would identify the events as

$$(\bar{t}, r, 0, z) = (\bar{t}, r, 2\pi, z). \tag{11.12}$$

$(\bar{t}_0, r_0, \phi, z_0)$ for $r_0^2 < a^2$. The positive sign of the $g_{\phi\phi}$ component of (11.11) when $r < a$ shows that the character of ϕ is timelike and the *imposed* closure characteristic of the ϕ coordinate, (11.12) creates the closure of the curve that advances from $\phi = 0$ to $\phi = 2\pi$. An essential point is that this identification is not equivalent to (11.9).

It is helpful to express in a diagram Figure 11.1 the transition of the curve structures as we pass from the spacelike to null to timelike curves.

The transition is seen by the light cones: when the curves are outside of the light cones at successive events, the curves are spacelike and when inside, the curves are timelike and since the curves are closed, for the latter we do indeed have CTC's for these inner curves. This situation is similar to that expressed in the CTC figure in [89] for the Gödel universe. However, the latter is cloaked in

the camouflage of exotic gravity whereas we remind ourselves that the present example is simply flat spacetime. It arose by virtue of an unusual coordinate transformation (11.10) while insisting that in the new system, the ϕ coordinate continue to be identified at 0 and 2π. It is well to reflect on the fact that this phenomenon is absent before the transformation where the only *physical* closed curves are spacelike. Since a mere manipulation of coordinates with an identification of points can produce CTC's that have such a similarity to the classical Gödel CTC's, we are led to question the legitimacy of the claim that there is any truly physical aspect to the CTC concept.

We can see the effect of the coordinate transformation in another useful manner. For this, we transform the light cones of Figure 11.1 back into the original (t, r, ϕ, z) coordinates that displayed flat space in the standard transparent form. Now the evolution of the light cones is seen in Figure 11.2.

The curves $t + a\phi = \bar{t}_0$, $r = r_0$, $z = z_0$ are seen as helices in this plot and these helices are inside the light cone for $r_0 < a$ and they are outside the light cone for $r_0 > a$. However, the paths now being helices do not close as opposed to the previous circular paths that are closed paths. Thus we have removed the CTC feature by no longer identifying the points as previously when ϕ became timelike. Now the successive periodic encounters in spatial position occur at successively *later* times as we are familiar in real life.

The same type of situation occurs in the Gödel [95] spacetime. Here the metric is expressed in the confusing form with two timelike coordinates $\bar{t}, \bar{\phi}$ everywhere as

$$ds^2 = a^2\left(d\bar{t}^2 - d\bar{r}^2 + \frac{1}{2}e^{2\bar{r}}d\bar{\phi}^2 + 2e^{\bar{r}}d\bar{t}d\bar{\phi} - d\bar{z}^2\right). \tag{11.13}$$

As discussed previously, the essential 3+1 nature of the spacetime is camouflaged. For clarity, it is necessary to display the 3+1 character explicitly which we do with the transformation

$$\begin{aligned} \bar{t} &= t + \frac{r\phi}{2}(1 - \ln r) + \frac{1}{2}\ln r \\ \bar{r} &= r\phi \\ \bar{\phi} &= -\frac{1}{2}e^{-r\phi}\ln r \\ \bar{z} &= z. \end{aligned} \tag{11.14}$$

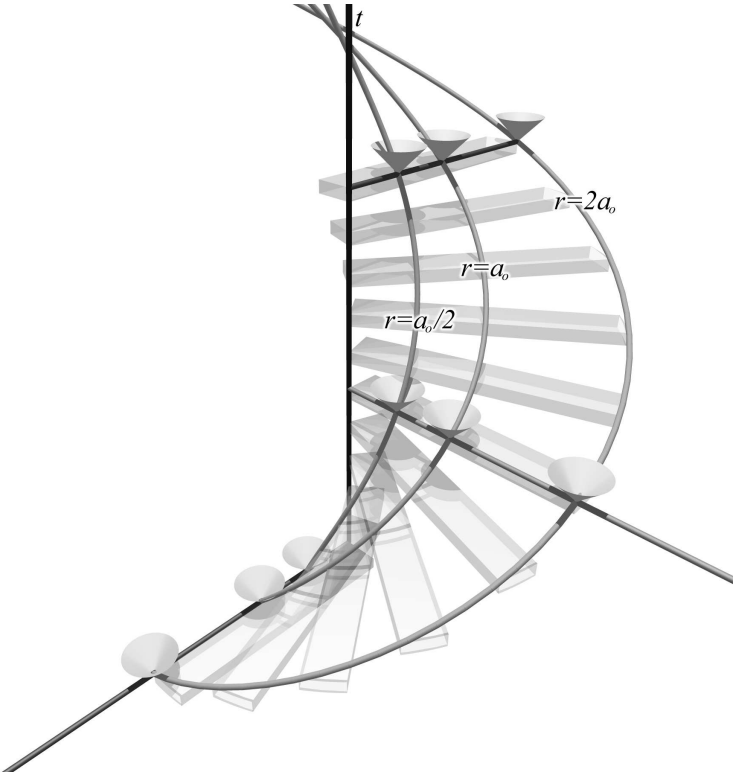

Figure 11.2: As in Figure 11.1, the boxes are used as visual aids to illustrate the evolution of the curves. By contrast with the previous figure, the boxes here are at constant t. In the (t, ϕ) coordinate system, the spacelike, null and timelike curve are seen as a unified family of curves advancing monotonically in time t. Evolving curves never close in terms of t and hence there are no CTC's with the periodic time restriction removed. Here, the fixed $\bar{t} = \bar{t}_0$ surface is actually helicoidal.

The Gödel metric now takes on the form

$$\begin{aligned}\frac{ds^2}{a^2} =& dt^2 - \left[\phi^2 + \frac{1}{8r^2}\left(r\phi \ln r - 1\right)^2\right] dr^2 \\ &- \left[\frac{3}{4}r^2 + \frac{1}{8}(r \ln r)^2\right] d\phi^2 - dz^2 \\ &- \frac{1}{4}\left(8r\phi + r\phi(\ln r)^2 - \ln r\right) dr d\phi + r dt d\phi.\end{aligned} \tag{11.15}$$

As a result of this transformation, ϕ dependence appears in the metric. With such explicit dependence, it would not be suitable to impose periodic identification in ϕ: doing so would induce a metric discontinuity which is unacceptable for a spacetime.

To view the imposed periodicity properly, we determine how the $\bar{\phi}$ in (11.13) appears in terms of its appearance in the unbarred coordinates of (11.15).

Since we have two timelike coordinates in the barred system, we can have time continuing to advance even when we hold one of the two timelike coordinates fixed. The periodic identification in the barred coordinates

$$(\bar{t}, \bar{r}, 0, \bar{z}) = (\bar{t}, \bar{r}, 2\pi, \bar{z})$$

is transformed to

$$(t, 1, \phi, z) = (t + 2\pi(1 - \phi)e^{\phi}, e^{-4\pi e^{\phi}}, \phi e^{4\pi e^{\phi}}, z).$$

We note that in this unbarred coordinate system that displays the 3+1 character of the matric, there is no longer even a suggestion of any identification of spacetime points.

We now re-examine the general metric form (11.1) and consider the transformation for the curve in the case where (r, z) are held constant. Since f, ν and w are functions only of r and z, these functions are also kept constant on this curve. As a result, the differentials transform as

$$dt = d\bar{t} - w d\varphi \tag{11.16}$$

$$d\Phi = \frac{w^2 f - r^2 f^{-1}}{2fw} d\varphi - d\bar{t}. \tag{11.17}$$

For this curve, the line element takes the form

$$\begin{aligned} ds^2 =& \frac{f}{(w^2f^2+r^2)^2}\Big((w^2f^2-r^2)^2dt^2 \\ & - 8f^2w^2r^2d\Phi dt - 4f^2w^2r^2d\Phi^2\Big). \end{aligned} \tag{11.18}$$

Note that here, t is a timelike coordinate and Φ is a spacelike coordinate regardless of whether (11.5) or (11.7) holds. This is the type of behavior with which we are familiar and comfortable. Moreover, We see that the azimuthal coordinate Φ is maintained explicitly as a truly angular coordinate throughout, unlike the case with the Gödel approach. Here, we see that there is no room for ambiguities of interpretation, making these coordinates particularly valuable.

Since we have two timelike coordinates in the barred system, we can have time continuing to advance even when we hold one of the two timelike coordinates fixed. As a particular simple choice, we let $\bar{t} = 0$. The curve equation in parametric form becomes

$$\begin{aligned} t &= -w\varphi \\ \Phi &= \frac{(w^2f - r^2f^{-1})\varphi}{2fw} \end{aligned} \tag{11.19}$$

with parameter φ. When we eliminate φ between the two equations, we see that Φ is a linear function of t with proportionality factor that depends on the particular (r, z) chosen.

Now working in the familiar system of cylindrical-like polar coordinates, we are able to connect the mathematics with our experience. We can use t as a reliably transparent timelike coordinate because it is the *sole* timelike coordinate. Similarly, we can appreciate Φ as the azimuthal angle coordinate as opposed to φ. Thus we ascribe periodicity to Φ, identifying the values 0 and 2π. On the other hand, there is no reason to ascribe periodicity to t. We *choose* to have time flow monotonically without repitition as it does in conventional flat space, as is our experience in nature. The spatial points are retraced *ad infinitum* and they do so at successively later times. They do so here as in the previous examples in Figure 11.2.

11.4 The role of our experience in nature

There is an interesting issue involved here. In the study of CTC's, our experience in nature has already been imposed prior to any analysis. We demand that curves always evolve into the *forward* light cone, that time for a physical observer is monotonic in its evolution. As well, we recognize that once the forward direction of time is set, we demand that it always continues to flow in that direction, *in conformity with our experience*. To be noted is that if we were to allow a reversal in the flow of time, CTC's would be trivially created. Since we have already injected rules based upon our experience, in our view there is no logic in accepting periodicity in the time coordinate for a physical solution. Our experience in nature is that while we readily re-visit spatial locations, we do not re-visit points in time.

The essential confusion that arises with the Gödel spacetime is the notion that an angular coordinate, once set logically to have a certain periodicity when spacelike should by necessity maintain that periodicity when it becomes timelike. While it is certainly a *choice* that can be made, there is nothing that makes it a necessity. Indeed, to adopt that choice is to force the realization of a CTC in a particular spacetime. However, the choice we would argue as more natural, is *not* to force the periodicity and this does not yield a CTC in this different spacetime. It is important to note that there is nothing in the field equations to guide us in one direction as opposed to the other. We would argue that the choice of a system of coordinates in which there are two timelike coordinates with one held fixed for a timelike curve is an unfortunate choice from the point of view of clarity. Our choice is to select coordinates that maintain their timelike or spacelike character. In so doing, we recognize that the imposition of periodicity is a choice rather than a necessity.[a]

Returning to the issue of whether or not the CTC's of Gödel are geodesic, we would argue that the authors in [87] were justified in raising this point. To be a geodesic curve is to be traceable *without*

[a]This also applies to the spacetime referred to in section 8 of [96]. Having the coordinates run from $-\infty$ to $+\infty$ does not alter the fact that to have a CTC of interest, one must return to the same spacetime point after a journey in the forward light cone. For closure, the points with $\phi = 0, 2\pi, 4\pi$ etc. are identified. Again, this is a choice and not a necessity.

any extraneous elements, i.e. to "fall freely". Had it been the case that the Gödel CTC's were geodesic, then one might have argued that the deviation from our normal experience would not be so radical. However, once it is seen that they are not geodesics, there is the immediate requirement for an agency to force the particular spacetime trajectory, i.e. a mechanism of sorts to provide the required non-gravitational force, which could be labeled a "time machine". This is more serious in that to recreate the conditions for total closure of the system, the elements of the time machine must also follow closure in time. In so doing, there will be a reversal in entropy flow that further compounds the demands upon one's credulity, since the Second Law of Thermodynamics is an additional strong element of our experience in nature.

It is our contention that the essence of CTC creation stems from the process of identification of spacetime points, that the CTC is a mathematical man-made choice rather than the result of general relativity and the solutions of the field equations.

11.5 Gott's moving cosmic strings

This applies to even the modern versions such as the example put forward by J. R. Gott [92]. This case was examined by Tieu as first discussed in [97]. In the Gott system, a CTC is constructed using two moving cosmic strings and the mathematics of Lorentz boosts. A pair of simple examples will illustrate its main feature and how the existence of such CTC's arises.

First, consider a flat 1+1 spacetime in which a strip $-1 < \bar{x} < 1$ is removed, i.e. the points identified are $(\bar{t}, -1)$ and $(\bar{t}, 1)$. For this example, we shall call the region $\bar{x} < -1$ the negative side and the $\bar{x} > 1$ the positive side. If one applies a Lorentz boost from $(\bar{t}, \bar{x})$ to (t, x) *before* identifying the points, the two edges of the cut as seen in the new coordinate system will "slip" as shown in Figure 11.3.

It is seen in Figure 11.4 that any object from the positive side, crossing the identified strip to the negative side will be displaced back in t value. In a sense, making the transition over the identified points e_1 and e_2 in this direction allows travel into the past. In spite of this, causality is not violated because any attempts to close the object's

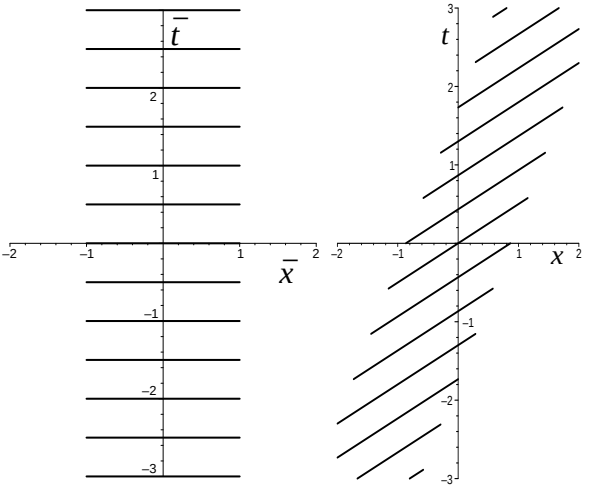

Figure 11.3: In the left figure, the identifications after the removal of the strip $-1 < \bar{x} < 1$ are shown using horizontal line-segments. The right figure illustrates the same identification of points after the Lorentz boost.

world line would require another crossing through the identified strip. However, traveling through this strip in the opposite direction would have the t value *increased* by the same amount. Thus, no events in the future of e_2 coincide with e_1, i.e. it is impossible to return to the initial event via a timelike trajectory.

On the other hand, if the identification were made before the Lorentz boost was applied, the spacetime would be continuous. All events would be mapped smoothly from one coordinate to another without any jump in "time". This will be the key feature to be employed in the next example. The choice of appropriate order of identification will be discussed later.

Next, we consider a 2+1 system $(\bar{t}, \bar{x}, \bar{y})$, as opposed to the 1+1 dimensional system in the previous example. Instead of having one strip $-1 < \bar{x} < 1$, $-\infty < \bar{t} < \infty$ removed for $\bar{y} = 0$, we will consider two strips removed: the "front" strip being $-1 < \bar{x} < 1$, $-\infty < \bar{t} < \infty$, $\bar{y} = y_1$ where y_1 is a positive constant and the "back" strip being $-1 < \bar{x} < 1$, $-\infty < \bar{t} < \infty$, $\bar{y} = -y_1$. Consider a Lorentz boost with velocity $+\beta_s$ in the positive $\bar{x}$-direction for half of the

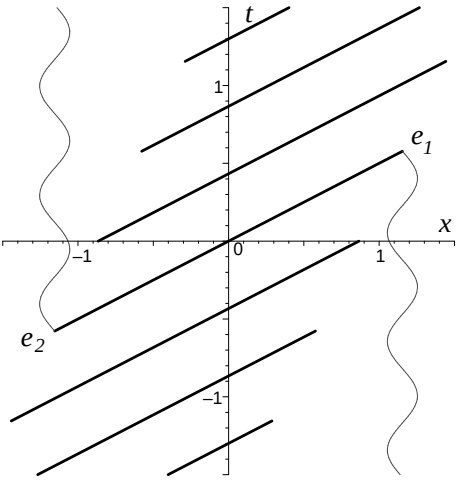

Figure 11.4: This shows a possible worldline of a massive object as it crosses the identified strip. Events e_1 and e_2 are identified. In this coordinate, the t-value of e_2 is less than that of e_1 so one can say that e_2 occurred before e_1.

space $\bar{y} \equiv y \geq 0$ and another boost in the negative $\bar{x}$-direction for the other half, $y < 0$. Following this, we stitch the two half-planes together. This is possible because the two half-spaces are flat and hence, they can be stitched together [92].

That there is a violation of causality can be seen as follows: A traveller starts at an event E_1 on the right side of the front strip (with the front being $y = y_1$) and crosses over the $+\beta_s$ Lorentz-boosted strip to travel "back in time" to event E_2 as seen in Figure 11.5. With y_1 sufficiently small, he could proceed via a timelike path to the back strip ($y = -y_1$) at event E_3. At this point, he crosses over the $-\beta_s$ Lorentz-boosted strip to event E_4. From E_4, he could follow a timelike trajectory to return to his original position in space and time at event E_1. Thus, his worldline is a CTC. This illustration captures the essential mechanism of the Gott-produced CTC [92].

The primary difference between this scenerio and that of Gott is in the choice of coordinate system. Gott chose a coordinate system where each cosmic string is at rest (in the barred coordinates) and thus E_1, E_2, E_3, E_4 and all intermediate events are simultaneous in

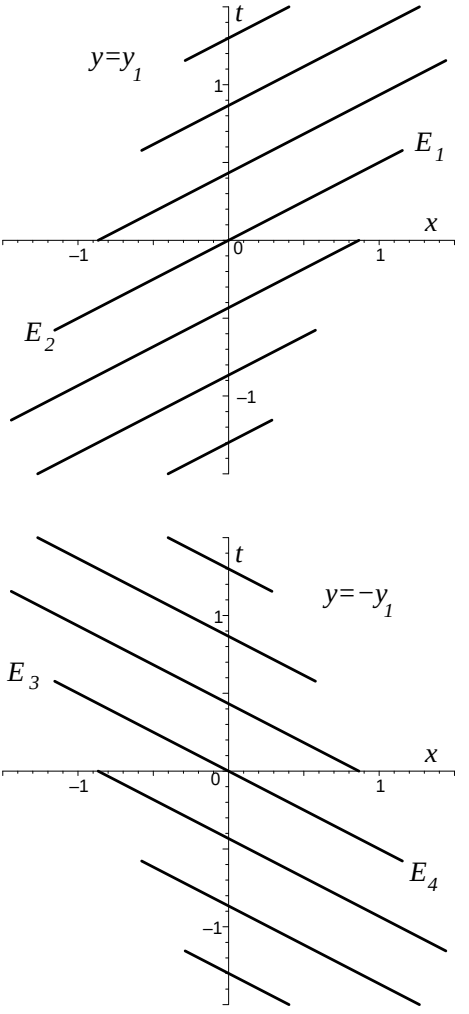

Figure 11.5: These are two Lorentz-boosted strips in opposite directions: the front being $y = y_1$ and the back being $y = -y_1$. Points E_1 and E_2 are identified as with points E_3 and E_4.

that system. Either system of coordinates is acceptable.

The presence or absence of a CTC rests in identifying respectively the points *after* or *before* the Lorentz boost is applied. Gott chose the first form of identification and hence he realized a CTC. The question arises as to which approach is the more natural one in dealing with such a system. Regardless of the viability or non-viability of cosmic strings, one would expect continuity and axial symmetry of the spacetime around one cosmic string at rest, i.e. the "wedge" that is removed in the construction of a cosmic string from non-singular flat spacetime should not be detectable. This is just as in the case of an axially symmetric cone which does not display the wedge that one would create to paste it together into a cone from the original plane. As a string is Lorentz-boosted, this continuity should be maintained even though the axial symmetry is lost. On the other hand, if there *were* a discontinuity, there would be no preferred location for it to appear because of the ultimate axial symmetry. Thus, the more reasonable scenerio from the point of view of physics (at least to the extent that one is inclined to regard these constructs as physical) is in the identification of points *before* the Lorentz boost. In so doing, one rules out the closed timelike curve as envisaged by Gott.

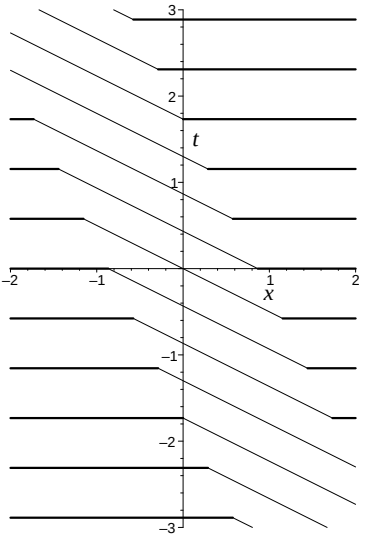

Figure 11.6: The lines illustrate the identification of points along the Lorentz-boosted strip.

This page intentionally left blank

Chapter 12

The Direction of Physics Research

In a very interesting recent book *The Trouble With Physics* [98], the author, L. Smolin provides his views on the misdirection of current physics research. His delineation of the sociology of theoretical physics research is astute and insightful, delivered with great panache. After a particularly scathing assessment of the string theory community and what is in his view a rather limited set of achievements, he leans towards his own favored directions in quantum gravity, decrying the dearth of revolutionary fervor in the physics community. Ironically, from our perspective, what is particularly valuable about this work is that it encapsulates the commonality of mind sets of what has become the dominant string theory group "A" and the supposed independents "B".

What A and B share is a preoccupation, perhaps "obsession" would be a more apt description, with "unification", which has become the mantra of theoretical physics research. In [98], that author recalls the usually described progression, how Maxwell unified electricity with magnetism, how S. Weinberg, A. Salam and S. L. Glashow unified the weak and electromagnetic interactions and how gauge theory brought further unification into all but gravity, the lone hold-out in the dream of ultimate unification.

Taken for granted is that gravity must fall in line with the rest of physics as a quantum phenomenon with its spin-two graviton quantum and that quantum gravity is where physics theory is to be

focused.

While we have taken a brief excursion (Chapter 7) into the Planck scale where any quantum effects from gravity would be expected to show up if at all, we stress that there are very solid reasons to resist the pouring of all efforts into quantum gravity. Some efforts are laudable but balance, moderation and perspective have been lost. First and foremost is that the experimental underpinnings for theoretical research in some of the leading-edge areas is largely absent, and in the case of quantum gravity, totally absent. Physics drifts towards metaphysics when the experimental support becomes more and more remote. Particularly distressing is the human factor: with no objective standards for success that experiments provide, the politics of physicists takes over, a point well-documented in [98]. With human beings there will always be some politics but with no objective guidelines, politics can become tyranny.

At the turn of the previous century, there was adequate experimental evidence that the existing theoretical tools in physics were lacking. The remarkable edifice of Maxwell's classical electromagnetism was exhibiting serious deficiencies:

- 1. There was the photoelectric effect, the abrupt onset of electron emission from a metal surface only when light of a particular frequency was shone upon it. If the frequency was too low, there were no electrons emitted, regardless of the intensity of the light.

- 2. There was the "ultra-violet catastrophe" in black-body radiation emission: rather than increase monotonically, the spectrum for the intensity of the radiation fell sharply towards the ultra-violet end of the spectrum.

- 3. There was the mysterious behavior of electromagnetic radiation from atoms: accelerated charges did not emit a continuum of emissions at the atomic level but rather emitted at discrete frequencies. This quantification of emissions had no place in classical electromagnetic theory. (It is fortunate for us that classical electromagnetism breaks down at the atomic level as otherwise, the electrons that constitute the atoms of our bodies would radiate continuously as they spiraled into our nuclei,

and our structure would be severely compromised.)[a]

There was also the observation of "zitterbewegung", the jerky motions of suspended granules in a fluid. These *observed* phenomena spurred Planck, Bohr and Einstein towards their remarkable achievements. By contrast, what failures of existing theory do we have today to tell us experimentally that there is something wrong with the classical picture of gravity: a *continuous* curved spacetime provided by Einstein's original general theory of relativity? None of which we are aware. It is interesting to contemplate the influence that Einstein's two "mistakes" have had on contemporary physics thinking, the Principle of Equivalence and the injection of the cosmological constant into the field equations. One might raise one's eyebrows when reading of the Principle of Equivalence as a mistake. Nevertheless, a mistake it is, at least in its most common form. Simply put, an accelerated reference system is *not* equivalent to a gravitational field. True gravitational fields such as those produced by real bodies cause the paths of neighboring free-fall particles to converge or diverge in particular manners depending upon the orientation whereas accelerated reference frames cannot exact the same effects. While this is fairly widely understood, nevertheless this basic fact continues to be ignored or minimized in its importance. As a result, a certain misconception commonly enters current thinking to the effect that Einstein had "unified" acceleration with gravity (as presented for example, in [98]). If the Equivalence Principle really were precisely true, there would have been a basis for this claim. But it is only an approximate truth, gravity is *locally* equivalent to an accelerated reference frame. One does not eliminate gravity by free fall.

The essence of the phenomenon that we call "gravity" arises in general relativity as the curvature of spacetime. This curvature reveals itself in an invariant manner through the non-vanishing of the Riemann tensor. One commonly encounters the misconception that as a result of the Equivalence Principle, one has produced true gravity by proceeding to an accelerated reference frame while still actu-

[a]Actually the Pauli exclusion principle also plays an assisting role in maintaining our integrity, preventing the electrons from piling themselves together into the ground state innermost shell, even given the quantification of atomic energy levels.

ally being in flat spacetime where the curvature is zero. However, the Riemann tensor is zero in flat spacetime, curvature is absent in flat spacetime, whether it is calculated in the standard inertial frame or in the accelerated frame. Unappreciated by many is that special relativity, Einstein's theory of space and time in the absence of gravity, can be formulated in accelerated reference frames, albeit not trivially, instead of in the usual inertial reference frames (see, for example, [4]). It has been argued that the Principle of Equivalence should be seen as the guide along the way for Einstein to develop his general theory of relativity rather than as a true principle of physics. Synge has been a very strong advocate for this stance [17].

In fact, we would argue that Einstein's work should have been seen more rightly as moving gravity in the opposite direction from unification. Consider Newtonian gravitation. There is a force

$$\frac{GmM}{r^2}$$

between two masses M and m separated by a distance r just as there is a force

$$\frac{e_1 e_2}{r^2}$$

between two charges e_1 and e_2 separated by a distance r. This in itself could be viewed as a kind of unification. However, general relativity removes gravity from the category of force and repositions it as the curvature of spacetime. All particles and fields in nature apart from gravity exist *within* spacetime but gravity *is* spacetime. We developed the analogy in Chapter 6, of the particles and fields in nature as actors on the stage whereas gravity is the stage itself, a fundamental disunity of concepts. In fact the analogy is even better: we can (and should) view the stage as flexible (recall from Chapter 2 that there are no truly rigid bodies in nature by special relativity), so that the heavier actors bend the stage more than the lighter actors. Thus, more massive bodies produce more curvature of the stage and by analogy, stronger gravity, than do the lighter bodies. Our proposal: Einstein's theory with gravity incorporated into the curvature of spacetime on the left side of his field equations and with all other fields and matter embodied on the right hand side in the form of the energy-momentum tensor, be seen more accurately as the *dis*unification of nature.

It would seem reasonable that in the absence of experimental phenomena to have us think otherwise and believe that gravity must be quantized, that this counter way of thinking is actually the conservative stance. While some might view such a counter-position as radical, it is in fact more in harmony with the essence of general relativity.

A key indictment of string theory that is provided in [98] concerns "background dependence", that the spacetime in which string theory exists does not evolve but rather is a fixed structure in which the strings live. That author's complaint derives from the dynamical nature of spacetime in general relativity, rendering preordained spacetime theory regressive. For him, any legitimate theory must be "background independent", that spacetime itself must "emerge" from the theory. We regard this stance as problematic for a variety of reasons:

1. In general, the theory of general relativity describes how spacetime *curvature* gets created and modified, usually in response to dynamical material sources.[b] The theory does not lead to the creation of spacetime itself except for some very special circumstances such as emerging Friedmann universes from the big-bang singularity. Even there, the issue is not straightforward because an essential singularity comes into play. Once the Friedmann universe emerges from the singularity, the spacetime is already present. It is just that in the case of the "closed" universe, it is minute in size and grows from that emergence. To reject a theory of elementary particles because it does not embody such a special feature is, in our view, rather unreasonable.

2. Whether or not one finds fault with string theory, the quest for an improved theory of elementary particles is surely a laudable endeavor. Even if it should finally turn out to be the case that an ultimate theory will actually be able to describe both the emergence of spacetime and the particles within it (a circumstance that we

[b]There are exceptions. For example, there is a purely vacuum (i.e. source-free) gravitational wave spacetime of Bondi, F. A. E. Pirani and I. Robinson [99] with non-zero curvature.

find dubious at best), to reject less ambitious theorizing *a priori* on the basis that it does not do both is not, in our view, conducive to scientific progress.

3. Indeed, with rare exceptions, the history of scientific progress is one of useful incremental steps, each presenting helpful, at times valuable, levels of understanding. As an example, what we now regard as a very naive theory of gravity, Newton's theory does a splendid job of describing planetary motion to a very satisfactory level of accuracy for most human needs. General relativity adds a refining step, treating the planets as particles moving on the geodesics of the *fixed* spherically symmetric Schwarzschild spacetime created by the Sun. As we have discussed earlier, this provides an excellent explanation for the residual precessional motion of the orbit of Mercury.

4. From our own experience with the modeling of particles as solitons [38], the effects of general relativity had significance only for situations of charged particles in which the charge to mass ratio e/m of the soliton was approximately 1 in units where $G = c = 1$. However, in these units, the particles in nature have an e/m ratio of the order 10^{22}, a far cry from 1. If this is any guide, it would appear that general relativity does not play a discernible role in the construction of elementary particles in nature.

As another example of theorizing gone astray, we point to another issue discussed in [98]. This concerns the well-understood phenomenon of Lorentz contraction that we studied in Chapter 2, now applied to elementary particles. The concern was expressed that a particle would be seen to have its size approach zero as the speed of an observer viewing the particle approaches the speed of light. Since zero size for physical particles is a troubling aspect to many (including ourselves), it was proposed that just as there is a maximum velocity c in nature, there is also a minimum size, with the Planck length presented as the logical candidate.

In our view, this argumentation is symptomatic of an essential misdirection in thinking about relativity. Here, what is unappreciated is that in relativity, size is a matter of perception and as such,

it loses the kind of invariant significance that it had in pre-relativity physics. It is only the *proper* length, the length read in the rest frame of the object, that has particular significance and this is a maximum. It bears no intrinsic physical concern for us that the perceived size shrinks toward zero with increasing velocity. (In fact even the perception would be out of reach well before the calculated length would be at the Planck scale.) Perception, while certainly of interest, and in some cases very important (such as in measurements of stellar and galactic velocities) is not intrinsic. The lesson of relativity is that perception is relative. What would be of concern in the case of particles would be a *proper* length that was zero. In fact this concern was part of what led us to pursue the Einstein/Rosen program of soliton modeling for elementary particles.

In an interesting philosophical book of theorizing about physics, Weinberg frames the nature of the quest in the very title, *Dreams of a Final Theory*[100].

Steven Weinberg (1933–) shared the Nobel Prize with A. Salam and S. L. Glashow for their important contributions to the unification of the weak and electromagnetic interactions. All three of these outstanding theoretical physicists continued with their important studies in various areas of physics.

He writes " ...already in today's theories we think we are beginning to catch glimpses of the outlines of a final theory." Physicists can become so enmeshed with their work to the point of having their vision over-extended. History has taught us that earlier pronouncements along these lines have proved futile and it is most rational in our view to believe that the chase to an ultimate theory is a chase to infinity. Indeed even if physics were to actually achieve that final theory, mere mortals could never prove that they have captured this ultimate truth. The essential point is that the glory and the joy are in the journey itself. Perhaps Robert Browning said it best [120]: "*Ah, but a man's reach should exceed his grasp, or what's a heaven for?*"

This page intentionally left blank

Chapter 13

Summary with Concluding Commentary

We focused on various goals in this book. We wished to present the reader with a simple introduction to the basics of special and general relativity, theories that have changed our view of the physical world in the most profound ways. As for special relativity developed in Chapter 2, the very solidly supported Einstein theory of spacetime exempting gravity that reveals its correctness on a daily basis in laboratories throughout the world, our presentation had two foci. There was the standard approach of most texts but the best thus far in our opinion, the classical treatise of Landau and Lifshitz [3], is beyond the level of probably most general scientifically inclined readers. It is also somewhat terse in its presentation. The aim was to simplify this kind of treatment and expand upon parts that require more explanation.

Then there was the Bondi approach [4] in Chapter 3, a wonderfully intuitive treatment that could be understood fully by junior high school students and which is not nearly as well-known as it should be. We expanded upon Bondi's development of special relativity, most notably in showing how the so-called twin paradox is resolved within the Bondi framework.

It must be stressed that we focused upon the essentials. The reader is encouraged to pursue more advanced treatments for other topics in special relativity.

In Chapter 4, we turned our attention to the basics of general

relativity. We built the mathematical foundation for the theory in the form of tensor calculus. The Principle of Equivalence connecting gravity with accelerated reference frames was discussed as the guide for Einstein towards his new theory of gravity, general relativity. It was emphasized that the Equivalence Principle is frequently misrepresented and that gravity is only locally equivalent to an accelerated reference frame, that the essence of gravity lies in the curvature of spacetime whereas the mere act of accelerating does not create spacetime curvature. The energy-momentum conservation equations of special relativity, generalized to a form applicable to arbitrary coordinates, and thence to include gravitational fields, were used to build the Einstein field equations, the basic equations of general relativity. The dynamical equation for free bodies in general relativity, the geodesic equation, was shown to follow as a generalization of free motion in special relativity for arbitrary coordinate systems. The general relativistic equations for electromagnetism, the Einstein–Maxwell equations and the force law for a charged particle followed from the pattern of general covariance, the expression of equations in tensor form, applicable to arbitrary coordinate systems.

While there are many known solutions of the Einstein field equations, the number of solutions of direct physical relevance is rather limited. Probably of most importance is the very simple solution representing the spacetime metric exterior to a spherically symmetric ball of matter, the Schwarzschild solution. We used this example in Chapter 5 to illustrate the aspects of distance and time measurements peculiar to general relativity, aspects of particular importance in what followed later in Chapter 10. These concerned the finite time that is perceived by a local observer to record a particle reaching the surface at $r = 2m$ in contrast to the infinite amount of time that a distant observer would attribute for this particle to reach this radius. With the Schwarzschild spacetime, we saw the nature of phenomena peculiar to strong gravitational fields: an event horizon (the surface at $r = 2m$ in the Schwarzschild case), a black hole (the region within the event horizon) and a spacetime curvature singularity (where the matter reaches infinite density), in this case at $r = 0$. We described in brief, the tests of general relativity, classical and modern.

Just as there are electromagnetic waves as described by Maxwell theory, general relativity predicts the existence of gravitational waves.

In Chapter 6, we described how these waves arise in the theory and how they are necessary to accord with the basic premise of special relativity, the finite velocity for the propagation of information. We displayed the field equations in the approximate linearized form of general relativity where nonlinearities are discarded as is generally appropriate for the case of weak gravitational fields.

While the standard assumption is that gravitational waves carry energy and a formalism has been developed using energy-momentum pseudotensors to provide a measure for this energy, we presented a contrary view: while gravitational waves exist in nature, they do not carry energy in vacuum. In addition to the less-than-satisfactory non-tensorial aspect of the existing standard formalism, we presented reasons in support of our idea. The proposal stems from our simplifying hypothesis that energy, including the contribution from gravity, is localized in non-vanishing regions of the energy-momentum tensor, T^{ik}.

Energy localization has always been a problematic and contentious issue in general relativity. In our view, our hypothesis addresses the problems regarding energy in general relativity but it must be said that our hypothesis is seen by most researchers as highly controversial. The idea of accepting the concept of waves that do not convey energy is understandably a large leap for most to fathom. The hypothesis also does not sit well with particle physicists who view gravity as just another field and the graviton as just another particle. This underlines the divide that exists between most particle physicists and general relativists. By and large, most general relativists regard gravity as the manifestation of spacetime curvature, i.e. an inherent geometrical property of spacetime itself rather than a field such as the electromagnetic field that resides within spacetime. If our hypothesis should prove to be correct, the existence of a graviton, the quantum of the gravitational field, is brought into question.

In Chapter 7, we surveyed the scales of dimension in nature and what could turn out to be the smallest physical scale, the Planck scale. We noted the various interactions in nature that relate to the various scales. The strongest interaction, aptly named the strong interaction, binds the nucleons, the neutrons and protons, in the atomic nuclei. It is of short range as the so-called weak interaction that mediates decay processes such as the decay of a neutron into a proton,

an electron and an anti-neutrino. Intermediate in strength between these two interactions is the electromagnetic interaction which is of long range. It is directly related to our everyday experience and has significance over the entire range of scales. Finally there is the weakest interaction, the gravitational interaction, which is of long range and of the prime focus in this book. Placing gravity within the lexicon of interactions could carry with it the implication that gravity does not have a particularly special aspect, fundamentally separating it from the other interactions in its role as the embracer of all particles and fields. This is unfortunate as it masks the fundamental split in the views concerning the essence of gravity.

We discussed the nature of phenomena at the various scales, from the largest cosmological scale down to the tiny scales of modern particle physics. At that point, we introduced the Planck scale, first developed by forming a combination of the fundamental constants c, $\hbar$ and G into one of the dimension of length. This turned out to be of the order 10^{-35} m, a size so incredibly removed by some 20 orders of magnitude from that of the already minute elementary particles. We also noted that the Planck scale could more usefully be derived by linking the characteristic size of a body at which point its gravity becomes very strong, namely its Schwarzschild radius $2Gm/c^2$, to its Compton wavelength, $\hbar/mc$ which characterizes its quantum mechanical aspect. From the Planck length, there followed the Planck mass and the Planck time. We noted that this standard approach ignores the spin and charge of a particle, fundamental quantized aspects of matter. To remedy this neglect, we replaced the Schwarzschild radius by that which arises for a particle with both spin and charge. This was quantized in the elementary approach reminiscent of Bohr's first quantum mechanical model of the hydrogen atom. As a result, a spectrum of Planck states emerged bringing in an effective fine-structure constant α. Remarkably, the spectrum limit went hand-in-hand with an α value of $1/128$ which is extremely close to the α value governing high energy radiation in Z-boson production and decay. Whether this is a mere coincidence or whether this concordance might have deeper significance remains to be seen.

In Chapter 8, we left the realm of the smallest of dimensions and turned to the largest, the scale of the astronomical elements. In preparation, we used a series of scaling models to assist in the

visualization of the relationship between the sizes and separation distances between the elements. To visualize the solar system, we considered the Sun reduced to the size of a pin head and found that the planets on this scale were microscopic specks located many meters away. The nearest star was kilometers distant by this measure. This scaling procedure gave us a sense of the vast empty space between the elements. However, by using a similar scaling model with galaxies, we also had a sense of the greater relative density of clusters of galaxies.

We touched briefly upon the early historical ideas about cosmology and turned to general relativity to guide us toward the modern era of cosmological research. The key transition due to Hubble came from viewing the universe as a static structure to one of dynamic expansion. Friedmann's cosmological solutions of the Einstein equations provided the theoretical framework for this transition. We discussed the three types of Friedmann models in terms of their curvature characteristics. We then considered the modification to the field equations that Einstein had initially imposed in order to maintain the then commonly believed static character of the universe, the addition of the cosmological term Λ. Upon learning from Hubble of the expansion of the universe, Einstein reportedly declared the addition of the Λ term his greatest blunder. We mentioned how this term has been resurrected in modern research to account for what is now generally believed to be the acceleration in the expansion of the universe. While many researchers regard the presence of the Λ term in the field equations as a natural geometrical element and that the notion of it being absent amounts to highly improbable fine-tuning, we took a different approach. We showed that it was more logical to place the Λ term on the right side of the field equations as a form of exotic energy-momentum rather than on the left side with the Ricci tensor and Ricci scalar. This removed the fine-tuning argument and provided a more pedestrian perspective on the cosmological term.

In more recent times, we see the pendulum shifting with researchers regarding the term as belonging on the right hand side of the field equations as a kind of strange matter. It has been named "dark energy" with the property of repulsion providing an acceleration to the expanding universe.

We then turned to the motions of the stars in the majestic spiral galaxies in Chapter 9. We noted that the rotational velocities of the

stars did not fall off with distance from the axis of rotation but rather tended to remain essentially constant as they were tracked in a line perpendicular to the axis. These velocities, plotted on graphs, are referred to as "rotation curves". It was determined that the general flatness of these rotation curves could only be accounted for within the confines of Newtonian gravitational theory if there were vast amounts of extra unseen matter distributed in spherical halos around the visible galactic contents. This unseen matter, displaying its presence solely by its apparent gravitational effect, was designated "dark matter". Some have opined that the name itself is a euphemism for ignorance while a large body of the physics community has embraced dark matter with considerable enthusiasm. Regardless, the question of the existence and nature of dark matter has become one of the most important issues in contemporary physics.

Some have attempted to explain the phenomenon of the large velocities in the spiral galaxies by an *ad hoc* modification of the Newtonian law of attraction and others by the addition of new fields. We reasoned that since general relativity is the premier theory of gravity, an attempt should first be made to analyze the apparently anomalous velocity problem within Einstein's theory. It is understandable that such an attempt had never been made before because of the pervasive bias that Newtonian gravity should always suffice where the gravitational fields are weak and the velocities are non-relativistic. However, we noted that the nonlinearities of general relativity can lead to unexpected results even when fields are weak in the case where the velocities are driven by gravity itself.

We found this to be true for a stationary axially symmetric rotating collection of pressure-free fluid (i.e. dust) as an idealized model for a rotating spiral galaxy. Using the solutions of Einstein's equations, we found that we were able to model a variety of the known galactic rotation curves with considerable accuracy without invoking the massive halos of dark matter. This work was met with both enthusiasm by those who doubted the existence of dark matter from the outset and by criticism from a variety of researchers. In Appendix A, we related the essentials of the various critical arguments and we provided our counter-arguments to each critic.

An important point should be emphasized: while Newtonian gravity is a simple theory that demands the large quantities of unseen

matter to realize the large stellar velocities as seen in flat rotation curves, Einsteinian gravity is far richer. General relativity can accommodate these rotation curves without dark matter. However, suppose there actually were large halos of dark matter surrounding the visible matter in galaxies. In that case, the question arises as to how general relativity would discriminate between the presence and absence of the extra matter. We found the answer in terms of the degree in dispersion of the rotation curves, the extent to which rotation curves would vary as plotted in successive parallel planes above and below the galactic symmetry plane. We noted that only very scant dispersion data are presently available. Hopefully in time we will have the data to follow this interesting avenue of investigation.

Because of the criticism that we received in our galactic modeling, we sought another source of confirmation that general relativity could accommodate the larger-than-expected astronomical velocities without the dark matter demanded by Newtonian gravity. For this, we turned in Chapter 10 to an idealized spherically symmetric model of a cluster of galaxies. A system of pressure-free particles was investigated in the process of collapse at the stage when the velocities were still small and the gravitational field was still weak throughout. It is well known that the perception of velocities of such elements will differ for distant observers as compared to those who are adjacent to the falling particles. We provided this analysis using the known exact solutions for this type of system. The results were interesting: we found that we could model an idealized Coma cluster of galaxies that would reveal the observed velocities for us as distant observers without dark matter. The beauty of this system is that there are no singularities present at the weak-field non-relativistic velocity stage.

We then considered the various items of claimed evidence for the presence of dark matter, indicating the reasons for our reservations. In the times to come as in the past, new findings will point the way in either direction and the true picture will eventually emerge.

While in the previous chapters we indicated how general relativity could be used to describe gravitational phenomena of interest, in Chapter 11 we turned to a subject in which general relativity has not been used constructively for physics. This concerned closed timelike curves, paths in spacetime for which an observer could travel into his or her past. The idea of such curves was greeted with con-

siderable hoopla by many as it represents a potential realization in hard science of the phenomenon that has appeared in science fiction over many years, namely the time machine. In various examples, we showed that the constructed closed timelike curves actually resulted from a mathematical choice rather than a physical necessity. The choice concerned the identification of spacetime points. The physical choice that was available was the one that respected the inexorable continued flow of time.

In Chapter 12, we discussed the direction of physics research. We focused upon some of the key assumptions that are leading the direction of current research and posed some alternative approaches.

Appendix A

Critical Challenges and Our Replies

An issue first raised privately to us by some colleagues and later in [101], [102] concerns the nature of the matter distribution in our galactic model. They have noted that given the existence of the discontinuity of the function N_z that we had pointed to in [63], a significant surface tensor S_i^k can be constructed with a surface density component given by

$$(8\pi G/c^2)S_t^t = \frac{N[N_z]}{2r^2} - \frac{[\nu_z]}{2} \tag{A.1}$$

to first order in G, i.e. G^1. The square bracket notation [..] denotes the jump over a discontinuity of the given function, here at $z = 0$. Using (9.7), this becomes

$$(8\pi G/c^2)S_t^t = \frac{N[N_z]}{2r^2} + \frac{N_r[N_z]}{2r}. \tag{A.2}$$

It was claimed that this necessarily implied the existence of a singular *physical* surface of mass in the galactic plane above and beyond the continuous mass distribution that we had found, thus rendering our model unphysical.

Having received this challenge, we calculated the surface mass that was said to be present in the four galaxies that we had studied by integrating (A.2) over the surface without paying heed to the actual sign of the result. Suspicions were aroused from the discovery

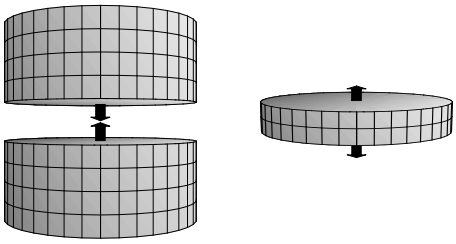

Figure A.1: Normal vectors used to calculate flux.

that (A.2) in each case gave a numerical value slightly less than the mass that we had derived from the volume integral of our entire *continuous* mass density distribution using (9.18), (9.15) and (9.12).[a] With echoes from undergraduate mathematics courses, this pointed to a plausible explanation: in our case, *with our choice of model*, there is no *physical* mass layer present on the $z = 0$ plane. *The surface integral of this singular layer is merely a mathematical construct that indirectly describes most of the continuously distributed mass by means of the Gauss divergence theorem.* To see this, consider the vector $\mathbf{F}$ defined as[b]

$$\mathbf{F} \equiv A(r, z)\mathbf{e}_r + B(r, z)\mathbf{e}_z \tag{A.3}$$

where

$$(8\pi G/c^2)B \equiv \frac{NN_z}{2r^2} + \frac{N_r N_z}{2r} \tag{A.4}$$

as a first option. We choose $A(r, z)$ so that

$$\int \nabla \cdot \mathbf{F} dV \equiv (8\pi G/c^2)M \tag{A.5}$$

where M is the total mass. As a more transparent second option, we choose

$$(8\pi G/c^2)B \equiv \frac{NN_z}{r^2} \tag{A.6}$$

where we define

$$\nabla \cdot \mathbf{F} \equiv (8\pi G/c^2)\rho. \tag{A.7}$$

[a] It should be noted that the two terms in (A.2) were found to contribute equally.

[b] $\mathbf{e}_r$ and $\mathbf{e}_z$ are unit vectors in the r and z directions.

From these definitions, we deduce the form of $A(r, z)$ in order to produce the density as expressed through N in (9.12). We calculate the mass over the cylindrical volume defined by $-\infty < z < \infty$, $0 < r < r_{galaxy}$. By the Gauss divergence theorem, the volume integral of ρ, via (A.7) is equal to the integral of the normal component of $\mathbf{F}$ over the bounding surfaces. However, the integration must be over a continuous domain and since the $\mathbf{e}_z$ component is discontinuous over the $z = 0$ plane, the volume integral must be split into an upper and a lower half (see Figure A.1). The two new surface integrals together would constitute the jump integral of (A.2) in the first option if one were to be cavalier about the directions of unit *outward* normals, as we shall discuss in what follows. The surfaces above and below the galaxy give zero because of the exponential factors in z and the final small contribution comes from the cylinder wall via the A function.

In our solution, the actual *physical* distribution of mass is not in concentrated layers over bounding surfaces: the Gauss theorem gives the value of the *distributed* mass via equivalent purely mathematical surface constructs as we are familiar from elementary applications of this theorem. Physically, the density is well defined and continuous throughout, except on the $z = 0$ plane. In fact the limits as $z = 0$ is approached give the same finite values from above and below. While the field equations break down at $z = 0$, the density for a physically viable model is logically defined by this limit at $z = 0$. However, with the chosen form of solution, the density *gradient* in the z direction is discontinuous on the $z = 0$ plane. This gradient undergoes a reversal for a galactic distribution with diminishing density in both directions away from the symmetry plane. It is most convenient to achieve this with an abrupt reversal as we have done. There is no indication that this choice alters the essential physics.

Thus we have shown via the Gauss divergence theorem, that the supposed surface layer is merely a re-expression of the integrals that constitute the *continuous volume distribution* of mass. Indeed if one were to reject this interpretation and insist that these surface integrals reveal additional mass in the form of a layer, then the Gauss theorem would indicate that this mass must be negative. Indeed various authors (e.g. [102], [96]) have referred to negative mass layers. However, as Bondi had emphasized in his writings, negative mass repels rather than attracts. Therefore we had set out to test the via-

bility of the presence of such negative mass to see if repulsion rather than attraction was in evidence. We considered a test particle in our model that was comoving with the rotating dust apart from having a component of velocity U^z normal to the $z = 0$ plane. The geodesic equation in the z direction reduces to

$$\frac{dU^z}{ds} = \frac{N_r N_z (U^z)^2}{2r} \tag{A.8}$$

We had computed the complete N series for the galaxy NGC7331 (see [63]). We then focused upon points in the range $r = 0.1$ to 30 and points above the $z = 0$ symmetry plane $z = 0.001$ to 1 for the right hand side of (A.8). All of the points gave a negative value as expected for the z acceleration (i.e. attraction) of a particle in the region above the symmetry plane. However, if the $z = 0$ surface actually harboured a *physical* negative mass surface layer, indeed one of numerical value comparable to the positive mass of the normal galactic distribution, then at the very least, one would have expected to witness a *repulsion* of the particle as the test particle approached the boundary. The absence of this occurence adds further support to our original model [63] as being free of surface layers of mass.

It is true that our choice of solution leads to a discontinuity in the z-derivative of N across the $z = 0$ plane. It is well to reiterate and emphasize the argument: it goes hand-in-hand with the physically natural density *gradient* discontinuity across the symmetry plane. (This is even more benign than the density discontinuity in the constant non-zero density Schwarzschild sphere solution matched to the exterior vacuum Schwarzschild solution.)

To see this, consider the essential characteristics of our model which consists of dust with reflection symmetry about the $z = 0$ plane. The density naturally increases symmetrically as this plane is approached from above and from below with the same absolute value but opposite sign from symmetry. In all generality, the density z-gradient will be different from zero as this plane is approached and because of reflection symmetry, this gradient will of necessity be discontinuous.

The density gradient is governed by the behavior of odd derivatives of N with respect to z. However, the density itself is governed by N_z^2 (9.12) which has the same limit as z approaches 0 from above or below. Thus, we define the value of ρ at $z = 0$ by this common

limit and hence the singularity is removable. It is only with delicate fine-tuning that this discontinuity can be avoided and this will be the case only if the density gradient is adjusted to be precisely zero as the $z = 0$ plane is approached from above and below.

As an exercise in response to critical comments [64], we achieved this approximately by choosing $\cosh(\kappa_n z)$ functions in place of exponential functions to span the region in a sandwich encompassing the symmetry plane and employing the usual exponential functions beyond this sandwich. This led to the issue of matching the N and N_z functions along the external/internal region joins and it was achieved by using many different k_n parameters for the external exponential functions as opposed to the original 10 internal parameters of the original model. In [102], it was claimed that a matching could not be achieved but these authors had not realized that we used different and many parameters for the outside regions. Since then, we have refined the fit further by employing hundreds of external parameters and the improved fit is shown in Figure A.2. However, it must be stressed that the generic situation would be one in which the density gradient is discontinuous at $z = 0$.

In a follow-up paper by these authors [103], they pasted a finite thickness layer of density and stress as a sandwich about the $z = 0$ symmetry plane. This is of some interest in building more general galactic models. However, there remains the assertion that when the sandwich is reduced to zero thickness, a surface layer arises which, by the right choice of parameter, results in having the layer consist of negative mass. The fact that test particles are attracted towards rather than repelled from this layer, regardless of the assumed sign of the parameter, negates this interpretation.

In this paper, the authors draw a connection with the benign $z = 0$ density gradient discontinuity plane and the struts that occur in the literature to stabilize the superposition of static bodies in general relativity in otherwise empty space. It is instructive and of value to consider the contrast between our problem and this static strut situation. Two bodies in a line released at rest will not stay at rest unless they are supported by a strut under pressure. Weyl showed that the line singularity revealing this strut is related to the gravitational force between the bodies in the Newtonian approximation. The singularity is evidenced by the breakdown of "elementary

flatness" along the z axis between the bodies. This expression refers to fact that as one focuses on ever smaller regions, the geometry approaches ever closer to that of zero curvature. This is familiar with our experience with a hollow spherical shell. As we take smaller and smaller slices of the shell, the pieces resemble more and more closely the segments of a perfectly flat sheet. Regular regions of spacetime have this property. Otherwise, there is said to exist a "singularity".[c] By contrast, if the two bodies are in the proper state of *constant* circular motion about each other, they will continue in this state *without* any intervening strut in Newtonian gravity. In the general case, general relativity achieves this in approximation since the general relativity field is necessarily dynamic at higher orders. However, we see from the dust solution that by making the rotating system axially symmetric, stationarity is achieved (in fact perfectly as evidenced by exact solutions in the literature). This is a relatively simple step up from the Newtonian two-body orbit example. It underlines the freedom from essential singular distributions that become possible by virtue of rotation.

In a critique [104], the well-known expression of the field equations in the harmonic gauge in Cartesian coordinates

$$\partial_k \partial^k h^{ab} = \frac{16\pi G}{c^4} \tau^{ab} \tag{A.9}$$

(τ^{ab} includes the energy-momentum tensor of the matter plus the non-linear terms in the Einstein equations) is invoked. (A related line of reasoning was followed in [101]). In [104], the author presents the standard description of the post-Newtonian perturbation scheme to conclude that the solution to the galactic problem must be the

[c] A simple but excellent example is provided by the surface of a cone. If any point other than the vertex point is chosen as the center of a circle drawn on the surface of the cone, the circumference of the circle divided by its diameter has the value π. However, if the (singular) vertex point is chosen, the ratio is not π. Note that the surface of a cone has no intrinsic curvature: if the cone were to be slit open, it would flatten perfectly onto the plane. By contrast, a spherical surface or the surface of a saddle would not flatten on the plane as does a cone. The sphere and the saddle surface have intrinsic curvature. In the general situations with intrinsic curvature, at non-singular points, the circumference over the diameter *approaches* π as the size of the circle is allowed to approach 0. The flat space behavior is only approached. The non-singular points satisfy the demand for elementary flatness. In mathematical language, the π ratio in the limit is said to indicate the presence of a "local Minkowski tangent space".

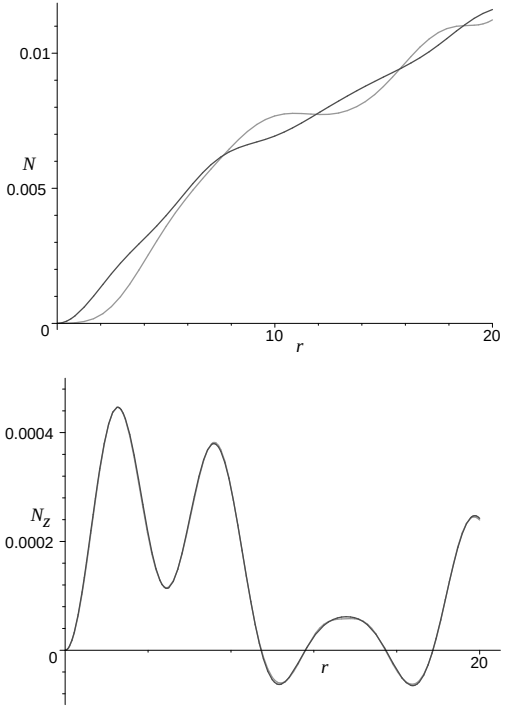

Figure A.2: Matching conditions for N and $\partial N/\partial z$ at $z = z_0$.

usual Newtonian one and that all corrections must be of higher order. Firstly, we did not use this scheme (as noted as well in the paper [105]). Just as one would not logically choose Cartesian coordinates in the harmonic gauge to describe Friedmann cosmologies, one would not normally choose these for our stationary axially symmetric galactic problem. Our problem is greatly simplified with cylindrical polar coordinates comoving with the matter. Secondly, for the gravitationally bound system under study, the metric components are of *different* orders in G[d]. If one were to take the approach suggested in [104], the equations (A.9) could be schematically expressed as

$$\nabla^2 h_{(1/2)} = 0, \quad \nabla^2 h_{(1)} = GT + h^2_{(1/2)} \tag{A.10}$$

where tensorial superscripts have been suppressed and the lower case numbers refer to orders in G. In this manner, we would have incorporated the non-linear structure of our system within the framework of the scheme suggested by [104]. The novel aspect is that the lowest order equation (of order $G^{1/2}$) in (A.10) has zero on the RHS and the second equation that would normally be the Newtonian Poisson equation, differs in that it has non-linear terms. Thus, the structure of our solution does not proceed as in the standard approach of (A.9). In the latter standard approach, the lowest order base solution is the Newtonian solution whereas in the galactic problem, the lowest order equation is the Laplace equation for which an order $G^{1/2}$ solution is necessary (see the paper [106] where this component is inappropriately chosen to be zero) and the next order (order G^1) equation for the density (A.11),

$$\frac{N_r^2 + N_z^2}{r^2} = \frac{8\pi G\rho}{c^2} \tag{A.11}$$

has non-linear terms in the metric in the form of the squares of the derivatives of an order $G^{1/2}$ metric tensor component N. Thus, our situation is unlike standard iterative perturbation scheme applications as envisaged in [104]. Hence there is no basis to draw the conclusions that are expressed therein.

Further in [104], the author refers to "extra matter" in the symmetry plane of the galaxy and muses whether our model "could be

[d] In general, first order perturbations lead to linear equations. However, there are situations where this is not the case. For example in fluid mechanics, certain approximations still lead to non-linear equations as in Burger's equation.

somehow fixed". However, in [64] we presented the evidence that our solution embodies the physically natural density gradient discontinuity at the plane of symmetry and that it does not contain extra matter. Moreover, we showed that if there were to be a surface layer of mass such as had been claimed, it would be negative mass but this was negated by the attraction rather than repulsion of test particles near the symmetry plane, as we discussed above.

The author concludes with an argument to attempt to provide dark matter through general relativity in the form of a "geon" [e] where general relativity would be required and he deduces that this is impossible [104]. While we are in agreement with him that it is indeed impossible with geons (but from a different line of reasoning, see [107], [108]), the argument is irrelevant because the galactic field is weak and hence geons are *a priori* out of the question, even if they were viable in principle.

With regard to the issue of gauge, it was argued in [101] that asymptotically flat solutions are unattainable with a lead-off $G^{1/2}$ order metric component. However, we have shown that they are readily attainable in conjunction with the physically desirable N_z discontinuity and are approximately attainable with the smoothed fine-tuned solution discussed above. Moreover, they are precisely attainable when an essential singularity is invoked[f]. A key point is that the equations have an inherent non-linearity as a result of the fact that the metric components are of different orders and the different orders are a necessary consequence of the problem being a gravitationally bound one.

In the paper [106], the author brings up the covariant vorticity and (vanishing) shear tensors for the rotating dust and poses the latter characteristic as being inconsistent with a galaxy that has differential rotation. However, a rotating dust cloud cannot physically rotate rigidly as does a disk of steel which has internal stresses. The answer to [106] is that the vanishing *covariant* shear is analogous to the vanishing *covariant* acceleration of a freely moving particle. In

[e] A geon is a theoretical construct of a concentration of wave fields, be they electromagnetic or gravitational, of such high density that they gravitate into a ball, displaying mass as does a ball of normal material.

[f] This was almost achieved in [109]. Their axis singularity prevented global asymptotic flatness. However, exact solutions with compactified singularities of the Weyl type are likely to rectify this deficiency.

the case of the latter, it is only under very special conditions that the motion will be one of constant velocity. The generic motion will be conical or more complicated. This could have been recognized in [106] where the correct *differential* local angular velocity $cN(r,z)/r^2$ is displayed. Also in [106], when the author chooses a solution for which the N function is taken to be zero at order $G^{1/2}$, he is being inconsistent with the demand that this is a gravitationally bound problem with rotation. Finally this author treats the transformation $\phi \to \bar{\phi} = \phi + \omega(r,z)t$ as a 'global' transformation to the 'co-moving frame'. However it is the original coordinate set that constitutes the co-moving frame and moreover, this transformation has value strictly as a local transformation.

The paper [109] arrives at our equations (9.7), (9.8) (with w set to zero) apart from the exponential ν factor which they later note can be taken to be a constant scaling factor. These authors find the same order of magnitude reduction of galactic mass that we had found [63] starting from their exact solution class. This provides some vindication for our analysis. It should be noted that their scaling factor is actually incorporated in our solutions within the computed amplitudes of our basis expansion functions. To be particularly noted in [109] is that their solution class is fine-tuned as the density gradient is precisely zero at $z = 0$. The price that is paid to achieve this degree of smoothness is the incorporation of an axial singularity. These authors justify the singularity by identifying it as a jet. While jets are observed in various galaxies in their formative stages, they are not known to be present in the essentially stationary galaxies that are being modeled with this class.

Quite apart from the issue of jet interpretation, in [109], given their non-separable solution, it is possible that the proper circumference of the azimuthal direction in their case pinches off to zero before r becomes zero. It would then be more reasonable to regard this r value as the true physical center of the galaxy for their case and the smaller r values should be excised from the spacetime. As well, for their strong field region, proper rather than coordinate radial distance should be recognized for the physical interpretation of distance. These arguments are related to the comments made in the recent comprehensive paper by Wiltshire [110].

A detailed analysis of the exact van Stockum [68] spacetime is

provided in the paper [111] from which the authors derive and transfer supposed restrictions onto our work. However in doing so, they miss the point that we are analyzing in generality the *weak-field* stationary axially symmetric dust spacetimes. This is necessarily approximate as a result and it fits the physical situation at hand. No physical galaxy is exactly stationary; a galaxy evolves with time. Hence restrictions derived on an exactness basis from some particular *stationary* solution are not relevant to us. Our solutions are approximate, with sufficient accuracy for the physical situation at hand. Only if galaxies were exactly stationary, would exact analyses of stationary spacetimes be relevant. But they are not exactly stationary.

Moreover, we have now considered various velocity fall-off scenarios beyond the HI regions and have extended our rotation curves to match these assumed fall-offs.[g] These are shown in Figure A.3. To accommodate this expanded region, this requires a completely new and enlarged set of parameters from the small set that we used originally. These are used to follow the relatively flat region and then merge into the fall-off region. It is to be noted that since the velocity continues to be constructed in the form of (9.18), there is continuity of the curve and its derivatives apart from the value at precisely $z = 0$ discussed previously. The kink in Figure A.3 is only apparent, arising from the practical need to fit the subtle transition into a compressed graph.

From (9.12), we see that the density vanishes when N_r and N_z are both zero, i.e. when N is a constant. Also, from (9.15), when N is a constant, the velocity falls as $1/r$. Therefore, at first glance one might believe that by choosing the continued velocity curves beyond the HI region in the form A_0/r, one would be tracking r into the vacuum, identifying N with a constant A_0 and hence accumulating no further mass. However, while in plotting rotation curves at a given z, it is only a net independence in r for the N function that is required to give an A_0/r fall-off in V. The z dependence in the N function can still be present. Thus, the density will still not be

[g]It is to be noted that this is in keeping with our desire to work with globally dust models, thus avoiding transition issues for the metric and its derivatives in going from dust to total vacuum. Also to be stressed is that the points beyond the HI region are not based upon observed data but rather are artificially imposed to induce smooth matter fall-offs of various forms.

zero and mass will still be accumulated as one tracks in the radial direction. This is evident in Figure A.4. Faster fall-off rates with r would improve the trend towards vacuum further. Slower fall-offs such as of the form $1/\sqrt{r}$ which might be suggested from Newtonian gravity are clearly not adequate to merge towards vanishing density.

A fall-off of the form $1/\sqrt{r}$ would be appropriate to impose for test particles in the field of a massive body such as is the case in the solar system. However, here we have seen that in the case of a continuous gravitationally bound source, general relativity presents a dynamical system with behavior that does not match the Newtonian picture. It is also not necessary to envisage such slow rates since we are basing our analysis on the preferred theory of gravity, namely general relativity. The challenge is to find a general relativistic solution that merges properly into near vacuum and we have met that challenge.

We display the accumulated mass for the Milky Way in a highly extended cylindrical volume of 300 Kpc in size in Figure A.4. It is to be noted that even assuming a Newtonian-like fall-off of the form $1/\sqrt{r}$, there is a far less amount of accumulated mass up to a radius of ten times the visible radius than is envisaged by the use of Newtonian as opposed to general relativistic galactic dynamics. An even slower accumulation of mass is seen for the $1/r$ fall-off. For such a fall-off, the accumulated mass is approximately $35 \times 10^{10} M_{\odot}$ at a radius of 300 Kpc and a linear extrapolation to $r = 900$ Kpc yields a value of $39.2 \times 10^{10} M_{\odot}$, a very modest increase in comparison to Newtonian modeling. Moreover, the faster fall-offs of $1/r^2$ and $1/r^4$ yield very minor mass increases out to very large radii as can be seen in Figure A.4.[h]

This fortifies our contention that general relativity obviates the need for overwhelmingly dominant massive halos of exotic dark matter.

It is to be noted that our models are *globally* dust[i] and therefore there is no basis for a matching with the vacuum Kerr metric as-

[h] Note from this figure that the accumulated mass at 30 Kpc is approximately the same for the various fall-off scenarios as well as the value stated in Chapter 3 where we used only 10 parameters and where we did not focus on the bahavior of the model beyond the 30 Kpc edge of the HI region.

[i] We make this choice for the composition and distribution of the matter for the purpose of mathematical simplicity.

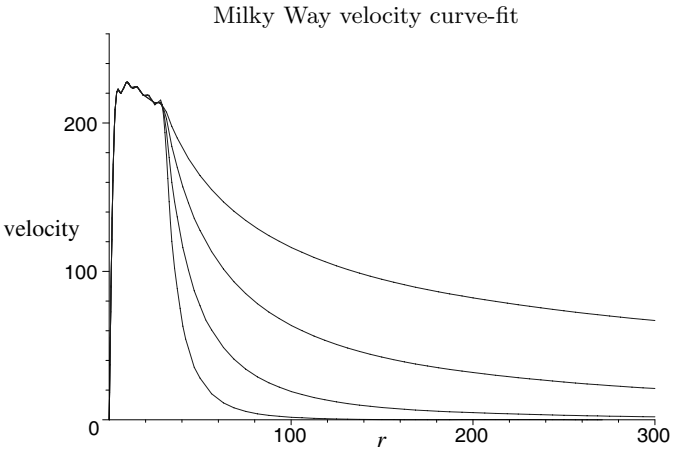

Figure A.3: Milky Way velocity curve fit: beyond the HI region, the velocity can be modeled in many different manners: here $V \propto 1/\sqrt{r}$, $V \propto 1/r$, $V \propto 1/r^2$ and $V \propto 1/r^4$ are illustrated.

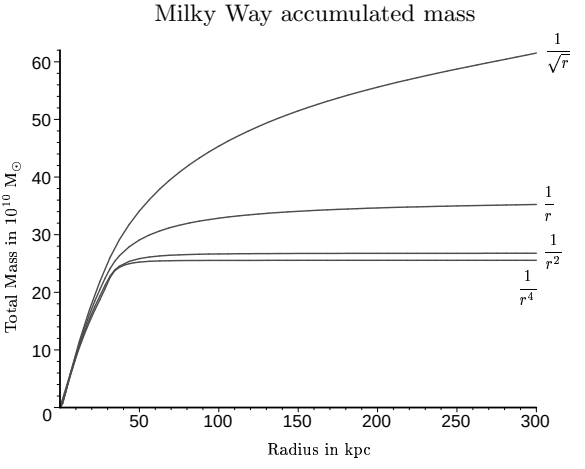

Figure A.4: The Milky Ways's accumulated mass as a consequence of velocity fall-off beyond the HI region.

ymptotically. Our models are asymptotically flat with a well-defined mass. The Tolman integral dictates the value of this mass and since there is no stress and the fields are weak, the Tolman mass to lowest order is simply given by the coordinate volume integral of the density.

In the paper [112], the observed solar neighborhood density data are compared with the values given by our model. These authors find that our density is less by a factor of approximately six. Firstly, it is to be noted that there are considerable error bars on their data and it is possible that our neighborhood could be one of enhanced density. Secondly, it must be emphasized that our model is adjusted in the simplest terms to account for the overall mean velocity distribution with a mere ten parameters.[j] This scant input can hardly be expected to account for the distribution of hundreds of billions of stars and their motions. This is beyond the capacity of even Newtonian theory let alone general relativity. By necessity, great simplifications are necessary in practice. Moreover, as additional support for our model having the correct overall characteristics, our integrated density over the visible region falls within predicted limits.

Even if one were to assume a logical basis for making their comparison between local observed data and our globally derived very approximate data, it should be noted that the vertical distribution of observed stars in the local galactic disk has a sharp peak [113]. This fits in well with the presence of a density gradient discontinuity in our models. Moreover, it should be noted that for mathematical tractability, we have assumed that the density peaks all occur at $z = 0$. However, there will naturally be some variation in the location in z for these peaks. Due to the rapid decline in density, there will be large local variations in density and therefore the criticisms in [112] about our distributions are further seen to be inappropriate.

In [114], the author models the galaxy via Newtonian physics with a surface mass layer. Clearly, given the freedom to impose the properly adjusted internal stresses, virtually any kind of approximate velocity distribution can be simulated. This is inadequate for the problem at hand on two counts: firstly, a good model should be stress-free (i.e. purely free-fall gravitationally driven) and secondly,

[j] Even with the enlarged parameter set used to model the asymptotic fall-off, the parameter input is still quite modest.

the point is to develop an extended *volume* distribution and hence more akin to reality. This is what we have set out to do and this, within general relativity, the preferred theory of gravity. Moreover, there is the suggestion, sometimes reiterated by others, that we claim to have modeled galaxies without any dark matter whatsoever, this in spite of the fact that we have explicitly referred to dead stars, neutron stars, etc. as dark matter constituents of galaxies and we have presented mass-to-light ratios. What we are questioning is the existence of *exotic* dark matter, the supposed constituent of the massive halos 5 to 10 times the traditionally computed mass that are said to surround galaxies, matter that has no known counterpart to the matter that physicists identify in laboratories and particle accelerators. Our approach is in keeping with the spirit of Occam.

In [115], the authors fault our models as extended constructs that indicate enormous quantities of mass beyond the HI regions. This is a useful point of criticism in that we had not investigated earlier the asymptotic consequences of the particular parameter sets that we had chosen to model the observed rotation curves.[k] However, in [66], as first reported in [116], we assure more realistic fall-off scenarios[l] and we find that the accumulated mass profiles indicate that most of the mass of a galaxy is confined fairly close to the region of the visible disk with *modest* accumulations of mass beyond this region. General relativity achieves this with a pressure-free fluid model, unlike Newtonian gravity.

In a more recent paper [117], these same authors presented new points of criticism regarding our derivation of tangential velocity. They begin by considering "a zero-momentum particle $p_\phi = 0$" but in our co-moving system, the zero-momentum particle has $p^\phi = 0$ and p_ϕ differs from zero. Regardless, of essence is the relative velocity between the local inertial frames (that are carried by the geodesic dust particles) and the local non-rotating frames, and their deriva-

[k] Note however that their argument that mass accumulates linearly in r is faulty as a generalization. With the correct combination of parameters, the term that would lead to such an accumulation can be eliminated. Our examples in which we achieve minimal accumulation, provide the direct proof that this is the case.

[l] It is to be noted that in so doing, while the expansion parameters are no longer the same as in the earlier sets, we have determined that the net physical effects are of insignificant difference within the observed matter distribution in the two approaches.

tion still reproduces our result. It does so because by setting the covariant rather than the contravariant component to zero, they are setting the locally non-rotating frame, what some refer to as the zero angular momentum frame, to have zero velocity. Their frame is not co-moving with the particles. This is the opposite of our procedure but it achieves the same result, the same *relative* velocity. They claimed that our derivation "does not describe the tangential velocity of a dust particle moving on a geodesic," with the implication that we have failed to account for "derivatives in the metric in the connection term." However the velocities of our dust particles, their geodesic motions, are, as in many other accounts in the literature, simply zero relative to our co-moving frame. The geodesic equation with its connection coefficients, is already implicitly satisfied. It is this velocity, namely zero, relative to the *local* non-rotating frame that must be computed and this is done by the *local* transformation that we have effected, as was done by others in the literature. Further, the authors state: "Clearly the frame-dragging rate is normally much smaller than the circular geodesic rate for non-relativistic systems, so in general these cannot be equated, as seems to have been done by these authors in [astro-ph/0507619]"[63]. However dust particles follow geodesic paths and hence they transport locally inertial frames. Thus they are all in effect "frame-dragged" and this holds whether their velocities are high or low.

The earlier comments referring to [111] apply also to the paper [118]. Moreover, as described above, we have avoided the complications of merging from the dust regions to vacuum by dealing with models that are globally dust. Since the dust in our models become extremely diffuse with distance, the physical distinction between having the global dust and the vacuum is inconsequential. Also to be noted is that since the N function is ultimately connected to solutions of the Laplace equation, there will necessarily be a singularity of some form present. It is quite acceptable if it is the right kind of singularity and in our construction it is the case, modeling the physically desirable density gradient discontinuity. This is a singularity that has been misinterpreted by some of the authors discussed previously.

Indeed it must be stressed that singularities have to be interpreted. There are subtleties attached to their nature and they can

readily lead the unwary astray.

In an interesting approach from a very different direction [119], the author has pointed to relativistic inertial effects that do not have a Newtonian limit counterpart. He has suggested that in the weak field limit, these effects could match our results.

This page intentionally left blank

Appendix B

Radial Velocity Derivation Details

In the galactic cluster model, the essential details of the derivation of (10.23), the radial velocity of the falling matter as viewed by a distant observer, are now provided.

We take differentials of each of the equations in (10.20), and with (10.21), (10.22), solve for the differentials dR and $d\tau$. These differentials are substituted into (10.10) to derive the normal form of the metric in Schwarzschild-like coordinates (r, t) with terms of the form $g_{00}dt^2$ and $g_{rr}dr^2$, as well as a cross-term of the form $2g_{0r}drdt$. We need to eliminate this cross-term in order to join smoothly with the diagonal exterior vacuum Schwarzschild metric at the dust-vacuum interface. After effecting the transformation, we impose the vanishing of the cross-term g_{0r} and find a connection of the solution with $p'(r, t)$. The resulting equation contains e^{λ}.

In order to achieve the streamlined form for p' of (B.4) below, we require the expression for e^{λ} which, from (10.15), is equal to $(r')^2$ for $E = 0$. In turn, this requires r' which is easily found from a differentiation with respect to R of (10.17). This gives

$$r' = \alpha + \beta. \tag{B.1}$$

where

$$\alpha = \frac{rF'}{3F} = \frac{rM'(R)}{3M(R)} \tag{B.2}$$

$$\beta = \sqrt{\frac{F}{r}} = \sqrt{\frac{2M(R)}{r}}. \tag{B.3}$$

In this manner, r' (and hence e^{λ}) is expressed in terms of the useful quantities α and β. When this is substituted into the equation $g_{0r} = 0$, we are able to express p' in the useful form

$$p' = \frac{\left(\frac{3R\sqrt{F}\alpha}{2r} + \sqrt{r}\beta\right)}{(\alpha+\beta)(1-\beta^2).} \tag{B.4}$$

This is the expression for p' in all generality for transformation to the diagonal metric form in (r,t).

We now determine how p' as well as $\dot{p}$ fit into the expression for velocity. Recall that every dust element has its own R coordinate value which is fixed for all time (i.e. comoving coordinates). Therefore an expression of the radial motion of any given element is $dR = 0$. (Note that for motion in this case of spherical symmetry, $d\theta = d\phi = 0$ as well.) We can express this condition in terms of the Schwarzschild-like coordinates by taking differentials of the first of (10.20) and setting $dR = 0$ to get

$$p'(r,t)dr + \dot{p}(r,t)dt = 0. \tag{B.5}$$

Thus, the radial velocity dr/dt of the particles as witnessed by external distant observers is expressed in generality as the ratio of the partial derivatives of the first transformation function $p(r,t)$ as

$$dr/dt = -\dot{p}(r,t)/p'(r,t). \tag{B.6}$$

While seemingly simple, several steps are required to bring this important quantity into a useful form. We have already determined the useful form for $p'(r,t)$. To express $\dot{p}(r,t)$, we first apply $\partial/\partial t$ to (10.18):

$$8\pi\frac{\partial\rho}{\partial t} = \frac{F'^2\left(\frac{\alpha}{F} + \beta\left(\frac{F''}{F'^2} - \frac{1}{2F}\right)\right)\dot{p}}{r^2(\alpha+\beta)^2\left(\frac{3R\sqrt{F}\alpha}{2r} + \sqrt{r}\beta\right)}. \tag{B.7}$$

The derivation of (B.7) made use of (B.1),

$$\frac{\partial(\alpha+\beta)}{\partial t} = \left[\frac{F'}{2\sqrt{Fr}} + \frac{r}{3}\left(\frac{F''}{F} - \frac{F'^2}{F^2}\right)\right]\frac{\partial R}{\partial t} \tag{B.8}$$

(where (B.2) and (B.3) have been used) and the elimination of $\frac{\partial R}{\partial t}$ using

$$\dot{p} = \left[\frac{F'R}{2\sqrt{F}} + \sqrt{F}\right]\frac{\partial R}{\partial t} \tag{B.9}$$

which follows from the partial differentiation with respect to t of the first of (10.20). Finally, using (B.4) and (B.7) in conjunction with (B.6) (and with a cancellation of the factor $(\frac{3R\sqrt{F}\alpha}{2r} + \sqrt{r}\beta)$), we find

$$\frac{dr}{dt} = -\frac{(\alpha+\beta)(1-\beta^2)}{8\pi r^2\rho^2}\left[\frac{\alpha}{F} + \beta\left(\frac{F''}{(F')^2} - \frac{1}{2F}\right)\right]^{-1}\frac{\partial\rho}{\partial t}. \qquad \text{(B.10)}$$

This is the important distant-observer-based velocity of the elements of the distribution. It stands in sharp contrast to the very simple Newtonian-like expression

$$\beta = \sqrt{\frac{F}{r}}. \qquad \text{(B.11)}$$

This page intentionally left blank

Bibliography

[1] T. Paine, *The Rights of Man*, Penguin Classics, 1984.

[2] A. Einstein, *The Meaning of Relativity*, Princeton University Press, 1923.

[3] L. D. Landau and E. M. Lifshitz, *The Classical Theory of Fields* Fourth revised English edition, Pergamon Press, Oxford, 1975.

[4] H. Bondi in *Lectures on General Relativity, Brandeis Summer Institute in Theoretical Physics*, eds. S. Deser and K.W. Ford, Vol.1, Prentice-Hall, New Jersey, 1965.

[5] F. I. Cooperstock and R. S. Sarracino, *Nature* **264**, 529, 1976.

[6] R. P. Feynman, R. B. Leighton and M. Sands, *The Feynman Lectures on Physics*, Addison-Wesley, Reading, Mass., 1964.

[7] H. Goldstein, *Classical Mechanics*, second edition, Addison-Wesley, Reading, Mass., 1980.

[8] B. Spain, *Tensor Calculus*, Oliver and Boyd, Edinburgh, 1960.

[9] J. D. Jackson, *Classical Electrodynamics*, John Wiley and Sons, New York, 1962.

[10] J. Weber, *General Relativity and Gravitational Waves*, Interscience Publishers Inc., New York, 1961.

[11] R. Adler, M. Bazin and M. Schiffer, *Introduction to General Relativity*, McGraw-Hill, New York, 1965.

[12] A. Einstein and N. Rosen, *Phys. Rev.* **48**, 73, 1935.

[13] F. I. Cooperstock and G. J. G. Junevicus, *Nuovo Cimento* **16B**, 387, 1973.

[14] J. L. Synge, *Proc. Roy. Irish Acad.* **A53**, 83, 1950.

[15] N. Rosen, in *Relativity*, Proceedings of the Relativity Conference in the Midwest, eds. M Carmeli, S. I. Fickler and L. Witten, Plenum Press, New York, 1970.

[16] F. I. Cooperstock, S. Jhingan, P. S. Joshi and T. P. Singh, *Class. Quantum Grav.* **14**, 2195, 1997.

[17] J. L. Synge, *Relativity: The General Theory*, North-Holland, Amsterdam, 1966.

[18] I. I. Shapiro, *Phys. Rev. Lett.* **13**, 789, 1964; ibid. **26**, 1132, 1972.

[19] A. Papapetrou, *Proc. R. Soc. London A*. **209**, 248, 1951.

[20] J. M. Overduin in Y. Nambu, ed., Proceedings of JGRG17, the 17th Workshop on General Relativity and Gravitation in Japan (Nagoya, 3-7 December 2007). A more convenient reference for the non-expert, by J. M. Overduin is http://einstein.stanford.edu/SPACETIME/spacetime-index.html, "Spacetime: from the Greeks to Gravity Probe B".

[21] W. B. Bonnor, *Br. J. Appl. Phys.* **14**, 555, 1963.

[22] F. I. Cooperstock and D. W. Hobill, *Gen. Rel. Grav.* **14**, 361, 1982.

[23] F. I. Cooperstock, *Ann. Phys. N.Y.* **282**, 115, 2000; F. I. Cooperstock and S. Tieu, *Found. Phys.* **33**, 1033, 2003.

[24] J. M. Nester, *Phys. Rev. Lett.* **83**, 1897, 1999.

[25] J. M. Aguirregabiria, A. Chamorro and K. S. Virbhadra, *Gen. Rel. Grav.* **28**, 1393, 1996.

[26] N. Nissani and E. Leibovitz, *Phys. Lett. A* **126**, 447, 1988.

[27] M. J. Dupre, preprint, 2008.

[28] W. B. Bonnor, *Commun. Math. Phys.* **51**, 191, 1976.

[29] M. Gurses and F. Gursey, *J. Math. Phys.* **16**, 2385, 1975.

[30] F. I. Cooperstock, *Ann. Phys. N.Y.* **47**, 173, 1968.

[31] H. Bondi, M. G. J. van der Burg and and A. W. K. Metzner, *Proc. Roy. Soc.* **A269**, 21, 1962.

[32] J. Madore, *Ann. Inst. Henri Poincare'* **12**, 365, 1970.

[33] F. I. Cooperstock and D. W. Hobill, *Phys. Rev.* **D20**, 2995, 1979.

[34] A. Papapetrou, *Ann. Phys.* (Leipzig), **20**, 399, 1957; **1**, 185, 1958.

[35] L. Bel in *Relativistic Astrophysics and Cosmology*, eds. J. Buitrago *et al*, World Scientific, Singapore, 1997.

[36] F. I. Cooperstock and V. Faraoni, *Class. Quantum Grav.* **10**, 1189, 1993.

[37] N. Rosen, *Phys. Rev.* **55**, 94, 1939.

[38] F. I. Cooperstock and N. Rosen, *Int. J. Theor. Phys.* **28**, 423, 1989; F. I. Cooperstock, *Developments in General Relativity, Astrophysics and Quantum Theory* eds. F. I. Cooperstock, L. P. Horwitz and J. Rosen, IOP Publishing Ltd, Bristol, England, 1990.

[39] C. S. Bohun and F. I. Cooperstock, *Phys. Rev.* **A60**, 4291, 1999.

[40] F. I. Cooperstock and V. Faraoni , *Mod. Phys. Lett. A* **18**, 1037, 2003; *Int. J. Mod. Phys.* **D12**, 1657, 2003.

[41] E. T. Newman *et al.*, *J. Math. Phys.* **6**, 918, 1965.

[42] J. K. Webb *et al.*, *Phys. Rev. Lett.* **82**, 884, 1999.

[43] M. T. Murphy *et al.*, *Mon. Not. R. Ast. Soc.* **327**, 1208, 2001.

[44] P. C. W. Davies, T. M. Davis and C. H. Lineweaver, *Nature* **418**, 602, 2002; S. Carlip,*Phys. Rev. D* **67**, 023507, 2003; M. J. Duff,*hep-th*/0208093; S. Carlip and S. Vaidya, *hep-th*/0209249; M. Fairbarn and M. H. G. Tytgat, *hep-th*/0212105; S. Das and G. Kunstatter, *hep-th*/02012334.

[45] J. W. Moffat, *Int. J. Mod. Phys. D* **2**, 351, 1993.

[46] A. Peres,*quant-ph*/0209114.

[47] E. R. Harrison, *Nature* **204**, 271, 1964.

[48] P. S. Wesson, K. Valle and R. Stabell, *Astrophys. J.* **317**, 601, 1987.

[49] J. M. Overduin, *The Universe Seen Darkly*, *Space Sci. Reviews*, to appear, 2008.

[50] H. Kragh, *Conceptions of Cosmos*, Oxford University Press, 2008.

[51] J. M. Overduin and F. I. Cooperstock, *Phys. Rev. D* **58**, 043506, 1998.

[52] S. P. Starkovich and F. I. Cooperstock, *Astrophys. J.* **398**, 1, 1992.

[53] S. S. Bayin, F. I. Cooperstock and V. Faraoni, *Astrophys. J.* **428**, 439, 1994.

[54] M. Israelit and N. Rosen, *Astrophys. J.* **342** , 627, 1989.

[55] D. Wiltshire, *Phys. Rev. Lett.* **99**, 251101, 2007.

[56] T. Padmanabhan, arXiv/08072356.

[57] M. Milgrom, *Astrophys. J.* **270**, 365, 1983.

[58] M. Milgrom and J. D. Bekenstein, *Astrophys. J.* **286**, 7, 1984.

[59] J. D. Bekenstein, *Developments in General Relativity, Astrophysics and Quantum Theory* Eds. F. I. Cooperstock, L. P. Horwitz and J. Rosen, IOP Publishing Ltd, Bristol, England, 1990.

[60] L. Clewley et al, astro-ph/0310675; M. Wilkinson and N. Evans, *Mon. Not. R. Ast. Soc.* **310**, 645, 1999.

[61] J. R. Brownstein and J. W. Moffat, astro-ph/0506370.

[62] A. S. Eddington, *Proc. Roy. Soc.* **A 102**, 268, 1922; *The Mathematical Theory of Relativity*, Cambridge Univ. Press, Cambridge, U.K., 1923.

[63] F. I. Cooperstock and S. Tieu, astro-ph/0507619.

[64] F. I. Cooperstock and S. Tieu, astro-ph/0512048.

[65] F. I. Cooperstock and S. Tieu, *Mod. Phys. Lett. A*. **21**, 2133, 2006.

[66] F. I. Cooperstock and S. Tieu, *Int. J. Mod. Phys. A* **22**, 2293, 2007.

[67] W. B. Bonnor, *J. Phys. A: Math. Gen.* **10**, 1673, 1977.

[68] W. J. van Stockum, *Proc. R. Soc. Edin.* **57**, 135, 1937.

[69] J. M. Bardeen, *Astrophys. J.* **162**, 71, 1970.

[70] A. Ashtekar and A. Magnon, *J. Math. Phys.* **16**, 341, 1975.

[71] J. Winicour, *J. Math. Phys.* **16**, 1805, 1975.

[72] J. C. N. de Araujo and A. Wang, *Gen. Rel. Grav.* **32**, 1971, 2000.

[73] L. B. Ford, *Differential Equations* McGraw-Hill, New York, 1955.

[74] L. Mestel, *Mon. Not. R. Astron. Soc.* **126**, 553, 1963.

[75] S. M. Kent, *Astron. J.* **93**, 816, 1987.

[76] F. Fraternali et al, astro-ph/0410375.

[77] G. Battaglia *et al*, *Mon. Not. R. Ast. Soc.* **364**, 433, 2005.

[78] G. Heald, PhD Thesis, University of New Mexico, 2006.

[79] F. I. Cooperstock and S. Tieu, *Mod. Phys. Lett. A.* **23**, 1745, 2008.

[80] J. P. Hughes, astro-ph/9709272

[81] J. R. Primack, astro-ph/0312549.

[82] R. Massey et al, *Nature* **445**, 286, 2007.

[83] A. Mahdavi et al, astro-ph/07063048.

[84] J. M. Overduin and P. S. Wesson, *Dark Sky, Dark Matter*, Institute of Physics Press, Bristol, UK, 2003.

[85] I. Ozsvath and E. Schucking, *Am. J. Phys.* **71**, 801, 2003.

[86] W. Kundt, *Zeit. fur Physik* **145**, 611, 1956.

[87] S. Chandrasekhar and J. P. Wright, *Proc. Natl. Acad. Sci. U.S.A.* **47**, 341, 1961.

[88] H. Stein, *Philos. Sci.* **37**, 589, 1970.

[89] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Spacetime*, Cambridge University Press, Cambridge, 1973.

[90] M. Visser, *Lorentzian Wormholes*, AIP Press, New York, 1995.

[91] F. J. Tipler, *Phys. Rev. D* **9**, 2203, 1974.

[92] J. R. Gott, *Phys. Rev. Lett.* **66**, 1126, 1991.

[93] W. B. Bonnor, *Int. J. Mod. Phys. D* **12**, 1705, 2003.

[94] J. L. Synge, *Relativity: The Special Theory*, John Wiley and Sons, New York, 1964.

[95] K. Gödel, *Mod. Phys.* **21**, 447, 1949.

[96] W. B. Bonnor and B. R. Steadman, *Gen. Rel. Grav.* **37**, 1833, 2005.

[97] F. I. Cooperstock and S. Tieu, *Found. Phys.* **35** , 1497, 2005.

[98] L. Smolin, *The Trouble With Physics*, Houghton Mifflin Co., Boston, 2006.

[99] H. Bondi, F. A. E. Pirani and I. Robinson, *Proc. Roy. Soc. A.* **251**, 519, 1959.

[100] S. Weinberg, *Dreams of a Final Theory*, Pantheon Books, New York, 1992.

[101] M. Korzynski, astro-ph/0508377.

[102] D. Vogt and P. S. Letelier, astro-ph/0510750.

[103] D. Vogt and P. S. Letelier, astro-ph/0611428.

[104] D. Garfinkle, *Class. Quant. Grav.* **23**, 1391, 2006.

[105] M. D. Maia, A. J. S. Capistrano and D. Muller, astro-ph/0605688.

[106] D. J. Cross, astro-ph/0601191.

[107] F. I. Cooperstock, V. Faraoni and G. P. Perry, *Int. J. Mod. Phys. D* **5**, 375, 1996.

[108] G. P. Perry and F. I. Cooperstock, *Class. Quantum Grav.* **16**, 1889, 1999.

[109] H. Balasin and D. Grumiller, astro-ph/0602519.

[110] D. L. Wiltshire, *New J. Phys.* **9**, 377, 2007.

[111] L. Bratek, J. Jalocha and M. Kutschera, astro-ph/0603791.

[112] B. Fuchs and S. Phelps, *New Astron.* **11**, 608, 2006.

[113] J. Holmberg and C. Flynn, *Not. R. Ast. Soc.* **313**, 209, 2000.

[114] V. Kostov, astro-ph/0604395.

[115] D. Menzies and G. J. Mathews, gr-qc/0604092.

[116] F. I. Cooperstock and S. Tieu, (22nd Pacific Coast Gravity Meeting, KITP, Santa Barbara, California, 2006).

[117] D. Menzies and G. J. Mathews, astro-ph/0701019.

[118] T. Zingg, A. Aste and D. Trautmann, astro-ph/0608299.

[119] L. Lusanna, gr-qc/0604120.

[120] K. L. Knickerbocker, Editor, *The Selected Poetry of Robert Browning*, Random House, Modern Library College Editions, 1951.

Acknowledgements

The reader will note that the word "I" occurs only once in this book and it is in this sentence. This is not because we prefer the "royal *we*", this in spite of having Queen Elizabeth's representative for British Columbia as our next door neighbor (literally). Rather, it is because much of our work has been developed over the years in collaboration with others and it is simplest to use the "we" consistently. We are indebted to these collaborators whose varied talents and interests have brought us into contact with a wide variety of research projects and from whom we learned many things.

Valerio Faraoni was the collaborator for the research that was incorporated in the chapter on extending the Planck scale. Steven Tieu was the collaborator for the research that was incorporated in the chapters on single galactic and galactic cluster models as well as the chapter on closed timelike curves. He also created all of the illustrations for this book. This work would not have been possible without his expert contributions to our research program. Our Senior Editor, Jenny Hogan, was our guiding light for organizing this book into the form that made all the difference. Our editor, Zhang Fang, saw the production to its conclusion with great care and imagination.

We are particularly grateful to the wonderful colleagues who so kindly read the early draft and provided many valuable comments and corrections. They are James Overduin, Alberto Chamorro, Larry Horwitz, Alan Shotter, William Bonnor, Pankaj Joshi, Phil Perry and of course Steven Tieu. Peter Gary brought various valuable references to our attention. We add the standard disclaimer on behalf of these colleagues: the responsibility for any errors is ours alone.

We remember our mentors who are no longer with us: John L. Synge, Achille Papapetrou and Nathan Rosen.

Finally, our deepest gratitude for her constant encouragement and support goes to the person who always gets the very last word, Ruth.

Index