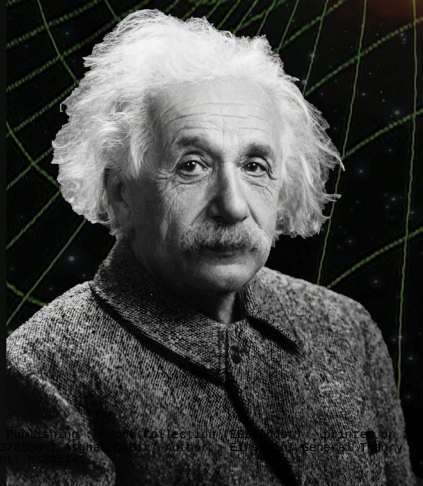
EINSTEIN'S GENERAL THEORY OF RELATIVITY

Asghar Qadir



Einstein's General Theory of Relativity

Einstein's General Theory of Relativity

by

Asghar Qadir

Cambridge
Scholars
Publishing

Einstein's General Theory of Relativity

By Asghar Qadir

This book first published 2020

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data A catalogue record for this book is available from the British Library

ISBN (10): 1-5275-4428-1
ISBN (13): 978-1-5275-4428-4

Dedicated To

My Mentors:
Manzur Qadir
Roger Penrose
Remo Ruffini
John Archibald Wheeler
and
My Wife:
Rabiya Asghar Qadir

Contents

List of Figures

Preface

This book was started more than 30 years ago and was ready in some rough form a couple of years after that. I was awarded a "Book Project" to write this book by the King Fahd University of Petroleum & Minerals nearly a quarter of a century ago and the book was ready in a typed form needing some serious editing over 20 years ago. You will notice my slow progress with it even then. It has "idled" since then until Ghulam Abbas, an ex-PhD student of Muhammad Sharif, an ex-PhD student of mine, complained publicly at a Conference about my unfairly withholding this book from the next generations of my students. I felt that he had a point and, fortuitously, I received an invitation by Cambridge Scholars Publishers to submit a book proposal. Knowing myself by now (being over 72 years I have had time to get to do so), I felt I had been remiss long enough and that it was only by committing myself that the book would ever see the light of day.

The book is based on my lectures on General Relativity since 1971, when I joined what was then the University of Islamabad, Pakistan, and later became the Quaid-i-Azam University, Islamabad, Pakistan. I taught the book at the local "M.Sc.", which is the equivalent of the senior years of the 4-year BS, and at the local "M.Phil.", which is the equivalent of the American MS and British M.Sc. It has been taught as a one-year, or two semester course. It is written so as to be able to teach students of the senior undergraduate of earlier postgraduate with a Physics background who have studied Special Relativity from my book *Relativity: An Introduction to the Special Theory* (World Scientific 1989) or equivalent, but do not have a sound background of Geometry. It can be used for students of Mathematics who have not studied Special Relativity but have a strong background of Geometry, by replacing the part on Geometry by chapters 2, 3, and parts of 5, 6 and 7. I would break the course off part way through Chapter 5, at section 5, and proceed for the rest of Chapter 5 and the next three chapters in the next semester. Chapter 8 of this book contains various recent developments and some other special topics (some of which could be left out from the course without any damage done to the rest of the course).

Let me also talk a bit about those to whom the book is dedicated. All my mentors said that if one cannot explain something simply, one has not understood it. My first mentor was my father, who not only started my education in Mathematics but was the person who, despite being a lawyer, first introduced me to the subject of Relativity, at the age of 9 and motivated me to try to understand the subject. I learned from him, also, that knowledge does not come by degrees but by curiosity — the desire to know. I found his understanding and knowledge of Mathematics better than many PhDs in the subject. He was very critical of anyone trying to argue by bald claims hidden behind layers of

jargon.

As regards my second mentor, I cannot imagine a better PhD supervisor than Roger Penrose. When I would say something stupid, he would not say it was stupid but that he did not understand it — *and he meant it*! When the discussion led to the correct version, he never pointed out that I had been wrong. Without appearing to guide me to the solution of the problem I had been interested in addressing when I joined him, by the end of the PhD he had got me to the stage of doing what I had wanted to achieve. He was not ready to take the "accepted wisdom" as correct, but judge it for himself each time. From him I learned to do the same. From him I also learned how Mathematics could lead, not merely to correct physical consequences, but to physical insight. He always gave credit for ideas freely and never claimed it for himself. He never put his name on a paper that was not significantly his.

From my third mentor, Remo Ruffini , I learned the importance of enthusiasm for the subject, especially in talking about it and communicating it. I had been involved in the attempt to find a Quantum Theory consistent with General Relativity. Remo got me interested in Relativistic Astrophysics. I was also fortunate to see a selfless appreciation for work on the development of ideas, rather than trying to grab credit for it. He had found the mass limit at which a collapsed object must become a black hole. When he went to China, he found that Fang Li Zhi had discovered the self-same limit, but had been unable to publish it in Western journals because this was at the time of the "Cultural Revolution". Fang had published it in China. Remo publicized the discovery by Fang as a contemporaneous independent discovery. I also learned from him the importance of using humour in and human interest in communicating serious Physics.

My fourth mentor, John Archibald Wheeler, started my love affair with Physics. He did not separate off parts of Physics but saw it as a unified whole. His "poor-man's way" of seeing results was an indispensable tool for his understanding. He needed to see things simply before going for long calculations to get the correct answer. As he said "I never start a calculation unless I know the answer". And *how* did he know the answer? By the poor-man's way. He also had a knack for catchy phrases and turns of expression. He invented the terms "black hole" and "big crunch" for example. His juxtaposition of opposites would express it all, as with "magic without magic". From him I learned the importance of saying things in a way which would catch the imagination and stay with the reader (or listener).

My wife is to blame for my still being around to write the book. If it had not been for her, it is highly unlikely that I would have actually got the book written, leave alone published, as I would have died long before.

I would be remiss not to thank all my students on whom I tried out my explanations and developed them to the point where most could follow what i taught. I must particularly thank two recent students of mine: Shameen Khattak for a very thorough proof-reading of the mathematical calculations in my book, eliminating many errors in the earlier draft; and Muhammad Usman for helping with handling the LaTeX required for typing the book and with diagrams.

Chapter 1

Introduction

If you say "Relativity", everybody thinks "Einstein" and if you say "Einstein", everybody thinks "Relativity". It may not be fair to Einstein to limit him to the theories of Relativity, as he was also the first person to believe in a quantum of energy and got a Nobel Prize for that work, along with his prediction for Brownian motion. Nor, for that matter, is the theory of Relativity solely developed by Einstein. The names of Poincaré and Hilbert are often mentioned as co-founders for the development of the Special and General Theories, respectively (and the name of Marcel Grossmann strangely suppressed for the latter). I will try to explain the development of the unrestricted theory, following a historical perspective, and explain why the theory should genuinely be regarded as Einstein's creation, despite all the contributions of other researchers. However, the essential purpose of this book is to explain the unrestricted, or general, theory so that the reader can actually follow the latest developments in the theory. But first some words about the first, *restricted*, or special, theory (being restricted to constant velocity).

When Special Relativity was developed, the misnomer of the theory created a lot of confusion. What Einstein had developed was a theory that said that the simultaneity of two events that occur, was not only dependent on the positions of two observers, but also on their *relative* velocity. That an observer near one event would see that event before a more distant one, did not take an Einstein to know — being rather blatantly obvious. The German name for the theory was "the relativity of *simultaneity*". In fact, the theory goes on to discuss those quantities that *do not* depend on relative motion. This is discussed more fully in my book on Special Relativity [1]. I will not review the Special Theory here, but do need to contrast the views in it, rather than the results, with Newton's views.

In Newton's view of the Universe, space and time are "absolute" entities in themselves. Space exists, whether it is occupied or not; whether anybody sees it or not; whether the person seeing it is moving or not — it just "is". This ran counter to the usual thinking based on Aristotle's metaphysics. To make sense of this belief, Newton invoked the existence of God as a universal observer. Despite the fact that this thinking got ingrained into us, if one thinks about it afresh, it *does* seem strange — what is meant by the existence of nothing? Newton also assumed that "time flows at a constant rate". Again, this (now) trite observation contains in it the question of what is meant by "the flow of

time", as if it were a stream flowing? When a particle in a stream is seen to move some distance in a unit of time at one stage, and a different distance at another stage, we say that the rate of flow of the stream has changed. If it is not seen to change, we say that the rate of flow is constant. How can "the rate of flow of time", then, mean *anything*? All Einstein did was to challenge these mystical beliefs, and replace them by assumptions relating to actual observation of physical quantities in some (thought) experiment. In this sense Einstein only cleared up confusion caused by unnecessary assumptions.

The unfortunate name of the theory led people to take it that Einstein had somehow argued that *everything* — even ethical values — is relative. Since Relativity was regarded as "scientifically proved", it was claimed that all certainty in life and reality was lost. He was regarded as a new Shakespearian Prospero who had made the World tempestuous saying "We are such stuff as dreams are made on" (The Tempest Act 4, Scene 1). Ironically, it was the Quantum Theory that actually destroyed Victorian certainty, by saying that all physical predictions are only probabilistic and not deterministic. Probability was already used for Statistical Mechanics, but only as a way of getting approximate results for something that could be known more precisely in principle. Quantum Theory insisted that it *could not* be known. Though Einstein had been one of its founders, he strongly disagreed with this probabilistic interpretation of the theory. Nevertheless, to an epitaph for Newton:
'Nature and her laws lay hid in night,
God said "Let Newton be!" − and all was light!',
someone added the couplet for Einstein:
'But not for long, the Devil howling "Ho!
Let Einstein be!", restored the status quo.'
That couplet would have applied better to Niels Bohr and Werner Heisenberg, who had pioneered the view of an inherent probability in the laws of Nature. But I suppose, "Let Bohr and Heisenberg be", would not go down that well, as it loses the meter.

Actually, Einstein was very clear that accelerated motion is *not relative*, in that it does not depend on the velocity of the observer. He talked of this by giving the example that if a train is moving smoothly a passenger will feel nothing but if the train speeds up or slows down, the passenger will be pushed back or to the front. It must have taken a lot of imagination for Einstein to think of a train of those days moving smoothly. But then, there were no planes in those days. (The Wright brothers had taken their maiden flight but that was pretty well all.) Galileo had a better example for the relativity of uniform linear motion with a boat moving in a calm sea, which was presumably modified by Einstein to a more "modern" example for the times. My point is that Einstein had realized that accelerated motion would be detectable from within a closed laboratory. This led him to focus on another mystical belief of Newton's view — to do with his law of gravity. Newton's view was that the force of gravity of a mass is instantaneously felt at a distance. Thus, if the Sun were to suddenly disappear, the Earth would be released from its orbit and go flying off at a tangent. Just imagine: you are seeing the Sun in the sky, and then suddenly, 8 minutes 20 seconds later, you see it go shooting off and disappear. (Bear in mind that it would take light that long to reach the Earth from the Sun.) Of course, that is an absurd example, as the Sun could not suddenly cease to exist, so one may not bother about it. However, what if the Sun were just accelerated

away? How could the information reach the Earth instantaneously, as nothing can go faster than light? There must be gravitational disturbances that travel at the speed of light — gravitational waves!

Einstein noted one other point. In Newton's laws the mass appears in two ways: in the second law of motion as inertia; and in the law of gravitation as a sort of gravitational charge. Attempts had been made to try to find a difference between the two, but (as I shall be mentioning shortly) had given a null result. Einstein realized that this could not be accidental, they must really *be the same*. He, therefore, stated "the principle of equivalence" that the gravitational and inertial masses are identical. That means that at one point one cannot distinguish between a gravitational effect and the effect of acceleration. In modern terms, if one is in a closed laboratory, it will not be possible to tell, by any experiment, if the laboratory is being accelerated by a rocket or being held up against the pull of the gravity of a planet. Conversely, if the laboratory is in a lift, by cutting the cable holding the lift up and letting the lift fall freely, we will have "switched off gravity". He later described this realization as "the happiest thought of my life". Imagine him pondering this matter with soft music playing in the background. As he ponders and comes closer to the realization, the music speeds up and grows in volume. Then, when he is struck by *the thought*, the music reaches a crescendo with a clashing of cymbals. General Relativity has arrived!

Unfortunately, the physical theory required some mathematics that Einstein had failed to pick up in his stay at the ETH Polytechnic at Zurich − namely Geometry. Herman Minkowski had taught a course on the subject that Einstein had studiously bunked. When Minkowski had recast Einstein's Special Theory in geometric terms as kinematics in a 4−d spacetime, Einstein had rejected the development, saying that it lost the physical understanding and obfuscated it in mathematics. Now, when he needed to make a workable theory from "the happiest thought of his life", he did not know how to do so. He went to his friend Marcel Grossmann, to teach him the required mathematics. After various failed attempts using other types of Geometry, such as Affine Geometry and Teleparallelism, they hit on Differential Geometry as the language for the theory. Two papers were published by them in 1913 and 1914 [2, 3], in which the theory was almost fully formulated using the (earlier) much hated Differential Geometry.

At the time most theoretical physicists used the Euler-Lagrange (EL) equations generalized to deal with the fields of James Clerk Maxwell. What remained for the theory was to formulate it in these terms, rather than what was regarded as the unfamiliar language of Geometry used in the Einstein-Grossmann papers. Einstein had been in correspondence with David Hilbert, who was abreast with Einstein's work so far. In 1915, the two of them independently took the next step of the field-theoretic formulation, in which a correction of the earlier papers was given. There were four by Einstein and one by Hilbert published in 1915 [4, 5]. The paper in which all errors were removed by Einstein appeared in 1916 [6]. Hilbert acknowledged Einstein as the originator of the theory in his paper, claiming only to axiomatize the foundations of Physics, but people claim priority for his work because Einstein's paper appeared (a bit) later. (Perhaps, Hilbert and the claimants being Christians, and Einstein being a Jew, had something to do with it. Being a Muslim, I can be objective, as no Muslim ever came close to contributing anything to Relativity till *very* much later.) It *is* Einstein's General

Theory of Relativity.

While there are many, and varied, successes of both SR and General Relativity (GR), there are problems with this theory. The most glaring is the fact that it singles out Gravity from the other fundamental forces of Nature. Einstein's "happiest thought" used Gravity to generalize the theory of uniform linear motion to arbitrary motion. In SR, forces were dealt with ignoring the philosophical problem that the force resulted in acceleration, which did not allow velocities to remain constant. One could argue that it was only the object that was accelerated and not the observer. However, by the spirit of Relativity, the object is an equally good observer. One could say that in the SR view *all forces are equally "bad"*. By using Gravity, arbitrary motion is incorporated, *but only Gravity is "good" and the other forces remain "bad"*. (It reminds one of George Orwell's *Animal Farm* where an animal revolution was started on the slogan that "All animals are equal". Then the pigs take over the revolution and the slogan is modified to "All animals are equal − but some are more equal than others". All forces are equal, but Gravity is the most equal force.) This deficiency bothered Einstein. At the time the only other fundamental force known was Electromagnetism, and he tried to extend GR to incorporate this force in a Unified Field Theory. Soon afterwards, Hilbert tried and then others joined in the attempt. Despite various claims there has been no philosophically satisfactory attempt that is free of problems. Of course, soon afterwards it became clear that there was another fundamental force responsible for the decay of heavy atomic nuclei and then one to hold the nucleus together. These are called the "weak" and the "strong" nuclear forces. Einstein never believed that they were fundamental and hoped to demonstrate that they were "effective forces" that were approximate descriptions of the interaction of Gravity and Electromagnetism.

Another problem was the relationship of Relativity to the Quantum Theory. Paul Maurice Dirac had developed a procedure to convert a classical field theory to a quantum version, the so-called "quantization of the classical field". This method proved extremely successful for the quantization of the electromagnetic field, leading to Quantum Electrodynamics (QED). When he attempted the quantization of the gravitational field, Dirac obtained meaningless answers. He was ready to ascribe them to the same cause as the meaningless answers provided for QED, yielding infinite probabilities. However, others managed to address the issue for QED and obtain correct answers by making the infinities irrelevant. The Quantum methods were applied to the nuclear forces to provide a beautifully elegant way of dealing with them as fundamental forces, and the procedure for rendering the infinities harmless worked well for them. Abdus Salam, Sheldon Glashow and Steven Weinberg managed to provide a unified theory of the electromagnetic and weak nuclear forces to a single "electro-weak" force. This force was compatible with the strong nuclear force, so that the three can be put together as "the Standard Model" of Particle Physics. There have been attempts to provide a "Grand Unified Theory" of the three forces put together as a neat whole package, but there are problems with the attempts that I will not go into here. However, the Quantum methods failed when applied to gravity and it was shown that they were always doomed to fail. There seems to be a much deeper tension between the two theories that precludes their marriage.

Despite its deficiencies, and all its difficulties, Relativity is "the only game in town" to provide answers for questions involving gravity alone. Further, it is not

as if one cannot get answers when other forces interact with gravity, but only that the methods are not *philosophically* satisfactory. Since both fundamental theories have philosophical problems, the hope is that when a correct theory arrives it will resolve all the problems. People on one end of the spectrum expect that one or more of he tenets of Relativity will need to be altered to make it compatible with the Quantum Theory; and those on the other end expect that Quantum Theory is the one that needs modification. Many have been claiming that Superstring Theory, or one of its derivatives like "brane theory" or "M-theory", will be "the Holy Grail". At the other end, many people follow Einstein's belief that Quantum Theory is flawed and when it is "corrected", or "completed", the true theory will be found. Generally, those who come to the problem from the Quantum side are of the former type and those who come from Relativity are of the latter type. Probably, both theories will need to be modified.

Almost immediately after his first complete formulation of GR, in 1916 he demonstrated that the theory necessarily required the existence of gravitational waves that travel at the speed of light [7]. The problem with detecting them is that they are about 10^{38} (a hundred trillion, trillion, trillion) times weaker than electromagnetism. With the recent discovery of gravitational waves in 2016, this prediction was verified a century after it was made! That seems to me to be something of a record.

Very soon after his final formulation of GR, in 1917 Einstein [8] tried to apply it to the Universe as a whole. At that time such discussion fell into the realms of Theology. Some monks had given physical arguments in favour of their favourite theological cosmology, but there was no interest in them among the physicists. Of course, physicists had always held their own religious beliefs but, since the time that Simon Laplace, had said "[Sire,] je n'ai pas eu besoin de cette hypothèse" ("[No, Sire,] I had no need of that hypothesis"), when Napoleon pointed out that there was no mention of God in his book on the Heavens, they did not bring it into their Physics. Einstein had obtained tests of GR that only gave very fine corrections. He probably wanted a situation where "GR would rule". As it transpired, this was a very fruitful line of enquiry and much work followed from it. The need for precise testing of Einstein's theories has not only contributed to our understanding of the physical world around us, it has driven technological development, leading to developments of telescopes and of laser interferometers in space. It used to be said that QED is the most precisely tested theory in Physics. GR has, since joined it.

Considering the wide variety of areas of Physics that Einstein made seminal contributions in, it is difficult to keep track of all that he did relevant to the development of SR and GR. I will not try to cite all his relevant papers, preferring to refer to two excellent biographies of Einstein and his work [11, 12] and the list of Scientific Publications of Albert Einstein on Wikipedia.

1.1 The Equivalence of Gravitational and Inertial Masses

It had already been noticed, by the nineteenth century, that the term "mass" appears in two separate contexts in Mechanics, with no reason to regard the term as identical in both contexts. One is the resistance offered to a force that tends to accelerate an object, called the "inertial mass". The other is as

a gravitational analogue of the electromagnetic charge, a sort of "gravitational charge", called the "gravitational mass". Despite the formal similarity between the gravitational and electromagnetic force laws there are three fundamental, physical points of difference between them. First, the gravitational charge is always positive (giving an attractive force only) while there are three possible types of charge, namely positive, negative and neutral (giving attractive, repulsive and zero forces). Second, in gravity like charges attract instead of repelling as in electromagnetism. Correspondingly the field intensity and hence the potential of gravitation has the opposite sign to electromagnetism. This point had been noted by Maxwell as a serious hurdle to extending his ideas for electromagnetism to gravitation. This means that the energy associated with gravity should be negative, if the energy associated with electromagnetism is taken to be positive. Third, the electric charge is quantized, in that it occurs in multiples of a third of the electron charge, while there does not seem to be a corresponding quantization of the gravitational charge. Any attempt to unify the forces of nature and Relativity with Quantum Field Theory must take these points into account. However, for our present purposes these differences are not so important. What is important is that there is no *a priori* reason why the inertial and gravitational masses *should* be identical.

An experiment to test the identity between the inertial and gravitational masses was performed by Baron Eötvös starting in 1886 (and going on to 1909). As with the Michelson-Morley experiment it was a crucial null experiment. Unlike that experiment the null result was expected and so there was no resistance to accepting its result. Only Einstein seems to have realized its significance. Since this was the only *fact* − the only piece of experimental evidence − on which GR is based, it is worth a more detailed discussion. It holds the same position for GR that the consistency of the speed of light does for SR. In fact, it is the basis of Einstein's "happiest thought of [his] life". In basing his new, unrestricted, theory on this fact, he elevated it to the status of a principle, calling it the "principle of equivalence". This is the crucial step in separating gravity from the other forces. Henceforth unifying gravity with the other forces would be something of a non-sequitur − at least if we require GR to hold. Einstein never seems to have realized that his "happiest thought" had turned into his "saddest thought" in his subsequent attempts to unify gravity and electromagnetism. It may be that this problem is at the base of those attempts at "quantizing gravity" that also unify unify the fundamental forces by modifying GR without touching Quantum theory. Because of its importance, this principle has gone on being tested to this day with the equivalence being maintained.

To try to distinguish between the gravitational and inertial masses it is necessary to allow the same body to experience a gravitational and an inertial force simultaneously. Let the gravitational and inertial masses be denoted by M and m respectively and the corresponding accelerations by $\mathbf{g}$ and $\mathbf{a}$ respectively. The net force on the body will then be $M\mathbf{g} + m\mathbf{a}$. If the forces are parallel or anti-parallel $\mathbf{a} \propto \mathbf{g}$ or $\mathbf{a} = k\mathbf{g}$, where k is a constant which may be positive or negative. The total force will then be $(M + km)\,\mathbf{g}$. Thus we will only be able to measure the combined, effective, mass $(M + km)$ and not be able to resolve it into the two separate masses. However, if the forces are at some oblique angle we will have two components with different combinations of M and m. We could use them to solve a pair of simultaneous equations in M and m so that we could distinguish between them. For this purpose it is necessary, therefore,

that the inertial force should not act straight up or down. The most convenient inertial force is the centrifugal force due to the Earth's rotation. Ofcourse, as we increase the magnitude of this force (by moving towards the equator) we decrease the angle and as we increase the angle (by moving towards a pole) we decrease the magnitude.

It is never as easy to measure the absolute values of two quantities as to measure small differences between them. The accuracy is much greater in the latter case. An example of this basic fact of experimentation is the relative ease with which the Michelson interferometer can measure path differences $\sim 10^3$ (or 10^{-7}m) while the path length can be measured with an accuracy of only $\sim$ 1mm (or 10^{-3}m). This fact is also repeatedly encountered in making Astronomical and Cosmological measurements, which I will not go into here. Eötvös also used this principle. He looked for a difference between the ratio $\alpha = M/m$ for different bodies. If the ratio is the same for all bodies it can be chosen, by appropriate choice of units, to be unity and hence gravitational and inertial masses will be identical. Since it is known that the two types of mass are more or less the same we know that α is close to unity for all bodies. Eötvös tried to measure the small difference $|1 - \alpha|$.

Notice that for the rotating Earth the magnitude, a of $\mathbf{a}$ (we will denote the magnitude of a vector by the same letter but not in boldface) is given by

$$a = R\omega^2 \cos\theta \ , \tag{1.1}$$

where R is the Earth's radius ($\approx 6.4 \times 10^8 m$), ω the angular frequency of the Earth's rotation ($\approx 7.3 \times 10^{-}5 rad/sec$) and θ is the latitude where the experiment is performed (being the complement of the usual polar angle in spherical polar coordinates in the Northern hemisphere). For a to be large θ must be small. However, as pointed out above, the quantity to be measured will be large when there is a large difference between the direction of $\mathbf{a}$ and $\mathbf{g}$. For concreteness take θ to be 30^o, $a \approx 0.3$ m/sec^2 while $g \approx 9.8$ m/sec^2. Thus, typically, $a \approx 3\%$ of g, which is a sizable value.

The experimental apparatus consisted of a rod, which can be represented by a vector $\mathbf{b}$ with two dense bodies attached, which we denote by "1" and "2". Rotating the rod through π radians interchanges "1" and "2" and reverses the vector $\mathbf{b}$, i.e. changes $\mathbf{b}$ to -$\mathbf{b}$. Denoting the gravitational and inertial masses of the two bodies by the previous symbols with the corresponding subscripts, the net torque due to the two forces (gravitational and centrifugal) is

$$\mathbf{T} = \mathbf{b} \times [(\mathbf{M_1} - \mathbf{M_2})\mathbf{g} + (\mathbf{m_1} - \mathbf{m_2})\mathbf{a}] \ , \tag{1.2}$$

(where $\times$ represents the usual "cross product" of vectors) while the resultant of the two forces on both bodies is

$$\mathbf{F} = (M_1 + M_2)\mathbf{g} + (m_1 + m_2)\mathbf{a} \ . \tag{1.3}$$

Thus the effective torque, which is the quantity that will actually be measurable, will be the component parallel to the resultant force (see Fig. 1.1)

$$\begin{aligned} T_{\|} &= \frac{\mathbf{T}\cdot\mathbf{F}}{|\mathbf{F}|} \\ &= \frac{\mathbf{b}\wedge[(M_1-M_2)\mathbf{g}+(m_1-m_2)\mathbf{a}]\cdot[(M_1+M_2)\mathbf{g}+(m_1+m_2)\mathbf{a}]}{2[\{(M_1+M_2)\mathbf{g}+(m_1+m_2)\mathbf{a}\}\cdot\{(M_1+M_2)\mathbf{g}+(m_1+m_2)\mathbf{a}\}]^{1/2}} \ . \end{aligned} \tag{1.4}$$

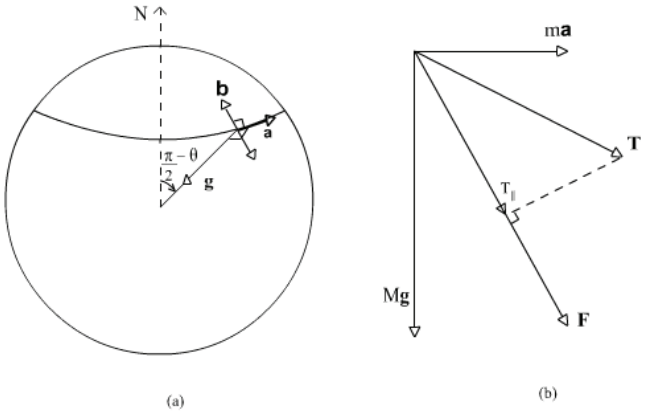


Figure 1.1: The Eötvos experiment. (a) At a point on the Earth's surface, at latitude θ a rod of length b is placed horizontally in the North-South direction. The centrifugal acceleration, a, is clearly orthogonal to the rod, b. (b) The resultant force, $F = Mg + ma$ does not act straight down. The component of the torque $\mathbf{T}$ along this force, $T_{||}$, is measured.

Now $\mathbf{g}.\mathbf{g} \sim 100, 2\mathbf{g}.\mathbf{a} \sim 5, \mathbf{a}, \mathbf{a} \sim 0.1$, in units of $(\mathrm{m/sec}^2)^2$. Thus the denominator may be approximated by $2(M_1 + M_2)g$, as the total gravitational mass is more or less the same as the total inertial mass, even if there is some slight difference between them. In the numerator, we can change the order of the scalar triple product so that the cross appears between the second and third terms. Since $\mathbf{g} \times \mathbf{g} = 0 = \mathbf{a} \times \mathbf{a}$, we are left only with a $\mathbf{g} \times \mathbf{a}$ term with coefficient $(M_1 - M_2)(m_1 + m_2) - (M_1 + M_2)(m_1 - m_2) = M_1 m_2 - M_2 m_1$. The scalar triple product $\mathbf{b}.\mathbf{g} \times \mathbf{a}$ will be maximum if $\mathbf{b} \perp \mathbf{g}, \mathbf{a}$. Setting this as the orientation of $\mathbf{b}$ and writing Eq.(4) in terms of α,

$$T_{||} \approx \frac{m_1 m_2(\alpha_1 - \alpha_2)}{m_1\alpha_1 + m_2\alpha_2} bR\omega^2 \sin\theta\cos\theta \ . \tag{1.5}$$

Clearly, the maximum value of $T_{||}$ will be at $\theta \approx \pi/4$ rad (or 45^o latitude). Taking $b = 1$m, $m_1 = m_2 = 1$kg, the effective torque (in mks units) is given by

$$T_{||} \approx 0.017(\alpha_1 - \alpha_2)/(\alpha_1 + \alpha_2) \approx 0.0085(\alpha_1 - \alpha_2) \ . \tag{1.6}$$

The torsion is actually measured by rotating the rod through π radians and observing the deflection of a spot of light reflected from a mirror attached to the wire whose torsion is being measured. Thus $2T_{||}$ is measured. Eötvös found that $|\alpha_1 - \alpha_2| < 10^{-8}$. Later experiments have raised this accuracy to $\sim 10^{-14}$. Even now, there remain attempts to introduce scalar fields (or even vector fields), which will lead to a fifth force, and argue that the effect is so small that it has not yet shown up.

There had been claims of an observed difference attributed to a "fifth force". Despite the great excitement generated at the time there was no satisfactory evidence for it, as the various claims were mutually contradictory [79] and the claims were later withdrawn.

1.2 Field Theory

A crucial aspect of the development of General Relativity is that it is expressed in the language of field theory. Many of the failed attempts and competing theories were also field theoretic. Though I prefer to develop it as a theory of motion, to most people it is merely a field theory of gravity. I will, therefore, very briefly present the essentials of the subject of Field Theory here and leave the main part of the discussion to Chapter 4. For a more detailed discussion the reader is referred to Landau and Lifshitz (LL), *The Classical Theory of Fields* [14], and to the relativists' Bible, "Gravitation" [15] by Misner, Thorne and Wheeler (MTW). The field theory*par excellence* is Maxwell's theory of the electromagnetic field. Its remarkable power, elegance, success and simplicity have led to its current status of "role model". Maxwell had considered the possibility of extending field theory to incorporate gravity but concluded that it was impossible. As mentioned earlier, Einstein and others made various false starts at a field theory of gravity till the final formulation of Einstein and Hilbert. One of the most successful attempts to unify the two forces was made in two parts by Kalutza and Klein (which will be briefly discussed in Chapter 5), but it had problems. While Einstein was very enthusiastic about it at first, he later rejected it.

More recently the development of field theory has received fresh impetus from considerations of symmetry of the field in some context. By "symmetry" is meant that the field is invariant under some transformations. If these transformations form a group (as they generally do) the theory can be expressed in the form of the symmetry group. In 1918 Emmy Noether stated a theorem according to which, for each generator of the group there will be a conserved quantity (a "charge"). Of particular interest are transformations which modify the potential functions without altering the physical quantities. These are called *gauge* transformations. If the invariance is only under global transformations we have a not-so-interesting "global gauge symmetry". To illustrate the difference between the two, consider the rotation of an irregular object. A rotation through 2π radians will leave it invariant but through any other angle will change it. This is a global symmetry. There can be global symmetry of rotation through π radians. For example an ellipse when rotated about its centre through π radians is left invariant. A very much stronger symmetry is provided by a circle. Rotation through any angle, about its centre, leaves the circle invariant. This is a local symmetry. Local symmetry implies global symmetry but the converse is not true.

As a historical aside, it is of interest to note that despite her stupendous contributions in Mathematics and Physics, Emmy Noether could not be employed in a University in Germany on account of being a female. She had to teach totally uninterested (and uninteresting) students in a finishing School for "young ladies". It took Hilbert's efforts for her to be allowed to teach at the University of Göttingen.

The reason why these considerations became interesting is that they lead to non-trivial generalizations. For the above example of a local symmetry any two transformations will always commute. The group of rotations in two real dimensions, $SO(2)$, is Abelian. Similarly the group of unitary transformations in one complex dimension, $U(1)$, is Abelian. Now consider the symmetries of a sphere. It is invariant under rotations about any axis passing through its centre.

There are three independent generators. In general two arbitrary rotations will not commute, as anyone who has played with a Rubik's Cube, or Rubik's Revenge, will bear witness to. The group of rotations in three real dimensions, $SO(3)$, is non-Abelian. Similarly, the group of unimodular (with determinant 1), unitary transformations in two complex dimensions, $SU(2)$, is non-Abelian. (You may wonder why there is no $SU(1)$. The reason is that the generator of $U(1)$ is simply a phase, or complex number with magnitude 1. Hence its determinant is a phase. Making the phase angle 0 reduces to just the number 1.) Further, the symmetry group of transformations of the Euclidean plane into itself, E_2, is non-Abelian. Non-abelian gauge theories were first considered by Yang and Mills and the first example considered was $SU(2)$. Yang-Mills fields were used by Glashow, Salam and Weinberg to construct the unified electro-weak theory $SU(2)_W \otimes U(1)_Y$. Similarly, strong force has the symmetry group $SU_C(3)$. This gives the standard model symmetry group $SU_C(3) \otimes SU(2)_W \otimes U(1)_Y$. These types of considerations led to Supergravity theory, conformal field theory and then to Superstring theory. We will not go further into any of these developments.

1.3 The Lagrange Equations

In the early days Mechanics was developed to be able to predict the motion of all the bodies of the solar system known at the time. Solving for a planet in the field of an infinite mass Sun, is trivial and gives a wrong result, as the Sun does not have infinite mass. The method to correct for the finite mass was already developed by Newton in his *Principia*. In fact, Robert Hooke had proposed to him, the inverse square law for the force pulling the planets towards the Sun and Newton had generalized the idea to his law of universal gravitation, so that the planet would pull the Sun as well. For the purpose, one breaks the motion into two parts: one for the centre of mass and the other for each body orbiting about the centre of mass. However, for three bodies the 3 coupled differential equations could not be solved so simply. Lagrange developed the method of minimizing the "free energy", the difference between the kinetic and potential energies (called the Lagrangian), $L[q^i(t), \dot{q}^i(t)] = T[\dot{q}^i(t)] - V[q^i(t)]$, where $q^i(t)$ and $\dot{q}^i(t)$ $(i = 1...n)$ are the generalized positions and velocities for N particles subject to m constraints, $n = 3N - m$, as functions of time, so that L is a *functional*. A functional may have a constant value, but depend non-trivially on a function, or functions. Thus we can look for the form of the functions which give a minimum or maximum value of the functional. This is what we will be doing with the Lagrangian.

Hamilton, later, re-formulated Lagrange's mechanics in terms of the generalized positions and momenta, $(q^i(t), p_i(t))$ and demonstrated that the integral of the Lagrangian over a finite time interval, called the action S, must be minimized to obtain the path of a particle. This is called *Hamilton's principle of least action*. He further showed that there was a conserved quantity, H, which is the *sum* of the kinetic and potential energies, corresponding to the Lagrangian. This total energy is a constant of the motion as a function but is non-trivial as a functional. With the former way of looking at it we get Lagrangian mechanics, yielding the Lagrange equations, and with the latter Hamiltonian mechanic, yielding the Hamilton equations.

Hamilton's principle that the action, S, be minimal requires that its variation be zero, i.e.

$$\delta S = \delta \int_a^b L(q^i, \dot{q}^i) dt \ , \tag{1.7}$$

where a and b are initial and final times and $q^i(a), q^i(b)$ are given constant values, so that $\delta q^i(a) = 0 = \delta q^i(b)$. The variation can be evaluated inside the integral sign to give

$$0 = \int_a^b [(\partial L/\partial q^i)\delta q^i + (\partial L/\partial \dot{q}^i)\delta \dot{q}^i] dt \ , \tag{1.8}$$

where the Einstein summation convention, that repeated indices are summed over, is used here (and throughout the book). It can be demonstrated that the operators δ and d/dt commute. Thus we can integrate the second term in the integral by parts, to obtain

$$0 = \int_a^b [\frac{\partial L}{\partial q^i} - \frac{d}{dt}\frac{\partial L}{\partial \dot{q}^i}]\delta q^i dt + \frac{\partial L}{\partial \dot{q}^i}\delta q^i \Big|_a^b \ . \tag{1.9}$$

The last term here is zero as $\delta q^i(a) = 0 = \delta q^i(b)$. Thus the expression in the integral must be zero. This will generally be true (for all a, b) only if the integrand is zero. This requirement gives the *Lagrange equations*

$$\frac{\partial L}{\partial q^i} = \frac{d}{dt}\frac{\partial L}{\partial \dot{q}^i} \ . \tag{1.10}$$

It should be borne in mind that these are *necessary*, but not *sufficient*, conditions for extrema to occur. The sufficient conditions would come from a second variation, that is seldom undertaken. There are various directions for generalizing this analysis. We have assumed that the Lagrangian has no explicit time dependence. If there were explicit time dependence there would be an extra term in the Lagrange equations. In this case energy would not be conserved. In such non-conservative systems the extra term corresponds to energy dissipation or creation. Another generalization allows the Lagrangian to depend on higher derivatives of the generalized coordinates. The above equations are second order differential equations. They would then become higher order equations of motion. There is no evidence that such a generalization is required in Physics. Yet another generalization is to deal with less rigid constraints on the generalized coordinates and to leave one, or both, of the ends free. I will not discuss any of these extensions here as they are not at all relevant for a discussion of Relativity.

1.4 Extension of the Lagrange Equations to Fields

For a very large number of particles with few constraints, i.e. very large n, it becomes convenient to take the limit $n \to \infty$. Correctly speaking, we should take the infinity to be countable (like the natural numbers), but so as to be able to use calculus, we take the continuum limit. Thus every point has a different value of $q(t)$. We thus replace the generalized coordinates by a *field*, $\phi(t, \mathbf{x})$, where $\mathbf{x}$ has replaced the index label i. Thus the field is a function of time, at

each position vector $\mathbf{x}$, and hence is a function of position as well. Of course, one can generalize so that $\mathbf{x}$ may be a lower or higher dimensional vector than 3. We would, correspondingly have a lower or higher dimensional field theory. There is a complication that arises, here, in regard to the Lagrangian. To explain that I will need to digress a little bit on the theory of cardinal numbers.

The cardinal number is the transfinite extension of the "number of elements" of a finite set. By setting up one-to-one correspondences we can compare infinite sets. The cardinality of the natural numbers (the *counting* numbers) is denoted by $\aleph_0$, (read *aleph null*). A set with a one-to-one correspondence with this set is said to be *countable*. Now, by the theory of ordered sets it is known that he cardinality of the power set of a given set (the set of all its subsets) is strictly greater than the cardinality of the set, $|\exp(A)| > |A|$. It can be easily demonstrated that $|\exp(A)| = 2^{|A|}$. These statements are as true for transfinite as for finite sets. Thus $2^{\aleph_0} \equiv \aleph_1 > \aleph_0$. It can also be proved that the set of real numbers is uncountable and hence it has a cardinality greater than $\aleph_0$. From Gödel's theorem it can be shown that it is possible to choose $\aleph_1$ to be the cardinality of the set of real numbers. This choice is known as the continuum hypothesis and will be adopted henceforth. (It is possible to choose otherwise and develop a different transfinite Mathematics but we will not go into that here.)

Taking the continuum hypothesis the cardinality of the space of all functions is $\aleph_2 \equiv 2^{\aleph_1}$, as that is the set of all subsets of $\mathbb{R}$, the set of real numbers. The number of degrees of freedom of a system of N particles subject to m constraints is $n = 3N - m$. For a continuum the number of degrees of freedom is the continuous infinity, $\aleph_1$. The Lagrangian for a system of N function, for which the usual differential calculus can be used. When we deal with fields, the Lagrangian becomes a *functional* of the system, with infinitely many degrees of freedom. Replacing $q^i(t)$ by $\phi(t, \mathbf{x})$ we must replace $\dot{q}^i(t)$ by $\dot{\phi}(t, \mathbf{x})$. Notice that the dot refers to a total derivative and it not the partial derivative alone, $\dot{f} = df/dt = \partial f/\partial t + \mathbf{x}.\dot{\nabla} f$. The Lagrangian is then written as the functional $L[\phi(t, \mathbf{x}), \dot{\phi}(t, \mathbf{x})]$. The cardinality of the space of variables for the Lagrangian being higher, we can no longer use ordinary differential calculus. The differences between what is required and usual calculus are studied in *Functional Analysis*. I will not go into that here. However, to distinguish the *functional derivatives* from the ordinary derivative, I will follow the usual notation of replacing "∂" by "δ". For details the reader is referred to [16].

As before, we have taken it for granted that the Lagrangian has no explicit dependence on position or time. Its dependence only comes through the field, ϕ, and its time derivative, $\dot{\phi}$. Following the same procedure as before we arrive at the *EL equations*

$$\frac{\delta L}{\delta \phi} = \frac{d}{dt}\left(\frac{\delta L}{\delta \dot{\phi}}\right) . \tag{1.11}$$

It is worth mentioning, here, that the assumption that the Lagrangian has no explicit space or time dependence means that it is invariant under space and time translations. The conserved quantities given by Noether's theorem, in this case, are momentum and energy. Thus the above assumption is equivalent to the momentum and energy conservation laws!

1.5 Relativistic Fields

Field theory, as presented here up to now, does not accommodate Special Relativity. Time is given a special place. To be relativistic, a field theory must not refer to $\phi(t, \mathbf{x})$ and the equations must not end up with a d/dt operating on any quantity as Eq11 does. The field ϕ must be a function of the spacetime position vector, x^μ ($\mu = 0, 1, 2, 3$). Further, the only derivative available is the gradient of the field, $\phi_{,\mu} \equiv \partial\phi/\partial x^\mu$. Our Lagrangian will then have to be replaced by a *Lagrangian density*, $\mathcal{L}[\phi, \phi_{,\mu}]$. The formulation is completed by replacing Eq.(11) by

$$\delta S = 0 = \delta \int_a^b L dt = \delta \int_a^b \left(\int_V \mathcal{L} dV \right) dt = \delta \int_\Omega \mathcal{L} d\Omega \ . \tag{1.12}$$

Here $d\Omega$ is the "volume element" in spacetime and Ω is the total spacetime "volume" under consideration (see Figure 1.2). In Minkowski space, using Cartesian coordinates,

$$d\Omega = dx^0 dx^1 dx^2 dx^3 \ . \tag{1.13}$$

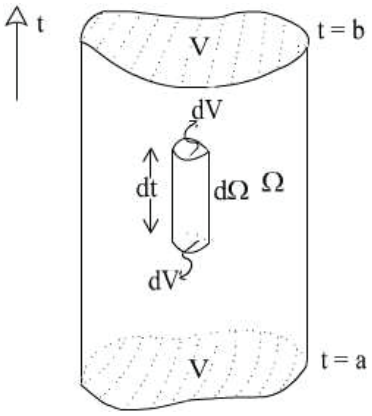


Figure 1.2: The "wotld-tube" represented by a 3-dimensional cylinder in 4-dimensional spacetime. The top and bottom "faces" are regions of 3-volume V at two times, $t = a$ and $t = b$. Inside this cylinder is a small 4-volume element.

A more general formulation will be presented later. Since that requires tensors and refers to concepts in curved spacetimes we will not go into it here. The three dimensional volume, V, traces, out a "world tube" in four dimensional spacetime.

We are now in a position to extend the EL equations to relativistic fields. We have

$$\delta\mathcal{L}[\phi, \phi_{,\mu}] = \frac{\delta\mathcal{L}}{\delta\phi}\delta\phi + \frac{\delta\mathcal{L}}{\delta\phi_{,\mu}}\delta\phi_{,\mu} \ . \tag{1.14}$$

We will need to integrate the second term by parts. Since this is not such a familiar procedure as before it needs to be elaborated a bit. For this purpose

we write the second term as

$$\left.\begin{aligned}\frac{\delta\mathcal{L}}{\delta\phi_{,\mu}}\delta\phi_{,\mu} =& \frac{\delta\mathcal{L}}{\delta\phi_{,0}}\delta\phi_{,0}\ \delta\left(\frac{\partial\phi}{\partial x^0}\right) + \frac{\partial\mathcal{L}}{\delta\phi_{,1}}\ \delta\left(\frac{\partial\phi}{\partial x^1}\right) + \frac{\partial\mathcal{L}}{\delta\phi_{,2}}\ \delta\left(\frac{\delta\mathcal{L}}{\partial x^2}\right)\\ &+ \frac{\delta\mathcal{L}}{\partial\phi_{,3}}\ \delta\left(\frac{\partial\phi}{\partial x^3}\right)\ .\end{aligned}\right\} \tag{1.15}$$

To evaluate the integral in Eq. (12), we need to integrate all four of the terms by parts, using the volume element given in Eq.(13). Let us just consider the first of the four terms. Integrating the term with respect to x^0 and using the fact that the partial derivative and the δ commute, the second expression can be treated as the function to be integrated and the first to be differentiated. Thus

$$\left.\begin{aligned}\int_\Omega \frac{\delta\mathcal{L}}{\delta\phi_{,0}}\delta(\frac{\partial\phi}{\partial x^0})(dx^3dx^2dx^1)dx^0 =& \int_V \frac{\delta\mathcal{L}}{\delta\phi_{,0}}(dx^3dx^2dx^1)\delta\phi|_a^b\\ &- (\frac{\delta\mathcal{L}}{\delta\phi_{,0}})_{,0}\delta\phi d\Omega\ ,\end{aligned}\right\} \tag{1.16}$$

where a is the initial time and b the final time. Since ϕ is fixed on the boundary, so $\delta\phi$ is zero at either end, and hence the first term vanishes. Doing the same for each of the other terms, it is obvious that Eq. (112) becomes

$$0 = \int_\Omega \left[\frac{\delta\mathcal{L}}{\delta\phi} - \left(\frac{\delta\mathcal{L}}{\delta\phi_{,\mu}}\right)_{,\mu}\right]\delta\phi d\Omega\ . \tag{1.17}$$

Since this integral is zero for arbitrary $\delta\phi$, the integrand must be zero. Hence we get the *EL equations*

$$\frac{\delta\mathcal{L}}{\delta\phi} = \left(\frac{\delta\mathcal{L}}{\delta\phi_{,\mu}}\right)_{,\mu}\ . \tag{1.18}$$

This formulation can be extended to a vector valued field, ϕ_r $(r = 1, ..., k)$. The r need not be a spacetime index. However special interest attaches to the case when the field is, itself, Lorentz covariant. In this case r will be a spacetime vector index, μ, or a tensor index like $\mu\nu$. Also, the field will be invariant under the full Poincarè group.

Another possibility is that r may be compounded of a spacetime index (or indices) and be invariant under some other symmetry. This is the case for the Yang-Mills field. To explain this I will first remind the reader of the Maxwell field. It is given by the four-vector potential, A_μ, and the Maxwell field tensor is the generalized curl of this field (as given in SR)

$$F_{\mu\nu} = A_{\nu,\mu} - A_{\mu,\nu}\ . \tag{1.19}$$

Under the gauge transformation

$$A_\mu \to \tilde{A}_\mu = A_\mu + f(x^\nu)_{,\mu}\ , \tag{1.20}$$

the field tensor remains invariant. Thus $F_{\mu\nu}$ is invariant under a further symmetry, which happens to be $U(1)$. This is called an *internal symmetry.* The corresponding Lagrangian density is

$$\mathcal{L} = \frac{1}{16\pi}(F_{\mu\nu}F^{\mu\nu} + j_\mu A^\mu)\ , \tag{1.21}$$

where j_μ is the four-vector current density.

In the case of the Yang-Mills field we have A^a_μ, where a is the index for the additional symmetry. Let $\mathbf{X}_a$ be the generators of the algebra of the additional symmetry group and let the algebra be given by

$$[\mathbf{X}_a, \mathbf{X}_b] = C_{ab}{}^c \mathbf{X}_c \ , \quad (a, b, c = 1, ..., p) \tag{1.22}$$

where $C_{ab}{}^c = -C_{ba}{}^c$ are the structure constants. Then the Yang-Mills field tensor is given by

$$F^a_{\mu\nu} = A^a_{\mu,\nu} - A^a_{\mu,\nu} + C_{bc}{}^a [A^b_\mu, A^c_\nu] \ , \tag{1.23}$$

where the brackets denote "commutator brackets" for operators

$$[\hat{P}, \hat{Q}] = \hat{P}\hat{Q} - \hat{Q}\hat{P} \ . \tag{1.24}$$

The Lagrangian density in this case is simply

$$\mathcal{L}_{YM} = -\frac{1}{16\pi}(F^a_{\mu\nu} F^{\mu\nu}_a + j^\mu_a A^a_\mu) \ . \tag{1.25}$$

1.6 Unsuccessful Attempts to Generalize Special Relativity

Through the period 1911-1915 Einstein (and others) made several unsuccessful attempts at developing an un-restricted, or general, theory of Relativity. The essential point had already been stated in 1907 with the principle of equivalence. This was the insight that accelerations may be incorporated into the theory of Relativity by introducing gravity. The fixed observer in a gravitational field is non-inertial while the freely falling observer is inertial. By virtue of the principle of equivalence (discussed earlier) gravity can be "switched off" by going from the non-inertial to the inertial frame. i.e. by going from the fixed to the freely falling frame. This idea was to ultimately provide a handle for implementing the generalization of the principle of Special Relativity (which will be discussed in the next section).

The first attempt at the required theory, made in 1911, was a scalar field theory in which the speed of light played the role of the gravitational field. The idea was that the Special Relativistic red-shift (and the corresponding time delay of signals) in the presence of a gravitational field, could be regarded as a reduction of light speed there. Thus the speed of light could be used as a measure of the gravitational field. Though this extension was formulated in special relativistic terms, it made the speed of light variable, violating the principle of the constancy of the speed of light. If taken seriously, it *could not* be relativistic. An alternate way of regarding it would be as changing the units of measurement of some other quantity, but it would then be physically equivalent (explained more fully in the next section) to a non-relativistic theory. Though it and was ultimately abandoned. Einstein pursued it till as late as 1913. This scalar field theory was limited to static gravitational fields. In 1912 Abraham tried to extend it to non-static fields. Einstein tried to go further with his flat spacetime scalar field theory, but with hindsight one can see that such an attempt would be foredoomed to failure.

Also in 1912 Einstein made a very important point. He demonstrated that acceleration could modify the geometry so that it was no longer Minkowskian. He considered a uniformly rotating circle. The circumference gets Lorentz contracted while the diameter does not (see SR or Einstein's "The Meaning of Relativity", which is available on the net for free). Thus the ratio of circumference to the diameter is no longer π and the spatial geometry is, therefore, non-Euclidean. Hence acceleration alters the geometry! This is the first time that curved spacetime geometries enter into Physics.

In 1913, in collaboration with Marcel Grossmann, Einstein proposed that the gravitational field be given by the metric tensor, $g_{\mu\nu}$, which enters into the definition of the proper time "line element"

$$ds^2 = g_{\mu\nu}(x^\rho)dx^\mu dx^\nu \ . \tag{1.26}$$

The metric coefficients are arbitrary functions of the spacetime position, subject to the requirement that $g_{\mu\nu}$ is a symmetric tensor. In the absence of any gravitational effects the line element given by Eq.(26) must reduce to the Minkowski line element. In that case the components of the metric tensor, in Cartesian coordinates, are

$$g_{00} = -g_{11} = -g_{22} = -g_{33} = 1 \ , \ g_{\mu\nu} = 0 \ (\mu \neq \nu) \ , \tag{1.27}$$

or in spherical polar coordinates

$$g_{00} = -g_{11} = 1 \ , \ g_{22} = -r^2 \ , \ g_{33} = -r^2 \sin^2\theta \ , \ g_{\mu\nu} = 0 \ (\mu \neq \nu) \ . \tag{1.28}$$

In this paper curved spacetime tensor analysis, which is the language for General Relativity, enters into Physics for the first time. However, Einstein and Grossmann did not arrive at the correct Lagrangian and gave fallacious arguments why tensorial equations were not applicable.

While Einstein's scalar field theory of 1911 was his first attempt, it is not as if there had been no attempt at a field theory of gravitation after Maxwell's (see section 1). In 1900 Lorentz had proposed a theory. In fact that theory was consistent with Special Relativity. Its only drawback was that it was totally unrealistic! Lorentz suggested that the attractive force between opposite electric charges may be marginally greater than the repulsive force between like charges. Thus, neutral matter composed of equal amounts of positive and negative charges, would attract other neutral matter. This residual attraction is what we call gravity. Since the inverse square law applies for electric charges and the constituents of matter are only microscopically separated, the residual force will obey the inverse square law over macroscopic distances. The reason why it is unrealistic is that there was absolutely no experimental evidence for the required asymmetry between electrical attraction and repulsion. By now the suggestion has been ruled out by experiment. Heaviside also speculated on developing a field theory of gravity by analogy with electromagnetism. Both of them ignored Maxwell's earlier objection (see section 1) that the analogy breaks down as gravitation gives a negative potential energy.

Nördstrom succeeded in writing down a Lagrangian which led to an internally consistent tensorial field equation in Minkowski space. Though unsatisfactory this was the only theory of gravity with any claims to acceptability in 1913. All other attempts were either incompatible with Special Relativity, or the principle of equivalence, or with experimental evidence. The next developments came in 1915.

1.7 The Principle of General Relativity

The principle of Special Relativity states that "all *inertial* frames are physically equivalent". Einstein sought a generalization in which the "inertial" could be dropped and accelerated frames incorporated into Relativity. The stumbling block is that we can *feel* acceleration. A more elaborate argument, due to Newton, attempts to prove that acceleration is absolute and not relative. Consider a bucket, containing water, suspended by a rope. The bucket is slowly twisted so that the rope is "wound up". When the water settles, its surface will be absolutely flat. Now, when the bucket is released the rope will unwind causing the bucket to rotate. The centrifugal force on the body of water will push the water outwards, causing the central region to be depressed and the edges to rise. Thus the surface of the water will be curved. This curvature is observable from *within* the bucket, without reference to anything outside. Since the rotation can be measured from within the rotating frame, it is absolute and not relative.

In the 1914 paper with Grossmann, using tensorial techniques, Einstein argued that the "centrifugal field" can be regarded as a gravitational field by virtue of the principle of equivalence. He demonstrated that the geodesic equations give the Newtonian equations of motion in the appropriate limit. In an earlier paper in the same year he and Fokker had used fully covariant tensors to demonstrate that the Nordström theory was a special case of the Einstein-Grossmann theory. However, even in early 1915, he continued to believe that full covariance was untenable and the tensors were invariant only under a class of transformations. He had almost arrived at the end of his journey.

It was in late November 1915 that Einstein finally concluded that full covariance, *with no restrictions whatsoever*, was required for the general theory of motion. Since the $4-d$ metric tensor describe is symmetric, it must have ten independent components. These components must all be present in an arbitrary gravitational field. By the end of the month he had derived the gravitational field equations and he and Hilbert had (independently) written down the correct Lagrangian. The unrestricted, or *general*, theory of Relativity had arrived! Its limitations, as a theory of motion, remained that it deals only with arbitrary motion in a *gravitational* field, but not with other fields like electromagnetism. To the end of his days Einstein was to try to incorporate electromagnetism satisfactorily into GR to obtain a Relativistic Unified Field Theory. (The cause of his dissatisfaction will be explained later.)

The generalized principle of Relativity is known as the *principle of general covariance*. It may be stated as follows: "All frames of reference are physically equivalent". Of course, put in this way much is left unsaid. It implies that all valid physical laws are expressible in tensorial form. Changes of frames of reference correspond to coordinate transformations that are, generally, nonlinear. It may be regarded as a restriction on what are acceptable physical laws. It should be stressed that this is the desired generalization of the principle of Special Relativity as the restriction to inertial frames has been dropped and *all* frames are allowed. It should also be pointed out, as a word of caution, that there can be changes of *coordinate systems* that do not correspond to changes of *frames of reference*. For example, changing from Cartesian to polar coordinates changes systems without changing frames. This difference, if kept clearly in mind, can avoid a lot of confusion.

1.8 Discussion of the Principles Underlying Relativity

The two assumptions on which General Relativity is actually based, the principles of equivalence and general covariance, have already been discussed in detail and will not be discussed further. There are infinitely many other, implicit, assumptions that have not been made explicit. For example, the assumption that there is a well defined "mass". Some of these assumptions entail how a variable in the theory will be measured. A change made will make apparent changes in the theory, without substantially changing it. As an example, consider the measurement of time. If measured by a water clock, which allows water to seep through a small hole in a container filled with a given volume of water, there will be a "law of speeding up of time with temperature", on account of the water evaporating, the stone slightly expanding and consequently the hole enlarging, and on account of the water becoming marginally less viscous. If measured by a pendulum clock on the Earth, there will be a "law of slowing down of time with altitude. Such theories are called "physically equivalent". These considerations have become important since the advent of GR because everybody wants to be an Einstein and provide a new competitor of GR. To deal with them, techniques were developed to experimentally test between genuinely different theories. I will discuss two of the older contenders.

First is the principle of simplicity, stating that physical laws should be simple in tensor form. Beyond the principle of general covariance it seems to be devoid of content, as "simplicity" is undefined. Naturally the simplest appearing laws will be tried first. If they are inadequate, complications will be introduced. When the complications start obscuring the significance of the law new concepts which make the laws seem simple, will be sought. As such it seems less a "principle" than a methodology for constructing theories *ab initio*. Taken in this spirit it has also been called "The Principle of Optimal Methodological Simplicity". Though reasonable it is merely a statement of the obvious procedure used. Let me elaborate this point further.

Till the end of the nineteenth century there was a belief in "the immutable laws of Nature". Many had been discovered and some still awaited discovery. Where Aristotelian science had taken "self-evident truths" as natural laws, however, Newtonian science took "self-evident definitions and concepts". In SR Einstein introduced the requirement of an *operational* definition of the quantities appearing in the theory. This did not entail a major change from the Newtonian view of science. GR did. I call such a theory "synthetic"; it did not evolve under the pressure of existing data but was constructed on account of logical and formal considerations. Its surprising success in not only explaining all known phenomena but predicting new phenomena, which were repeatedly verified, has made it the new ideal of science. Since then many new theories have been constructed *ab initio*. One such is the so-called "$f(R)$ gravity", which generalizes the Einstein-Hilbert Lagrangian to an arbitrary function of that Lagrangian. In its generality, this is not a scientific theory, as it introduces (continuously) infinitely many new parameters. Johann von Neumann is quoted as saying: "With 4 parameters I can fit an elephant an with 5 I can make him wiggle his trunk". I would add that given infinitely many parameters I can fit fairies and demons and have a parade of camels running through the eye of a needle. If it is regarded as a general formalism to provide one with machinery to churn out new theories, it is (at best) a mathematical method contribution.

The most commonly mentioned "additional" principle is Mach's principle. It is supposed to have originally motivated the development of General Relativity. The principle states that inertial mass exists because of "the distant stars". The reason for this belief is that a rotating frame can be seen to be non-inertial because the stars can be seen to be wheeling about the observer. The energy required to stop the rotation is not equal to the apparent energy of motion of the distant stars. Generally the "true" inertial frame will be that in which the distant stars are un-accelerated. To make sense of this proposal a scalar field connecting the distant stars to all matter was proposed. Of course, there is already Einstein's (tensor) field. One now needs to introduce a way of coupling the additional scalar field to the old tensor field. The first such proposal was the "Brans-Dicke scalar-tensor theory". The coupling was given by a dimensionless quantity written as ω. The larger the value of $\omega = 3$ the *less* is the coupling. It was claimed that there was reason to believe that $\omega = 3$. Later that claim had to be withdrawn. Tests to date have raised the value by a factor of more than a thousand. There have been other proposals since then. All precision tests have ruled out all scalar-tensor theories. No actual experiments can ever rule out any proposal totally, but only make it unbelievably messy to work with. That is what happened with the Aristotelian epicycles, the old "luminiferous aether" and Maxwell's "electromagnetic aether". Instead, one can rely on a thought experiment that proves a point of principle. Newton's bucket appears to negate the above argument. It was countered by pointing out that Newton's bucket experiment assumes the existence of a gravitational field *outside* the bucket which keeps the water surface of the un-rotated bucket flat. While the point is well-taken, the argument is invalid as explained below.

Consider a non-rotating self-gravitating fluid, i.e. the fluid has a gravitational field which holds the fluid together. In fact the gravitational force would tend to cause the fluid to collapse and become indefinitely dense. The fluid can maintain a finite size because the gravitational force will be opposed by the hydrostatic pressure. The size is determined by requiring that the two forces balance everywhere. Since both forces will act along the radial direction the fluid will take a spherical shape. Now consider a fluid rotating so that its surface has angular speed ω. The centrifugal force on any part of the fluid at distance r from the axis of rotation and having mass μ is $\mu r\omega^2$. Using spherical coordinates with the rotational axis as the polar axis, the centrifugal force on any point on the surface of a sphere of radius a will be $\mu a\omega^2 \sin\theta$, where θ is the polar angle. This will be zero at the poles and acquire a maximum value $\mu a\omega^2$ at the equator. Since this force will add to the hydrostatic pressure the fluid will acquire an equatorial bulge and take an elliptical shape. The "gas giant" planets of our Solar System, like Jupiter, show such a bulge. The elliptical shape of Jupiter could be observed by creatures on Jupiter, by using the geometric methods of Gauss (discussed in the next chapter), even if they could not see the stars. Though this argument shows that there is no necessity for Mach's principle it does not, by any means, disprove it. The general attitude seems to be to exclude Mach's principle on the basis of Occam's razor (that all unnecessary assumptions be dropped). Nevertheless many people still set great store by it. Very good discussions of the principle are available in [17] and [48].

As explained in my SR book [1], Einstein saw the need to use of Geometry for generalizing Relativity by considering a rotating disc. The circumference contracts as the motion is along the rim of the disc, but the diameter does not

as it is orthogonal to the motion. Hence the ratio of the circumference to the diameter is not π and so the Geometry of the space is non-Euclidean. Another argument given is SR is to consider a smooth rod moving on a smooth table-top, with a hole of the same size as the rod, at relativistic speeds (as such rods in textbooks of Mechanics are wont to do). An observer on the table will see the rod as shrunk and so it would fall through the hole. However, the observer on the rod (imagine her like a surfer) sees the hole as shrunk and expects to pass over the hole. Since this observer's frame changes from non-inertial to inertial, this expectation is wrong. The rod falls through the hole and she must see the table-top suddenly rise up in front like a tidal wave with the hole as a tunnel, so that she can go through the (shrunk) hole. Again, one sees the need for an invariant geometric description of the process. In the next two chapters this language will be developed before we go on to GR proper.

1.9 Exercises

1. The Lagrangian for a scalar field, ϕ, with four-gradient, $\phi_\mu = \phi_{,\mu}$, is

$$\mathcal{L} = \frac{1}{2}\phi^\mu\phi_\mu + \frac{m}{2}\phi^2 \ . \tag{1.29}$$

Derive the field equation for it. This is the relativistic energy conservation equation. When the classical energy conservation equation if quantized by Dirac's procedure, one gets the Schrödinger equation. When the relativistic energy conservation equation is quantized one gets the *Klein-Gordon equation*. What is the equation when the mass, m, is zero?

2. Derive the equations for a Lagrangian with the same first term as above and the second term: (a) $\frac{m}{3}\phi^3$; and (b) $\frac{m}{4}\phi^3$. If m has units of mass in the quadratic case, what would they it to be in these two cases? In what way would the theories for such potentials differ fundamentally from the quadratic case? [Hint: notice that the resulting EL equations for the quadratic case are linear.]

3. Take the first term in the original Lagrangian to be $\frac{1}{4}\phi^\mu\phi_\mu\phi^\nu\phi_\nu$ and derive the EL equations. What new feature arises in the EL equations here?

4. The Lagrangian, L, is normally taken as a function of a function, $y(x)$, and its first derivative, y'. Suppose it also depended on its second derivative, y''. Construct the Lagrange equations in this case by going through the variations from first principles. Generalise the Lagrange equations to n derivatives.

5. Extend the analysis in the previous example from the function $y(x)$ to the relativistic scalar field $\phi(x^\mu)$, and the Lagrangian, L, to a Lagrangian density, $\mathcal{L}$. Try to generalize this to a vector field $A^\nu(x^\mu)$.

Chapter 2

Review of Analytic Geometry in Three Dimensions

As explained in the previous chapter (and in [1]), the invariant language of Geometry is required to generalize Special Relativity. The "dialect" commonly used for the purpose is that of Tensors. In it rigor is occasionally sacrificed for the sake of calculational convenience. A notation by Roger Penrose provides the rigour of Differential Geometry with the convenience of Tensors. I have not understood why this "abstract index" notation has not caught on. (Perhaps it is because it had not been presented at the appropriate level, where such innovations are most readily accepted. If so this book should help to popularize the notation.) I will not go into rigorous Geometry here, but will only present as much as is required to introduce the abstract index notation. Even that is postponed to the next chapter. Before that I will review the theory of space curves and surfaces very briefly. This theory led to the development of modern Differential Geometry. It may be wondered why it is necessary to go into Mathematics so far removed from what is required for our present application. Apart from the argument for using a historical approach, outlined in the previous chapter, there are additional reasons in this case. Tensors, as commonly developed, are very abstract entities. To make them more concrete and easier to grasp, it is necessary to see how they came to be developed. Further, the literature on Relativity abounds with references to the theory of curves and surfaces and how to generalize it. While students of Mathematics should, probably, have studied it, Physics students have generally not. As such, physicists studying Relativity may miss the nuances in the literature and have to go back over various discussions. Also, many of the more formal requirements for the spaces considered in Geometry make sense only in the historical context. In this sense the theory of curves and surfaces discussed here, provides the motivation for the topics discussed in Geometry. Finally, as will be seen, it helps to develop tensor notation.

Carl Friedrich Gauss developed 3 dimensional geometry using calculus, as the first meaningful extension beyond Euclid. Before that, various people had pondered the fact that the axiom that parallel straight lines never meet, was neither proved, nor self-evident. As such, they felt that it should be possible to have a geometry without this axiom. Such geometries became popular as "non-

Euclidean geometries", but Gauss' was the first significant development. He first developed the theory of space curves using calculus and went on to develop the theory of surfaces embedded in 3-dimensional flat space. This breakthrough was soon followed by Nikolai Ivanovich Lobachevsky, who developed a slightly modified generalization of Gauss' theory without parallel lines meeting, giving a hyperbolic geometry. This becomes relevant for us later. The next development was by Bernhard Riemann, who died at the age of 40 after making major breakthroughs in most areas of Mathematics. He extended Gauss' geometry to n-dimensions and introduced the Tensor calculus. The purpose of this chapter is not to present the full theory of Differential Geometry but only the part relevant for us.

It is interesting to remark that there is a song by the Harvard mathematician and singer, Tom Lehrer, first performed in 1951 or 52, in which he makes fun of Lobachevsky for plagiarising the work of Riemann. Considering that the year that Lobachevsky presented his hyperbolic geometry, 1826, is the year of Riemann's *birth*, Tom Lehrer would have us believe that Lobachevsky was peeping over Riemannn's shoulders, as Riemann worked out his geometry in his mother's womb! Somehow, even the most liberal of Americans (Tom Lehrer being one) gets bitten by the media hype demonising a nation, race or ideology to pick on them unfairly. There is no doubt that the Russians love to claim priority of all ideas for the Russians, as the Indians do for the Indians, the Chinese do for the Chinese and as, most notably the English do for Isaac Newton. The only exception we know is the Pakistanis, who are likely to ascribe priority for any idea to *anybody* but a Pakistani, as they save their competitiveness for each other. Even the great Abdus Salam is not given credit by Pakistan for the work that won him worldwide acclaim.

2.1 Review of Three Dimensional Vector Notation

A point, P, in three dimensional Euclidean space, can be represented by a position vector, $\mathbf{x}$, relative to some other point, O, called the "origin". The distinction between Euclidean and non-Euclidean spaces will become clear later. Suffice it to say that Euclid's axioms hold in Euclidean spaces, but one or more of them do not hold in non-Euclidean spaces. For our present purpose it is enough to think of Euclidean space as the ordinary space. Vectors are generally represented by an arrow starting at O and ending at P. A 3-dimensional vector can also be represented algebraically in terms of three numbers, called coordinates. The most common are the Cartesian coordinates, see Fig. 2.1. Another very useful (and common) system is spherical coordinates, see Fig. 2.2. Slightly less useful, but still quite common, is the system of cylindrical coordinates, see Fig. 2.3. The common notations for these systems are (x, y, z), (r, θ, ϕ) and (r, θ, z) respectively. It is to be noted that the r, θ for cylindrical coordinates are very different from the r, θ for spherical coordinates. Particularly the cylindrical θ corresponds to the spherical ϕ (and is often represented by ϕ).

Generally, a vector can be written as a linear combination of three basis vectors. In Cartesian coordinates, with basis vectors ($\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$),

$$\mathbf{x} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k} \ . \tag{2.1}$$

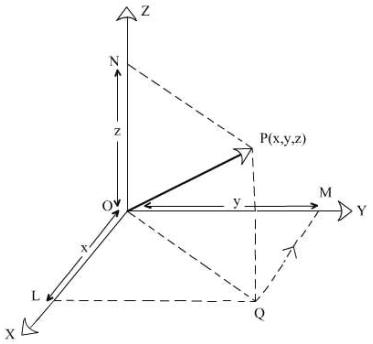


Figure 2.1: Cartesian coordinates $(x,\ y,\ z)$ are given by the projections OL, OM, ON of the position vector OP on the X, Y, Z axes.

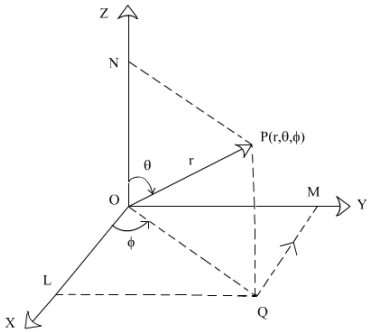


Figure 2.2: Spherical polar coordinates (r, θ, φ). The length of $\overrightarrow{OP}$ is r. The angle $\widehat{ZOP}$ is θ. The angle $\widehat{XOP}$ is φ. Here PQ is perpendicular to the XY-plane. Clearly $z = ON = r\cos\theta$, $x = OQ = r\sin\theta\cos\varphi$, $y = OM = r\sin\theta\sin\varphi$.

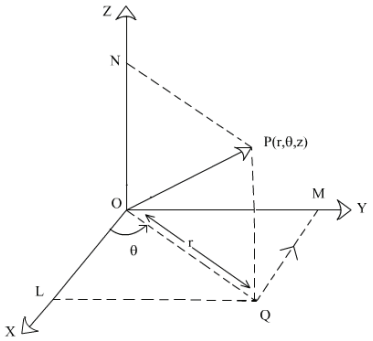


Figure 2.3: Cylindrical coordinates (r, θ, z). The length of OQ is r. The angle $\widehat{XOQ}$ is θ and z remains unchanged. Here $x = OL = r\cos\theta$, $y = OM = r\sin\theta$, $z = ON = z$.

Similarly, in spherical coordinates, with basis vectors $(\mathbf{l},\ \mathbf{m},\ \mathbf{n})$

$$\mathbf{x} = r\mathbf{l} + (r\theta)\mathbf{m} + (r\sin\theta)\mathbf{n}\ , \tag{2.2}$$

and in cylindrical coordinates, with basis vectors ($\mathbf{l}$, $\mathbf{m}$, $\mathbf{k}$)

$$\mathbf{x} = r\mathbf{l} + (r\theta)\mathbf{m} + z\mathbf{k} \ , \tag{2.3}$$

see Figs. (2.4)-(2.6). The basis vectors, here, are *orthonormal* i.e. a set of

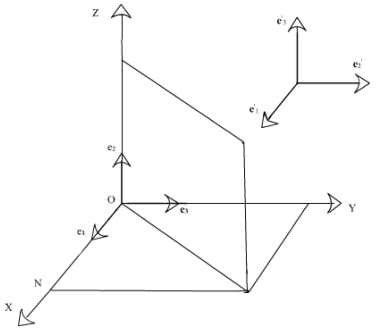


Figure 2.4: Cartesian basis vectors $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$ are unit vectors in the X, Y, Z directions. They are obviously invariant under translations as $\mathbf{e}'_1 = \mathbf{e}_1$, $\mathbf{e}'_2 = \mathbf{e}_2$, $\mathbf{e}'_3 = \mathbf{e}_3$.

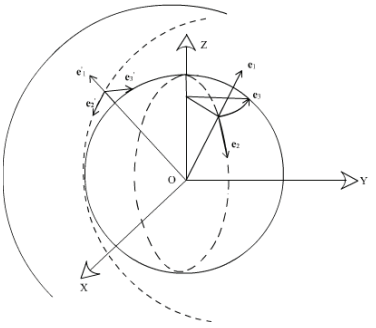


Figure 2.5: Spherical basis vectors are: $\mathbf{e}_1$ in the radial direction; $\mathbf{e}_2$ in the θ-direction (along a line of longitude on a sphere through the point); $\mathbf{e}_3$ in the φ -direction (along a line of latitude). Clearly under translation $\mathbf{e}'_1 \neq \mathbf{e}_1$, $\mathbf{e}'_2 \neq \mathbf{e}_2$, $\mathbf{e}'_3 \neq \mathbf{e}_3$.

orthogonal unit vectors. Notice that except for the Cartesian coordinate system, the coefficients of the basis vectors are often independent variables. So as to be able, later, to use partial differential calculus it is convenient to write the vector as a linear combination in which the coefficients are independent variables. In this case the basis vectors need no longer generally be orthonormal. They can remain orthogonal, as in the above coordinate systems, but need no longer be unit vectors. In fact their magnitudes may vary with position.

For an arbitrary coordinate system we have

$$\mathbf{x} = x^1\mathbf{e}_1 + x^2\mathbf{e}_2 + x^3\mathbf{e}_3 \ , \tag{2.4}$$

where the coordinates (x^1, x^2, x^3) are independent variables, and $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ are the basis vectors. We use the Einstein summation convention, whereby repeated

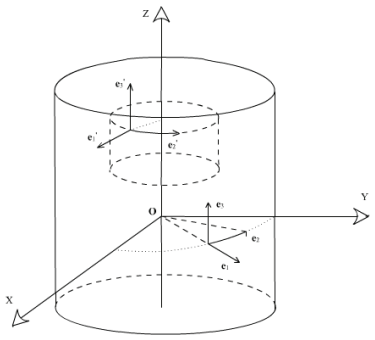


Figure 2.6: Cylindrical basis vectors are: $\mathbf{e}_1$ in the radial direction along a plane perpendicular to the axis of the cylinder; $\mathbf{e}_2$ along the θ-direction in the same plane orthogonal to $\mathbf{e}_1$ and $\mathbf{e}_3$ along the axis of the cylinder. Clearly under translation $\mathbf{e}'_1 \neq \mathbf{e}_1$, $\mathbf{e}'_2 \neq \mathbf{e}_2$, $\mathbf{e}'_3 \neq \mathbf{e}_3$.

indices will be summed over, with the modern provision that this pair of indices will be one superscript and one subscript. Clearly, it will be necessary to avoid an index appearing three or more times in a single expression to avoid ambiguity. (Where this becomes necessary it will be explicitly mentioned.) Thus Eq.(2.4) can be re-written as

$$\mathbf{x} = x^i \mathbf{e}_i \ , \quad (i = 1, 2, 3) \ . \tag{2.5}$$

The Cartesian system is orthogonal and normal, i.e.

$$\mathbf{e_i}.\mathbf{e_j} = \delta_{ij} \ , \tag{2.6}$$

where δ_{ij} is the *Kronecker delta* being 1 if $i = j$ and 0 if $i \neq j$. The former property guarantees normality while the latter gives orthogonality. For the other two coordinate systems mentioned the former property would not hold and hence Eq.(2.6) does not hold. The position vector is altered by a change of origin, called a *translation*. If we shift from O to O', by a vector $\mathbf{a}$, i.e. $\overrightarrow{OO'} = \mathbf{a}$, then

$$\mathbf{x} \to \mathbf{x}' = \mathbf{x} - \mathbf{a} \ . \tag{2.7}$$

The Cartesian basis vectors are unaffected by this transformation but the components of the vector are altered (see Fig. 2.7).

$$\mathbf{x}' = x^i \mathbf{e}_i - a^i \mathbf{e}_i = (x^i - a^i)\mathbf{e}_i = x'^i \mathbf{e}_i \ . \tag{2.8}$$

Vectors that are altered by translations are called *local vectors*. One generally talks of vectors that are not local, e.g. if P and Q have position vectors $\mathbf{x}$ and $\mathbf{y}$, respectively, the vector, $\mathbf{z}$, from P to Q will be origin independent

$$\mathbf{z}' = \mathbf{y}' - \mathbf{x}' = (\mathbf{y} - \mathbf{a}) - (\mathbf{x} - \mathbf{a}) = \mathbf{y} - \mathbf{x} = \mathbf{z} \ . \tag{2.9}$$

The position vector remains invariant under rotation of the axes but both its basis vectors and its components change,

$$\mathbf{x} \to \mathbf{x}' = x'^i e_i = (x^i \Lambda^j_i)(\Lambda^k_j e_k) = \mathbf{x} = x^i e_i = x^i \delta^k_i e_k \ , \tag{2.10}$$

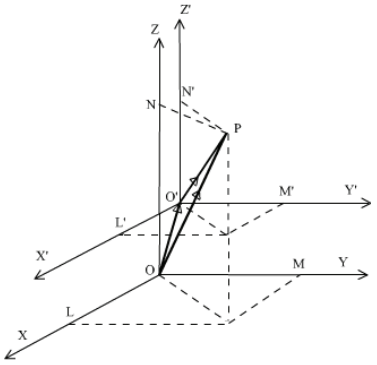


Figure 2.7: Translation of the origin from O to O' by a vector $\mathbf{a}$ changes the components of the vector $\overrightarrow{OP} = x$, from OL, OM, ON to $OL' \neq OL$,$OM' \neq OM$, $ON' \neq ON$ in general.

where Λ_i^j are the rotation matrices. These must be such that the transpose of the matrix is its inverse, i.e.

$$\Lambda_i^j \Lambda_j^k = \delta_i^k \ , \tag{2.11}$$

to maintain invariance of the vector. It is easy to see the sense in which invariance is maintained for a 2-d vector (see Fig. 2.8). In this case (as shown in SR)

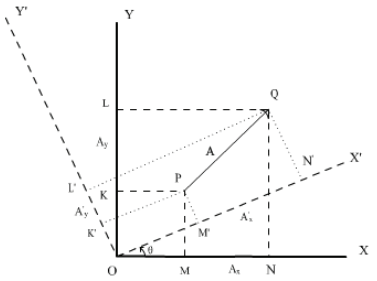


Figure 2.8: The change of basis due to rotation exactly balances the change of components of the vector.

$$\Lambda_i^j = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \ . \tag{2.12}$$

A *point* is a zero-d object and is determined by fixing all the coordinates. A *curve* is a 1-d object, meaning that points on the curve are generated by varying a single parameter, u. Thus it is represented by a variable position vector, $\mathbf{x}(u)$. A surface is a 2-d object, whose points are generated by varying two parameters, (u, v), represented by $\mathbf{x}(u, v)$. Fixing one of the parameters of a surface gives a curve and fixing the parameter of a curve gives a point.

2.2 Space Curves

In 3-d Cartesian coordinates a curve can be written down explicitly as a set of three functions of a single variable

$$\mathbf{x}(u) = (x(u), y(u), z(u)) \ . \tag{2.13}$$

The curve is continuous (or differentiable or analytic) if all three functions are. If it is (at least) differentiable, the tangent to the curve at a point P on it, at parameter value u_0, is given by

$$\mathbf{x}'(u_0) = d\mathbf{x}/du|_{u=u_0} \ . \tag{2.14}$$

Re-parameterizing the curve re-scales the tangent vector. Since re-parameterization is merely a change of the *description* of the geometrical object and not a *geometrical* change of the object, one wants the tangent vector to remain invariant under such changes. This may be assured by choosing a unit vector, i.e. a vector of unit magnitude. The *unit tangent vector*, $\mathbf{t}(u_0)$, is given by normalizing the vector given by Eq.(2.14),

$$\mathbf{t}(u_0) = \mathbf{x}'(u_0)/|\mathbf{x}'(u_0)| \ . \tag{2.15}$$

Repeating this procedure for all points along the curve we get a variable unit vector $\mathbf{t}(u)$, which is always tangent to the curve.

By re-parameterization, we choose a parameter, s, called the *affine parameter*, such that the derivative of $\mathbf{x}$ with respect to has unit magnitude. Derivatives with respect to this parameter will be denoted by dots. Then, by definition,

$$\mathbf{t}(s) = \dot{\mathbf{x}}(s) \equiv \frac{d\mathbf{x}}{ds} = \frac{du}{ds}\frac{d\mathbf{x}}{du} = \frac{\mathbf{x}'(u)}{ds/du} \ . \tag{2.16}$$

Comparing Eqs.(2.15) and (2.16) it is obvious that

$$ds/du = |\mathbf{x}'(u)| \ . \tag{2.17}$$

In principle Eq.(2.17) can be integrated to give s as a function of u. This procedure is called *quadrature*. Taking the definite integral from a point A on the curve, to another point B, where a is the value of u at A and b the value of u at B,

$$\int_A^B ds = \int_a^b \sqrt{\mathbf{x}'.\mathbf{x}'}du = \int_a^b \sqrt{d\mathbf{x}.d\mathbf{x}} \ , \tag{2.18}$$

There may of be no closed form expression giving the arc length s as a function of u. Even if one can find such a closed form, it may not be possible to invert it to give u as a function of s. Though it may not be *practicable* to replace u by s in Eq.(2.13), in principle the implicit function theorem says it *is* possible. Thus we will write all quantities as functions of s, bearing in mind that it is often preferable to use another parameter. We can re-express Eq.(2.17) also as

$$ds^2 = d\mathbf{x}.d\mathbf{x} \ . \tag{2.19}$$

called the *line element* or the *metric* (from the Greek word for "measure").

Of the infinitely many vectors orthogonal to a space curve at a given point, we want to select a "preferred" normal. To this end we generalize the procedure

for obtaining the tangent. In $2-d$ one takes the limit as a point Q on the curve approached the point P. Now we take two points R and Q approaching P from either side (see Fig. 2.9). Two points uniquely give a *line*. The three points P, Q and R uniquely define a *plane*, provided they are not co-linear. The vector orthogonal to $\mathbf{t}$, at P in the limiting plane as Q, $R \to P$ (called the *osculating plane*) is the preferred normal. Since $\mathbf{t}.\mathbf{t} = 1$, by differentiation it is clear that $\mathbf{t}.\dot{\mathbf{t}} = 0$ and hence $\dot{\mathbf{t}} \perp \mathbf{t}$. Obviously, $\dot{\mathbf{t}} \neq 0$, for $\mathbf{x}$ to represent a curve and not a straight line. Hence $\dot{\mathbf{t}}$ will lie along the normal, as is clear from Fig. 2.9. We define the *unit normal vector* by

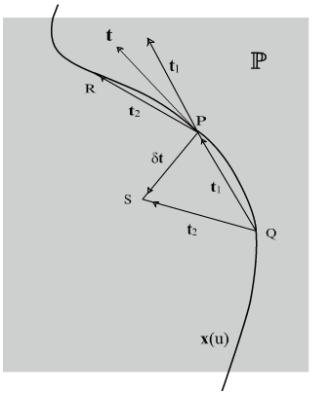


Figure 2.9: A curve, $\mathbf{x}(u)$, has three points Q, P, R on it. Take $\overrightarrow{QP} = \mathbf{t}_1$ and $\overrightarrow{PR} = \mathbf{t}_2$. If $|\mathbf{t}_1| = |\mathbf{t}_2|$, in the limit as $Q \to P$, the average of $\mathbf{t}_1$ and $\mathbf{t}_2$ is $\mathbf{t}$, the tangent vector. The difference is $\overrightarrow{PS} = \delta\mathbf{t}$. In the limit as Q, $R \to P$, $\delta\mathbf{t} \perp \mathbf{t}$. The three points P, Q, R define the osculating plane $\mathbb{P}$, in the limit as Q, $R \to P$. Both $\mathbf{t}$ and $\delta\mathbf{t}$ lie in $\mathbb{P}$.

$$\mathbf{n}(s) = \dot{\mathbf{t}}(s)/|\dot{\mathbf{t}}(s)| \ , \tag{2.20}$$

provided $\dot{\mathbf{t}}(s) \neq 0$. When $\dot{\mathbf{t}}(s) = 0$, integrating Eq.(2.16) gives

$$\mathbf{x}(s) = \mathbf{A}s + \mathbf{B} \ , \tag{2.21}$$

where $\mathbf{A}$ is a unit vector, i.e. we have a straight line.

Since $\dot{\mathbf{t}} = 0$ gives a straight line and $\dot{\mathbf{t}} \approx 0$ gives a curve that is nearly straight, $\dot{\mathbf{t}}$ is a measure of the curvature of the curve. Formally the *curvature*, κ, is defined by

$$\begin{aligned} \kappa^2(s) &= \dot{\mathbf{t}}(s).\dot{\mathbf{t}}(s) = \ddot{\mathbf{x}}(s).\ddot{\mathbf{x}}(s) \\ &= \frac{[\mathbf{x}'(u) \times \mathbf{x}''(u)] \cdot [\mathbf{x}'(u) \times \mathbf{x}''(u)]}{[\mathbf{x}'(u).\mathbf{x}'(u)]^3} \ . \end{aligned} \tag{2.22}$$

Thus we have

$$\dot{\mathbf{t}}(s) = \kappa(s)\mathbf{n}(s) \ . \tag{2.23}$$

The vectors $\mathbf{t}(s)$ and $\mathbf{n}(s)$ span the osculating plane, i.e. any vector in it can be written as a linear combination of these two vectors. They form an orthonormal basis for this plane.

A complete orthonormal basis for the 3-dimensional space, from the point P on the curve, is provided by defining the *unit binormal vector* orthogonal to the osculating plane

$$\mathbf{b}(s) = \mathbf{t}(s) \times \mathbf{n}(s) \ . \tag{2.24}$$

This basis is called the *moving triad* and is convenient for describing geometry from the point of view of an observer moving along the curve. For example, for a car moving along a hilly road with turns along which it climbs up or down, the tangent vector lies along the bonnet, the normal lies straight out the side window and the binormal straight out of the car roof (see Fig. 2.10). Though developed long before Relativity it is very definitely "relativistic in spirit" in that it gives the description according to the observer and there is no preferred observer.

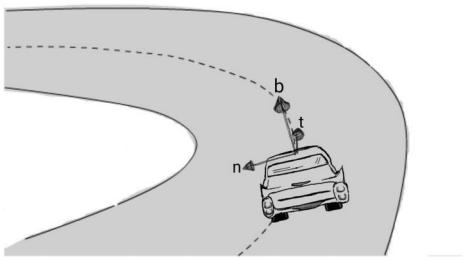


Figure 2.10: A car going up a mountain road is on an embankment, and therefore not level. It has a frame of three 1 metre long arrows which are mutually perpendicular stuck on its roof. The first points in the direction of motion of the car, $\mathbf{t}$, the second off the car to the left, $\mathbf{n}$, and the third straight out of the roof of the car, $\mathbf{b}$. Notice that $\mathbf{b}$ will not point straight "up" (shown by the dotted line) here as the car is not level. Also, $\mathbf{n}$ will not be horizontal.

Since all vectors can be expressed as linear combinations of the basis vectors we can, in particular, write the derivatives of the basis vectors as a linear combination of the basis vectors. In fact $\dot{\mathbf{t}}$ is already given in this form by Eq.(2.23). It turns out to be easier to work out $\dot{\mathbf{b}}$ first and then $\dot{\mathbf{n}}$. Differentiating

$$\mathbf{t}(s).\mathbf{b}(s) = 0 \ , \ \mathbf{b}(s).\mathbf{b}(s) = 1 \ , \tag{2.25}$$

with respect to s, using Eq.(2.23) and the fact that $\mathbf{b} \perp \mathbf{n}$, we see that $\dot{\mathbf{b}} \perp \mathbf{t}$, $\mathbf{b}$. Thus

$$\dot{\mathbf{b}}(s) = -\tau(s)\mathbf{n}(s) \ , \tag{2.26}$$

where the proportionality factor, $\tau(s)$, is called the *torsion* or the *second curvature*. (The reason for the former name will become clear shortly, while for the latter will be explained in the next chapter.)

We now come to the evaluation of $\dot{\mathbf{n}}$. For this purpose consider

$$\mathbf{t}(s).\mathbf{n}(s) = 0 \ , \ \mathbf{n}(s).\mathbf{n}(s) = 1 \ , \ \mathbf{b}(s).\mathbf{n}(s) = 0 \ . \tag{2.27}$$

Again differentiating these equations and using Eqs.(2.23) and now (2.26)

$$\mathbf{t}(s).\dot{\mathbf{n}}(s) = -\kappa \ , \ \mathbf{n}(s).\dot{\mathbf{n}}(s) = 0 \ , \ \mathbf{b}(s).\dot{\mathbf{n}}(s) = \tau \ . \tag{2.28}$$

The expansion of $\dot{\mathbf{n}}$ in terms of the moving triad would have the respective coefficients given by Eq.(2.28). Hence

$$\dot{\mathbf{n}}(s) = -\kappa(s)\mathbf{t}(s) + \tau(s)\mathbf{b}(s) \ . \tag{2.29}$$

Eqs.(2.23), (2.26) and (2.29) are known, collectively as the *Frenet-Serret formulae.*

Example: We will construct and compute all the quantities discussed for the *helix* defined by

$$\mathbf{x}(u) = (a\cos u, a\sin u, cu) \ . \tag{2.30}$$

Therefore

$$\mathbf{x}'(u) = (-a\sin u, a\cos u, c) \ ,$$

and hence

$$|\mathbf{x}'(u)| = \sqrt{a^2 + c^2} \ ,$$

Thus, by Eq.(2.17)

$$s = \sqrt{a + c^2}u \ ,$$

having chosen the zero of s to coincide with the zero of u. Thus Eq.(2.30) can be rewritten in terms of s as

$$\begin{aligned}
\mathbf{x}\,(s) &= \left(a\cos\frac{s}{\sqrt{a^2+c^2}}, a\sin\frac{s}{\sqrt{a^2+c^2}}, \frac{cs}{\sqrt{a^2+c^2}}\right) \ , \\
\mathbf{t}\,(s) &= \frac{1}{\sqrt{a^2+c^2}}\left(-a\sin\frac{s}{\sqrt{a^2+c^2}}, a\cos\frac{s}{\sqrt{a^2+c^2}}, c\right) \ , \\
\dot{\mathbf{t}}\,(s) &= \frac{-a}{a^2+c^2}\left(\cos\frac{s}{\sqrt{a^2+c^2}}, \sin\frac{s}{\sqrt{a^2+c^2}}, 0\right) \ .
\end{aligned}$$

Thus, by Eqs.(2.22), (2.23) and (2.24)

$$\begin{aligned}
\kappa^2\,(s) &= \frac{a}{a^2+c^2} \ , \\
\mathbf{n}(s) &= -\left(\cos\frac{s}{\sqrt{a^2+c^2}}, \sin\frac{s}{\sqrt{a^2+c^2}}, 0\right) \ , \\
\mathbf{b}(s) &= \frac{1}{\sqrt{a^2+c^2}}\left(c\sin\frac{s}{\sqrt{a^2+c^2}}, -c\cos\frac{s}{\sqrt{a^2+c^2}}, a\right) \ , \\
\dot{\mathbf{b}}(s) &= \frac{c}{a^2+c^2}\left(\cos\frac{s}{\sqrt{a^2+c^2}}, \sin\frac{s}{\sqrt{a^2+c^2}}, 0\right) \ .
\end{aligned}$$

From Eq.(2.26)

$$\tau(s) = c/(a^2 + c^2) \ .$$

The osculating plane is given by the position vector $\mathbf{X}(\lambda, \mu)$ which has the point of the curve on it and is spanned by the vectors $\mathbf{t}(s)$ and $\mathbf{n}(s)$ at the point. For an arbitrary point at parameter value s_0,

$$\mathbf{X}(\overline{\lambda}, \mu) = \mathbf{x}(s_0) + \overline{\lambda}\mathbf{t}(s_0) + \mu\mathbf{n}(s_0) \ .$$

Define

$$\lambda = \overline{\lambda} a/\sqrt{a^2 + c^2} \ ,$$

and use the parameter u, with u_0 corresponding to s_0. Then

$$\mathbf{X}(\lambda, \mu) = ((a - \mu)\cos u_0 - \lambda \sin u_0, (a - \mu)\sin u_0 + \lambda \cos u_0, c(u_0 + \lambda\sqrt{a^2 + c^2}/a)) \ .$$

The helix is depicted in Fig. 2.11. It is apparent that c measures how "fast" the helix spirals up. Clearly,

$$\tau/\kappa = c/a = \text{constant} \ . \tag{2.31}$$

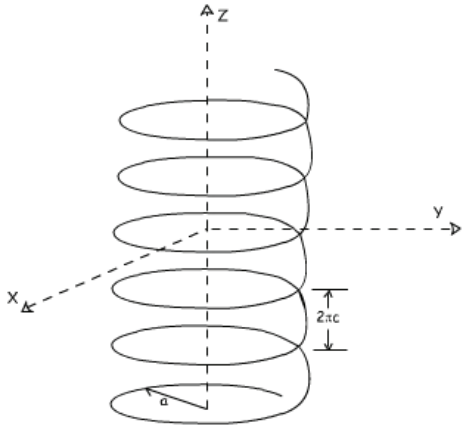


Figure 2.11: A helix is like a coil of spring. The radius of the spring is "a" and its "pitch" is $2\pi c$. The curvature is $\mathbf{a}$ for small c and the "twist" c.

In the case $c = 0$ we get a circle of curvature $1/a$ and zero torsion. The inverse of the curvature is called the *radius of curvature* and corresponds to the radius of a circle with the same curvature. The circle with this radius which lies in the osculating plane is called the *osculating circle*. Notice that if $a \to 0$ we get a straight line. In this case the torsion must also vanish. (We can have a curve with curvature and no torsion but not with torsion and no curvature.) As c is increased from zero, for a given a, the torsion increases till $c = a$, after which it decreases again. As is clear from Fig. 2.11, the helix is like a spring. Increasing c for a given a, for $c \ll a$, corresponds to stretching the string. This is what is measured by the physical "torsion" related to torques. Hence the name "torsion". A curve with zero torsion will have $\dot{\mathbf{b}}(s) = 0$ and hence $\mathbf{b}(s) =$ constant. Thus the osculating plane remains the same. Hence the curve always lies in the same plane. Such curves are called *planar curves*.

2.3 Surfaces

In Cartesian coordinates a surface can be written explicitly as a set of three functions of two variables

$$\mathbf{x}(u, v) = (x(u, v), y(u, v), z(u, v)) \ . \tag{2.32}$$

There is no unique vector tangent to a surface. However, it is clear that the vectors

$$\mathbf{x}_u(u_0,v_0) \equiv \left.\frac{\partial \mathbf{x}(u,v)}{\partial u}\right|_{u=u_0,v=v_0} , \ \mathbf{x}_v(u_0,v_0) \equiv \left.\frac{\partial \mathbf{x}(u,v)}{\partial v}\right|_{u=u_0,v=v_0} . \quad (2.33)$$

are tangent to the surface at the point P, given by the parameter values (u_0, v_0) on the surface (see Fig 2.12) if the vectors are linearly dependent, i.e. they are either parallel or one (or both) of them is zero, then

$$\mathbf{x}_u(u_0,v_0) \times \mathbf{x}_v(u_0,v_0) = 0 . \quad (2.34)$$

If Eq.(2.34) does not hold then $\mathbf{x}_u$ and $\mathbf{x}_v$, at P span a plane, called the *tangent*

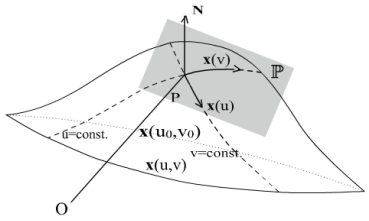


Figure 2.12: A surface $\mathbf{x}(u,v)$ has a point, P, on it. The tangent to the $v =$ constant curve on the surface, at the point P, is $\mathbf{x}_u$ and the tangent to the $u =$ constant curve at P, is $\mathbf{x}_v$. The plane, $\mathbb{P}$, spanned by $\mathbf{x}_u$ and $\mathbf{x}_v$ is the tangent plane. $\mathbf{N}$ is the unit normal, orthogonal to $\mathbb{P}$.

plane. A point at which Eq.(2.34) holds is called a *singular point*. As one avoids singular points, one needs to determine which points are singular. A common procedure to do so is to construct a (2×3) matrix, $\mathcal{M}$, consisting of $\mathbf{x}_u$ and $\mathbf{x}_v$, and look for points at which its rank is less than 2.

Consider a point Q at parameter values $(u_0 + du, v_0 + dv)$ on the surface, a short distance from P. The vector $\overrightarrow{PQ}$ is given by

$$d\mathbf{x} = \mathbf{x}_u du + \mathbf{x}_v dv + \frac{1}{2!}\left(\mathbf{x}_{uu}du^2 + 2\mathbf{x}_{uv}dudv + \mathbf{x}_{vv}dv^2\right) + \ldots . \quad (2.35)$$

If du and dv are infinitesimal the arc length along the surface will be given by the *first fundamental form*

$$\begin{aligned} ds^2 &= d\mathbf{x}\cdot d\mathbf{x} \\ &= \mathbf{x}_u\cdot\mathbf{x}_u du^2 + 2\mathbf{x}_u\cdot\mathbf{x}_v dudv + \mathbf{x}_v\cdot\mathbf{x}_v dv^2 \\ &= E du^2 + 2F dudv + G dv^2 . \end{aligned} \quad (2.36)$$

The first fundamental form is *positive definite*, i.e.

$$ds^2 \geq 0 \quad (2.37)$$

$$ds^2 = 0 \Longleftrightarrow Q = P \ \ (\text{i.e. } du = dv = 0) . \quad (2.38)$$

The former property is positivity while the latter is definiteness. This fact is geometrically obvious, but is worth following up more formally for later use,

when we will discuss more general spaces. To prove the result we will try to *disprove* it. For ds^2 to be negative at one place, being positive elsewhere, it must be zero somewhere. Suppose that this is possible. Then, dividing Eq.(2.36) by du^2 and writing dv/du as λ

$$E + 2F\lambda + G\lambda^2 = 0 \ . \tag{2.39}$$

This equation will have a solution only if the discriminant is not negative, i.e. if

$$F^2 \geq EG \ . \tag{2.40}$$

Inserting the definition of E, F and G the requirement is that

$$(\mathbf{x}_u \cdot \mathbf{x}_v)^2 \geq (\mathbf{x}_u \cdot \mathbf{x}_u) \cdot (\mathbf{x}_v \cdot \mathbf{x}_v) \ . \tag{2.41}$$

Let $x_u = |\mathbf{x}_u|$, $x_v = |\mathbf{x}_v|$ and θ be the angle between $\mathbf{x}_u$ and $\mathbf{x}_v$. The requirement is, then that

$$x_u^2 x_v^2 \cos^2\theta \geq x_u^2 x_v^2 \ . \tag{2.42}$$

Since Eq.(2.34) does not hold x_u, x_v, $\theta \neq 0$. Hence Eq.(2.42) can never hold. Hence Eq.(2.39) can never hold. Thus, for finite du, dv the metric is strictly positive, $ds^2 > 0$, and the only way that we can have $ds^2 = 0$ is for $du, dv = 0$ i.e. $P = Q$.

Notice that the first fundamental form depends only on the two tangent vectors that lie *along* the surface and not on a vector orthogonal to it. In other words it can be described from *within* the surface and does not require that one goes out of the surface. It can be expressed in index notation by writing (u, v) as u^a $(a = 1, 2)$ and the metric coefficients as the symmetric partitioned row matrix given by

$$g_{ab} = \left(E \ \ F \vdots F \ \ G \right) \ , \tag{2.43}$$

so that $g_{11} = E$, $g_{12} = g_{21} = F$, $g_{22} = G$. Then

$$\begin{aligned} ds^2 &= g_{11}\left(du^1\right)^2 + g_{12}du^1du^2 + g_{21}du^2du^1 + g_{22}\left(du^2\right)^2 \\ &= g_{ab}du^a du^b \ . \end{aligned} \tag{2.44}$$

A *surface curve* is defined by reducing the number of independent variables of the surface from two to one, i.e. by specifying a functional relation

$$\phi(u, v) = 0 \ , \tag{2.45}$$

between the two variables u and v. The function may, or may not, be separable in explicit form. For example a circle on a sphere or a helix on a cylinder are surface curves. Surface curves on curved surfaces may be planar, as in the former example (as may be seen by slicing the sphere by a plane to obtain the circle), or non-planar, as in the latter example.

Consider two surface curves

$$\phi_1(u, v) = 0 = \phi_2(u, v) \ , \tag{2.46}$$

which intersect at some point $P(u_0, v_0)$. Each curve will have a unit tangent vector that lies in the same tangent plane at P (see Fig. 2.13). Let them be

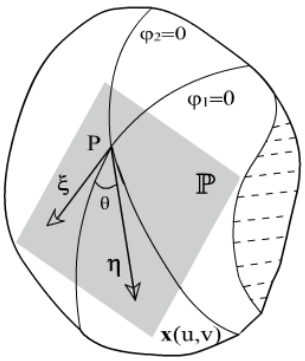


Figure 2.13: There are two curves $\varphi_1(u,v)=0$ and $\varphi_2(u,v)=0$, on the surface $x(u,v)$. They intersect at the point P. At this point they have unit tangent vectors $\underline{\xi}$ and $\underline{\eta}$ respectively. These vectors span the tangent plane $\mathbb{P}$ and make an angle θ.

denoted by $\underline{\xi}$ and $\underline{\eta}$ respectively. The *angle* between the two curves, θ, is then given by

$$\cos\theta = \underline{\xi}\cdot\underline{\eta}\ . \tag{2.47}$$

To compute θ we need to write both as derivatives of the position vector relative to the arc-length for each curve. So as to distinguish between the two parameters we denote the former by s_1 and the latter by s_2. Then

$$\begin{aligned}\cos\theta &= \frac{d\mathbf{x}}{ds_1}\cdot\frac{d\mathbf{x}}{ds_2}\\ &= \left(\mathbf{x}_u\frac{du}{ds_1}+\mathbf{x}_v\frac{dv}{ds_1}\right)\cdot\left(\mathbf{x}_u\frac{du}{ds_2}+\mathbf{x}_v\frac{dv}{ds_2}\right)\\ &= E\frac{du}{ds_1}\frac{du}{ds_2}+F\left(\frac{du}{ds_1}\frac{dv}{ds_2}+\frac{dv}{ds_1}\frac{du}{ds_2}\right)+G\frac{dv}{ds_1}\frac{dv}{ds_2}\ .\end{aligned} \tag{2.48}$$

We can, therefore, express the two vectors as two dimensional by writing them as

$$\xi^a=\frac{du^a}{ds_1}\ ,\ \eta^a=\frac{du^a}{ds_2} \tag{2.49}$$

and the angle as

$$\cos\theta = g_{ab}\xi^a\eta^b\ . \tag{2.50}$$

Thus the "dot", or scalar, product in Eq.(2.47) is replaced by g_{ab} when converting to index notation.

Consider the particular case when the two curves are the coordinate lines

$$\phi_1(u,v)\equiv v-v_0\ ,\ \phi_2(u,v)\equiv u-u_0\ , \tag{2.51}$$

where u_0 and v_0 are constants of integration. Thus we have

$$\frac{dv}{ds_1}=\frac{du^2}{ds_1}=0\ ,\ \frac{du}{ds_2}=\frac{du^1}{ds_2}=0\ . \tag{2.52}$$

Hence, from Eq.(2.48)

$$\cos\theta = F\frac{du}{ds_1}\frac{dv}{ds_2}=g_{12}\frac{du^1}{ds_1}\frac{du^2}{ds_2}\ . \tag{2.53}$$

Therefore if $F = 0$, $\theta = \pi/2$ and hence the coordinate lines are orthogonal. Conversely if the coordinate lines are orthogonal, $\cos\theta = 0$ and hence $F = 0$. Thus for orthogonal coordinate systems g_{ab} will be diagonal. Notice that since $\underline{\xi}$ and $\underline{\eta}$ are unit vectors

$$\frac{du^1}{ds_1} = \frac{du^2}{ds_2} = 1 \tag{2.54}$$

and hence Eq.(2.53) becomes

$$\cos\theta = F \ . \tag{2.55}$$

2.4 Coordinate Transformations on Surfaces

The parameters (u, v) on the surface perform the function of coordinates. As for the curve, the surface can be re-parameterized. However, there is no preferred set of parameters in this case. As such we must require that there be no difference made by such a re-parameterization, or coordinate transformation. This is what makes Geometry so useful in Relativity. Consider the transformation $(u, v) \to (\hat{u}, \hat{v})$ or $u^a \to u^{\hat{a}}$ which need not be a linear transformation. Then by the chain rule

$$\left.\begin{aligned} \mathbf{x}_{\hat{u}}\,(\hat{u},\hat{v}) =& \frac{\partial u}{\partial \hat{u}}\mathbf{x}_u(u,v) + \frac{\partial v}{\partial \hat{u}}\mathbf{x}_v(u,v)\ , \\ \mathbf{x}_{\hat{v}}\,(\hat{u},\hat{v}) =& \frac{\partial u}{\partial \hat{v}}\mathbf{x}_u(u,v) + \frac{\partial v}{\partial \hat{v}}\mathbf{x}_v(u,v)\ . \end{aligned}\right\} \tag{2.56}$$

The vectors $(\mathbf{x}_u, \mathbf{x}_v)$ provide a basis set for the tangent plane which we denote by $\mathbf{e}_a$ and the vectors $(\mathbf{x}_{\hat{u}}, \mathbf{x}_{\hat{v}})$ provide another basis which we denote by $\mathbf{e}_{\hat{a}}$. Then Eq.(2.56) can be re-written as

$$\mathbf{e}_{\hat{a}} = \delta^a_{\ \hat{a}}\mathbf{e}_a \ , \tag{2.57}$$

where $\delta^a_{\ \hat{a}}$ is the *Jacobian matrix*

$$\partial u^a/\partial u^{\hat{a}} := \delta^a_{\ \hat{a}} = \begin{pmatrix} \partial u/\partial\hat{u} & \partial u/\partial\hat{v} \\ \partial v/\partial\hat{u} & \partial v/\partial\hat{v} \end{pmatrix} \tag{2.58}$$

and the basis vectors are written collectively as a row matrix and hence *pre-multiply* the square matrix.

The coordinate transformation is permissible if it is invertible. Thus we require that the Jacobian matrix be non-singular, i.e. the *Jacobian determinant* (or *Jacobian*) $\det\left(\delta^a_{\ \hat{a}}\right) = \Delta$, must be non-zero. If at some point $\mathbf{e}_a$ is well-defined and $\Delta = 0$ then $\mathbf{e}_{\hat{a}}$ will not form a basis. In this case, in the re-parameterized form, the point will appear to be singular, while in the original form it will be non-singular. If the singularity is due to the choice of coordinates on the surface it is called a *coordinate singularity*, while if it is due to a feature of the surface it is called an *essential singularity.* Conversely, if there is a coordinate singularity, it can be removed by a coordinate transformation, but an essential singularity cannot be removed. The inverse transformation, where it exists, is written as $\delta^{\hat{a}}_a$. Thus,

$$\delta^a_{\ \hat{a}}\delta^{\hat{b}}_a = \delta^{\hat{b}}_{\hat{a}} \ , \ \delta^a_{\ \hat{a}}\delta^{\hat{a}}_b = \delta^a_b \ . \tag{2.59}$$

The inverse transformation can be used to express the old basis in terms of the new by

$$\mathbf{e}_a = \delta^{\hat{a}}_a\mathbf{e}_{\hat{a}} \ . \tag{2.60}$$

We can define a *dual basis* $\mathbf{e}^a$ such that

$$\mathbf{e}_a \cdot \mathbf{e}^b = \delta_a^b \ . \tag{2.61}$$

The dual basis is transformed by the inverse transformation, since the Kronecker delta is invariant under coordinate transformations,

$$\mathbf{e}^{\hat{a}} = \delta_a^{\hat{a}} \mathbf{e}^a \ , \ \mathbf{e}^a = \delta_{\hat{a}}^a \mathbf{e}^{\hat{a}} \ . \tag{2.62}$$

2.5 The Second Fundamental Form

So far we have dealt with the basis for the surface, or more precisely for the tangent plane to the surface. For the 3-dimensional space a third vector is required to complete the basis. For this purpose we take a vector normal to the surface at a point. By 'normal' we mean that the vector is orthogonal to the tangent plane at that point. For uniqueness we define the *unit normal* vector (see Fig 2.12)

$$\mathbf{N} = \frac{\mathbf{x}_u \times \mathbf{x}_v}{|\mathbf{x}_u \times \mathbf{x}_v|} \ . \tag{2.63}$$

Clearly $\mathbf{N}$ must not change under re-parameterization. This requirement is geometrically obvious. To verify it algebraically, consider the effect of re-parameterization on $\mathbf{x}_u \times \mathbf{x}_v$. Now

$$\begin{aligned} \mathbf{x}_{\hat{u}} \times \mathbf{x}_{\hat{v}} &= \left(\frac{\partial u}{\partial \hat{u}} \mathbf{x}_u + \frac{\partial v}{\partial \hat{u}} \mathbf{x}_v \right) \times \left(\frac{\partial u}{\partial \hat{v}} \mathbf{x}_u + \frac{\partial v}{\partial \hat{v}} \mathbf{x}_v \right) \\ &= \left(\frac{\partial u}{\partial \hat{u}} \frac{\partial v}{\partial \hat{v}} - \frac{\partial v}{\partial \hat{u}} \frac{\partial u}{\partial \hat{v}} \right) \mathbf{x}_u \times \mathbf{x}_v = \Delta \left(\mathbf{x}_u \times \mathbf{x}_v \right) \ . \end{aligned} \tag{2.64}$$

The magnitude of the above quantity is clearly given by the same Δ times the previous magnitude. Thus the transformed $\mathbf{N}$ remains invariant as the Δ in the numerator cancels the Δ in the denominator.

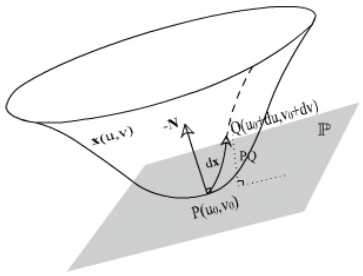


Figure 2.14: The height of Q above the tangent plane, $\mathbb{P}$, to the surface $\mathbf{x}(u, v)$ at the point P is $d_{PQ} = -d\mathbf{x} \cdot \mathbf{N}$, where P is at the parameter values (u, v) and Q is at $(u + du, v + dv)$.

The curvature of a curve is given by the rate of change of the tangent vector with the affine parameter. It may be regarded as the rate at which the slope rises above its previous level. For a surface curve the curvature will depend on the choice of curve and not on the nature of the surface. However, we can measure how fast the surface rises above its tangent plane and use this as the

value of the curvature of the surface. The curvature will be negative or positive according as the surface is above or below the tangent plane. This curvature is defined by measuring the height, d_{QP}, of $Q(u_0+du, v_0+dv)$ above the tangent plane at $P(u_0, v_0)$, see Fig. 2.14. We define the *second fundamental form* by

$$ds_n^2 = -2d_{QP} = -2d\mathbf{x}\cdot\mathbf{N}\ . \tag{2.65}$$

Notice that ds_n^2 is *not* positive definite in that it does not satisfy either the positivity requirement, given by Eq.(2.37), or the definiteness requirement, given by Eq.(2.38). For example, if on the surface it is possible to choose $d\mathbf{x}\perp\mathbf{N}$, for non-zero $d\mathbf{x}$ we can have $ds_n^2=0$. Further, if there is an obtuse angle between $d\mathbf{x}$ and $\mathbf{N}$, $ds_n^2<0$. Using Eq.(2.35) and bearing in mind that $\mathbf{x}_u\cdot\mathbf{N}=\mathbf{x}_v\cdot\mathbf{N}=0$ as $\mathbf{x}_u, \mathbf{x}_v\perp\mathbf{N}$, we get

$$\begin{aligned} ds_n^2 &= -\left(\mathbf{x}_{uu}\cdot\mathbf{N}du^2 + 2\mathbf{x}_{uv}\cdot\mathbf{N}dudv + \mathbf{x}_{vv}\cdot\mathbf{N}dv^2\right) \\ &= \widetilde{E}du^2 + 2\widetilde{F}dudv + \widetilde{G}dv^2 \\ &= \widetilde{g}_{ab}du^a du^b\ . \end{aligned} \tag{2.66}$$

In Relativity we use non-positive definite metrics by analogy with this (second fundamental) form.

The *normal curvature* of the surface is defined to be

$$\begin{aligned} \kappa_n &= ds_n^2/ds^2 \\ &= \frac{\widetilde{g}_{ab}du^a du^b}{g_{cd}du^c du^d} = \frac{\widetilde{E}du^2 + 2\widetilde{F}dudv + \widetilde{G}dv^2}{Edu^2 + 2Fdudv + Gdv^2} \\ &= \frac{\widetilde{E} + 2\widetilde{F}\lambda + \widetilde{G}\lambda^2}{E + 2F\lambda + G\lambda^2}\ , \end{aligned} \tag{2.67}$$

where $\lambda = dv/du$ gives the choice of a particular direction on the surface. Here $\lambda=0$ corresponds to the coordinate line $v=$ constant, while $\lambda=\infty$ corresponds to the coordinate line $u=$ constant. For a given space curve $\phi(u,v)=0$ we have, at $P(u_0,v_0)$,

$$\lambda = \frac{dv}{du} = -\frac{\phi_u(u_0,v_0)}{\phi_v(u_0,v_0)}\ . \tag{2.68}$$

Clearly κ_n depends upon properties of the surface, but it is also a function of the choice of curve, through the choice of λ.

A property of the surface, independent of the choice of the curve is the extremal (maximum or minimum) value of the normal curvature. This is obtained by setting the derivative of κ_n with respect to λ equal to zero. Differentiating Eq.(2.67) and putting it equal to zero gives

$$[(E+F\lambda)+\lambda(F+G\lambda)](\widetilde{F}+\widetilde{G}\lambda) = [(\widetilde{E}+\widetilde{F}\lambda)+\lambda(\widetilde{F}+\widetilde{G}\lambda)](F+G\lambda)\ . \tag{2.69}$$

Note that on expansion the last terms on either side cancel out, to yield

$$\left(\widetilde{E}+\widetilde{F}\lambda\right) = \frac{E+F\lambda}{F+G\lambda}\left(\widetilde{F}+\widetilde{G}\lambda\right)\ . \tag{2.70}$$

Inserting this expression into Eq.(2.67) and simplifying, the extremal values of the normal curvature $\kappa_\pm$, are given by

$$\kappa_\pm = \frac{\widetilde{F}+\widetilde{G}\lambda_\pm}{F+G\lambda_\pm} = \frac{\widetilde{E}+\widetilde{F}\lambda_\pm}{E+F\lambda_\pm}\ , \tag{2.71}$$

where the extremal directions satisfy Eq.(2.70). This gives

$$\lambda_{\pm} = \frac{\left(\widetilde{E}G - E\widetilde{G}\right) \pm \sqrt{\left(\widetilde{E}G - E\widetilde{G}\right)^2 - 4\left(E\widetilde{F} - \widetilde{E}F\right)\left(F\widetilde{G} - \widetilde{F}G\right)}}{2\left(F\widetilde{G} - \widetilde{F}G\right)} \ . \quad (2.72)$$

The $\kappa_{\pm}$ are called the *principal curvatures* and the directions, given by $\lambda_{\pm}$, the *principal directions*.

Eq.(2.71) can be directly solved, except if $\lambda_{\pm} = 0$ or ∞, when we need to go back and evaluate the principal curvatures from first principles. If a principal direction is along v = constant, so that $dv = 0$ in Eq.(2.67), or along u = constant, so that $du = 0$, in Eq.(2.67), i.e. $\lambda = 0$ or $\lambda = \infty$, we get

$$\kappa_{\pm} = \widetilde{E}/E \ \ , \ \ \widetilde{G}/G \ \ , \qquad (\kappa_+ \geq \kappa_-). \quad (2.73)$$

This will be the case if $F = \widetilde{F} = 0$ and $\lambda_+ = 0$, $\lambda_- = \infty$ (or vice versa). Since we could always re-parameterize our surfaces so that the u = constant and v = constant lines are the principal directions, whatever results of a geometrical nature hold in this choice, hold for all choices. In this choice, since $F = 0$, the principal directions are orthogonal. Hence the principal directions are always orthogonal. They would, generally, form a preferred basis on the surface.

At a given point the second fundamental form may be positive definite ($ds_n^2 > 0$), positive indefinite ($ds_n^2 \geq 0$), negative indefinite ($ds_n^2 \leq 0$), or negative definite ($ds_n^2 < 0$). It is said to be indefinite if it can change from positive to negative with changes of direction. In the first and last cases $\kappa_n \neq 0$ and does not change sign. Such points are called *elliptical*. In the second and third cases κ_n does not change sign but does take the value zero once. Such points are called *parabolic*. In the fifth case, κ_n changes sign and takes the value zero (twice). Such points are called *hyperbolic*. A direction along which $\kappa_n = 0$ is called an *asymptotic direction*. There is clearly no asymptotic direction for an elliptic point. For a parabolic point one of the principal directions is asymptotic as one of the principal curvatures is zero. For a hyperbolic point there are two asymptotic directions, neither of which is a principal direction. Generally the asymptotic directions are obtained by putting $\kappa_n = 0$ in Eq.(2.67), i.e.

$$\lambda_{\pm} = \frac{-\widetilde{F}_{\pm}\sqrt{\widetilde{F}^2 - \widetilde{E}\widetilde{G}}}{\widetilde{G}} \ . \quad (2.74)$$

There is no solution if

$$\widetilde{g} := \det\left(\widetilde{g}_{ab}\right) = \widetilde{E}\widetilde{G} - \widetilde{F}^2 > 0 \ , \quad (2.75)$$

one solution, namely

$$\lambda = -\widetilde{F}/\widetilde{G} \quad (2.76)$$

if $\widetilde{g} = 0$, and two solutions if $\widetilde{g} < 0$. Hence the point will be elliptic, parabolic or hyperbolic according as

$$\widetilde{g} \geq 0 \ . \quad (2.77)$$

If all points of a space are elliptic/parabolic/hyperbolic the space is said to be elliptic/parabolic/hyperbolic.

We now consider some illustrative examples so as to make the discussion of the theory of surfaces more concrete. The examples chosen are specially simple and useful in their own right for later developments.

2.6 Examples

Example 1: A sphere centered at the origin, with radius a, is given by

$$x^2 + y^2 + z^2 = a^2 \ . \tag{2.78}$$

This surface can be parameterized using spherical coordinates with $r = a$, $\theta = u$, $\phi = v$ (see Fig 2.15), to give

$$\mathbf{x}(u, v) = a(\sin u \cos v, \ \sin u \sin v, \ \cos u) \ . \tag{2.79}$$

The same parameterization could be used for a sphere of radius a with a centre at any point. Taking

$$\begin{aligned} \mathbf{x}_u &= a(\cos u \cos v, \ \cos u \sin v, \ -\sin u) \ , \\ \mathbf{x}_v &= a(-\sin u \sin v, \ \sin u \cos v, \ 0) \ . \end{aligned}$$

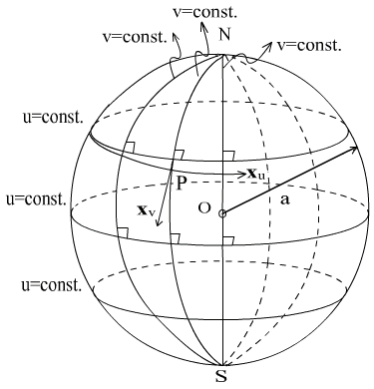


Figure 2.15: A sphere parameterized by spherical coordinates $u = \theta$, $v = \varphi$. Clearly $\mathbf{x} \perp \mathbf{x}$ everywhere.

The matrix for determining singularities is, therefore,

$$\mathcal{M} := \begin{pmatrix} \mathbf{x}_u \\ \mathbf{x}_v \end{pmatrix} = a \begin{pmatrix} \cos u \cos v & \cos u \sin v & -\sin u \\ -\sin u \sin v & \sin u \cos v & 0 \end{pmatrix} \ ,$$

which is of rank 2 except at $u = 0, \pi$ i.e. the North and South poles of the sphere. Since all points of the sphere are geometrically equivalent these must be coordinate singularities.. From Eq.(2.63)

$$\begin{aligned} \mathbf{N} &= \frac{a^2 \left(\sin^2 u \cos v, \ \sin^2 u \sin v, \ \sin u \cos u\right)}{a^2 \sqrt{\sin^4 u \cos^2 v + \sin^4 u \sin^2 v + \sin^2 u \cos^2 u}} \\ &= (\sin u \cos v, \ \sin u \sin v, \ \cos u) = \mathbf{x}/a \ . \end{aligned}$$

Thus the radius vector, $\mathbf{x}$, is normal to the tangent plane. From Eq.(2.36)

$$E = a^2 \ , \quad F = 0 \ , \quad G = a^2 \sin^2 u \ .$$

Therefore

$$g_{ab} = a^2 \begin{pmatrix} 1 & 0 \vdots 0 & \sin^2 u \end{pmatrix} , \tag{2.80}$$

$$ds^2 = a^2 \left(du^2 + \sin^2 u dv^2 \right) . \tag{2.81}$$

Taking second derivatives we have

$$\begin{aligned} \mathbf{x}_{uu} &= -a \left(\sin u \cos v, \ \sin u \sin v, \ \cos u \right) = -\mathbf{x} , \\ \mathbf{x}_{uv} &= a \cos u \left(-\sin v, \ \cos u, \ 0 \right) , \\ \mathbf{x}_{vv} &= -a \sin u \left(\cos v, \ \sin v, \ 0 \right) . \end{aligned}$$

From Eq.(2.66)

$$\widetilde{E} = a , \quad \widetilde{F} = 0 \quad \widetilde{G} = a \sin^2 u$$

Thus

$$\widetilde{g}_{ab} = a \begin{pmatrix} 1 & 0 \vdots 0 & \sin^2 u \end{pmatrix} , \tag{2.82}$$

$$ds_n^2 = a \left(du^2 + \sin^2 u dv^2 \right) , \tag{2.83}$$

$$\kappa_n = 1/a . \tag{2.84}$$

for all directions. The space is elliptic and has constant normal curvature, which is the inverse of the radius. There are no preferred directions, all directions being principal directions. The (u, v) system is an orthogonal coordinate system as $F = 0$.

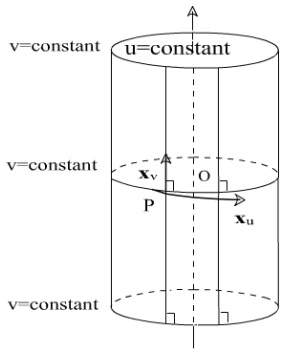


Figure 2.16: A cylinder parameterized by cylindrical polar coordinates $u = \theta$, $v = z$. Clearly $x_u \perp x$ everywhere.

Example 2 : A right circular cylinder with origin on the axis is given by (see Fig 2.16)

$$x^2 + y^2 = a^2 , \tag{2.85}$$

$$\mathbf{x}(u, v) = a \left(\cos u, \ \sin u, \ v \right) , \tag{2.86}$$

$$\begin{aligned} \mathbf{x}_u &= a \left(-\sin u, \ \cos u, \ 0 \right) , \\ \mathbf{x}_v &= a \left(0, \ 0, \ 1 \right) , \\ \mathcal{M} &= a \begin{pmatrix} -\sin u & \cos u & 0 \\ 0 & 0 & 1 \end{pmatrix} . \end{aligned}$$

which is always rank 2. Hence there are no singular points.

$$\begin{aligned}
\mathbf{N} &= (\cos u,\ \sin u,\ 0) = \mathbf{x}(u,0)/a\ , \\
E &= a^2\ , \quad F = 0\ , \quad G = a^2\ , \\
g_{ab} &= a^2 \left(\begin{array}{cc:cc} 1 & 0 & 0 & 1 \end{array} \right)\ , \qquad (2.87) \\
ds^2 &= a^2\left(du^2 + dv^2\right)\ . \qquad (2.88)
\end{aligned}$$

Clearly (u, v) form an orthogonal coordinate system.

$$\begin{aligned}
\mathbf{x}_{uu} &= -a(\cos u,\ \sin u,\ 0) = -\mathbf{x}(u,0)\ , \\
\mathbf{x}_{uv} &= 0 \qquad\qquad \mathbf{x}_{vv} = 0\ , \\
\widetilde{E} &= a\ , \quad \widetilde{F} = 0, \widetilde{G} = 0\ , \\
\widetilde{g}_{ab} &= a \left(\begin{array}{cc:cc} 1 & 0 & 0 & 0 \end{array} \right)\ , \qquad (2.89) \\
ds_n^2 &= a du^2\ . \qquad (2.90)
\end{aligned}$$

Thus

$$\kappa_n(\lambda) = 1/a\left(1 + \lambda^2\right) \qquad (2.91)$$

is the normal curvature. The principal curvatures obtained for an orthogonal coordinate system by setting $\lambda = 0, \infty$ or from Eq.(2.75), are

$$\kappa_+ = \widetilde{E}/E = 1/a \quad , \quad \kappa_- = \widetilde{G}/G = 0\ , \qquad (2.92)$$

and the corresponding principal directions are $v =$ constant and $u =$ constant lines. The space is parabolic and has only one asymptotic direction, namely $u =$ constant. The parameterization has been obtained from cylindrical coordinates by choosing $r = a$, $\theta = u$, $z = v$.

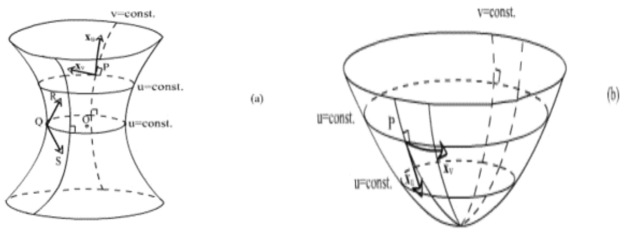


Figure 2.17: (a) The hyperboloid $x^2+y^2-z^2 = a^2$ parameterized by orthogonal (hyperbolic) coordinates u, v. There are no other "branches" of this hyperboloid. QR and QS give asymptotic directions at Q, along which the normal curvature is zero. $\mathbf{x}_u$ and $\mathbf{x}_v$ are shown as tangent to the coordinate lines at point P. (b) The hyperboloid $z^2 - x^2 - y^2 = a^2$ parameterized by orthogonal (hyperbolic) coordinates has two "branches". The one shown and its mirror image in the xy-plane. There are no asymptotic directions here.

Example 3: A right hyperboloid of one sheet, with origin at the centre (see Fig. 2.17 a) is given by

$$x^2 + y^2 - z^2 = a^2\ . \qquad (2.93)$$

Therefore

$$\begin{aligned}
\mathbf{x}(u,v) &= a(\cosh u\cos v,\ \cosh u\sin v,\ \sinh u)\ , \qquad (2.94)\\
\mathbf{x}_u &= a(\sinh u\cos v,\ \sinh u\sin v,\ \cosh u)\ ,\\
\mathbf{x}_v &= -a\cosh u(\sin v,\ -\cos v,\ 0)\ ,\\
\mathcal{M} &= a\begin{pmatrix} \sinh u\cos v & \sinh u\sin v & \cosh u\\ -\cosh u\sin v & \cosh u\cos v & 0\end{pmatrix}\ ,
\end{aligned}$$

which is always of rank 2. Hence there is no singularity.

$$\begin{aligned}
\mathbf{N} &= -\frac{(\cosh u\cos v,\ \cosh u\sin v,\ -\sinh u)}{\sqrt{\cosh 2u}}\ , \qquad (2.95)\\
E &= a^2\cosh 2u,\quad F=0,\quad G=a^2\cosh^2 u\ ,\\
g_{ab} &= a^2\left(\begin{array}{cc:c} \cosh 2u & 0 & 0\\ 0 & \cosh^2 u & \end{array}\right)\ , \qquad (2.96)\\
ds^2 &= a^2\left(\cosh 2u\ du^2+\cosh^2 u\ dv^2\right)\ , \qquad (2.97)\\
\mathbf{x}_{uu} &= a(\cosh u\cos v,\ \cosh u\sin v,\ \sinh u) = -\mathbf{x}(u,v)\ ,\\
\mathbf{x}_{uv} &= -a\sinh u(\sin v,\ -\cos u,\ 0)\ ,\\
\mathbf{x}_{vv} &= -a\cosh u(\cos v,\ \sin v,\ 0)\ ,\\
\widetilde{E} &= \frac{a}{\sqrt{\cosh 2u}}\ ,\quad \widetilde{F}=0\ ,\quad \widetilde{G}=\frac{-a\cosh^2 u}{\sqrt{\cosh 2u}}\ ,\\
\widetilde{g}_{ab} &= \frac{a}{\sqrt{\cosh 2u}}\left(\begin{array}{cc:c} 1 & 0 & 0\\ 0 & -\cosh^2 u & \end{array}\right)\ , \qquad (2.98)\\
ds_n^2 &= \frac{a}{\sqrt{\cosh 2u}}\left(du^2-\cosh^2 u dv^2\right)\ . \qquad (2.99)
\end{aligned}$$

Thus

$$\begin{aligned}
\kappa_n(\lambda) &= \frac{1}{a\sqrt{\cosh 2u}}\frac{\left(1-\lambda^2\cosh^2 u\right)}{\left(\cosh 2u+\lambda^2\cosh^2 u\right)}\ , \qquad (2.100)\\
\kappa_+ &= \frac{1}{a\left(\cosh 2u\right)^{3/2}}\ ,\quad \kappa_- = \frac{-1}{a\left(\cosh 2u\right)^{1/2}}\ , \qquad (2.101)
\end{aligned}$$

with the principal directions v = constant and u = constant. The space is hyperbolic and hence there are two asymptotic directions given by

$$\lambda_\pm = \left(\frac{dv}{du}\right)_\pm = \pm\sqrt{\frac{-\widetilde{E}}{\widetilde{G}}} = \pm\frac{1}{\cosh u}\ . \qquad (2.102)$$

Notice that at $u=0$ the directions are 45^o, or $\pi/4$ lines, but as u increases, the angles become steeper (since λ decreases as u increases). Clearly the (u,v) coordinates form an orthogonal system.

Alternatively, the right hyperboloid of two sheets is given by (see Fig. 2.17 b)

$$-x^2-y^2+z^2=a^2\ . \qquad (2.103)$$

Then

$$\begin{aligned}
\mathbf{x}(u,v) &= a(\sinh u\cos v,\ \sinh u\sin v,\ \cosh u)\ , && (2.104)\\
\mathbf{x}_u &= a(\cosh u\cos v,\ \cosh u\sin v,\ \sinh u)\ ,\\
\mathbf{x}_v &= a\sinh u(-\sin v,\ \cos v,\ 0)\ ,\\
\mathcal{M} &= a\begin{pmatrix} \cosh u\cos v & \cosh u\sin v & \sinh u \\ -\sinh u\sin v & \sinh u\cos v & 0 \end{pmatrix}
\end{aligned}$$

has a coordinate singularity at $u = 0$. Now

$$\begin{aligned}
\mathbf{N} &= \frac{(\sinh u\cos v,\ \sinh u\sin v,\ -\cosh u)}{\sqrt{\cosh 2u}} && (2.105)\\
E &= a^2\cosh 2u\ ,\quad F = 0\ ,\quad G = a^2\sinh^2 u\ ,\\
g_{ab} &= a^2\left(\begin{array}{cc:c} \cosh 2u & 0 & 0 \\ 0 & \sinh^2 u & \end{array}\right)\ , && (2.106)\\
ds^2 &= a^2\left(\cosh 2u du^2 + \sinh^2 u dv^2\right)\ , && (2.107)\\
\mathbf{x}_{uu} &= a(\sinh u\cos v,\ \sinh u\sin v,\ \cosh u) = \mathbf{x}(u,v)\ ,\\
\mathbf{x}_{uv} &= -a\cosh u(\sin v,\ -\cos v,\ 0)\ ,\\
\mathbf{x}_{vv} &= -a\sinh u(\cos v,\ \sin v,\ 0)\ ,\\
\widetilde{E} &= \frac{a}{\sqrt{\cosh 2u}}\ ,\quad \widetilde{F} = 0,\quad \widetilde{G} = \frac{a\sinh^2 u}{\sqrt{\cosh 2u}}\ ,\\
\widetilde{g}_{ab} &= \frac{a}{\sqrt{\cosh 2u}}\left(\begin{array}{cc:c} 1 & 0 & 0 \\ 0 & \sinh^2 u & \end{array}\right)\ , && (2.108)\\
ds_n^2 &= \frac{a}{\sqrt{\cosh 2u}}\left(du^2 + \sinh^2 u dv^2\right)\ . && (2.109)
\end{aligned}$$

Thus

$$\begin{aligned}
\kappa_n(\lambda) &= \frac{1}{a\sqrt{\cosh 2u}}\frac{\left(1+\lambda^2\sinh^2 u\right)}{\left(\cosh 2u+\lambda^2\sinh^2 u\right)}\ , && (2.110)\\
\kappa_+ &= \frac{1}{a\left(\cosh 2u\right)^{3/2}}\quad ,\quad \kappa_- = \frac{1}{a\left(\cosh 2u\right)^{1/2}}\ , && (2.111)
\end{aligned}$$

with the principal directions v = constant and u = constant. The space is elliptical as $\kappa = 1/\cosh^2 2u$. Thus there is no asymptotic direction.

It is worth remarking that the hyperboloid of two sheets is *not* a hyperbolic space in real variables, but it is in terms of complex variables.

2.7 Gauss' Formulation of the Geometry of Surfaces

Much of the theory of surfaces was developed by Gauss. He showed that the geometry of the surface could be described intrinsically, without reference to the third dimension (i.e. without using the third basis vector, $\mathbf{N}$), which forms the basis of the theory of curved spaces developed by Riemann and used by Einstein and Grossmann in GR. The essential point is that we do not have to regard the curved space as embedded in a higher dimensional flat space. Gauss also found properties of the surface invariant under coordinate transformations.

Since physical laws must be invariant under changes of reference frame, such invariant quantities are of prime importance. He also developed formulae for surfaces analogous to the Frenet-Serret formulae for curves. Now, instead of the moving triad depending on one variable, we have the two tangent vectors along the surface and one orthogonal to it depending on two independent variables. Thus there would be six equations instead of five. However, since the order of differentiation is irrelevant for analytic functions of two variables, the equations for $\mathbf{x}_{uv}$ and $\mathbf{x}_{vu}$ are identical. Hence there are only five independent equations in all.

We shall write the basis vectors along the surface as $\mathbf{e}_a(u^b)$ $(a, b = 1, 2)$, as discussed in section.4. These can be differentiated with respect to u^b to give $\mathbf{e}_{ab}(u^c)$. These vectors can be expressed as linear combinations of the basis at the point $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{N})$,

$$\mathbf{e}_{ab} = \{_a \ ^c \ _b\} \, \mathbf{e}_c + \gamma_{ab}\mathbf{N} \ , \tag{2.112}$$

where γ_{ab} are four coefficients such that $\gamma_{ab} = \gamma_{ba}$ and $\{_a \ ^c \ _b\}$ are *Christoffel symbols* (*of the second kind*), which are eight coefficients such that $\{_a \ ^c \ _b\} = \{_b \ ^c \ _a\}$. We need to evaluate all twelve of these coefficients, only nine of which are independent.

To evaluate the nine independent coefficients, of which six are independent, let us re-write Eqs.(2.36) and (2.66) in terms of the new notation as

$$\mathbf{e}_c \cdot \mathbf{e}_d \ = \ g_{cd} \ , \tag{2.113}$$

$$\mathbf{e}_{ab} \cdot \mathbf{N} \ = \ -\widetilde{g}_{ab} \ . \tag{2.114}$$

Now, since

$$\mathbf{e}_c \cdot \mathbf{N} = 0 \ \ , \ \ \mathbf{N} \cdot \mathbf{N} = 1 \ , \tag{2.115}$$

taking the scalar product of Eq. (2.112) with $\mathbf{N}$, using Eqs. (2.114) and (2.115) gives

$$\gamma_{ab} = -\widetilde{g}_{ab} \ . \tag{2.116}$$

Thus Eq.(2.112) becomes

$$\mathbf{e}_{ab} = \{_a \ ^c \ _b\} \, \mathbf{e}_c - \widetilde{g}_{ab}\mathbf{N} \ . \tag{2.117}$$

Taking the scalar product of Eq.(2.117) with $\mathbf{e}_d$, using Eqs. (2.113) and (2.114),

$$\mathbf{e}_d \cdot \mathbf{e}_{ab} := [d \ , \ ab] = \{_a \ ^c \ _b\} \, g_{cd} \ , \tag{2.118}$$

where $[d \ , \ ab]$ are the *Christoffel symbols of the first kind*. The order of appearance of these symbols is reversed in my presentation. Historically, as might be expected, the first kind preceded the second kind.

The Christoffel symbols will be evaluated generally. However, a clearer understanding of the process of calculation is provided by explicit computation in the earlier notation using $\mathbf{x}_u$, $\mathbf{x}_v$, E, F, G, etc. We first differentiate the equation defining E,

$$\mathbf{x}_u \cdot \mathbf{x}_u = E \ , \tag{2.119}$$

with respect to u and v, to give

$$\left.\begin{aligned} \mathbf{x}_u \cdot \mathbf{x}_{uu} = [1 \ , \ 11] = \tfrac{1}{2}E_u = \tfrac{1}{2}g_{11,1} \\ \mathbf{x}_u \cdot \mathbf{x}_{uv} = [1 \ , \ 12] = \tfrac{1}{2}E_v = \tfrac{1}{2}g_{11,2} \end{aligned}\right\} \ . \tag{2.120}$$

Similarly, using the equation defining G,

$$\mathbf{x}_v \cdot \mathbf{x}_v = G \ , \tag{2.121}$$

we get

$$\left.\begin{aligned} \mathbf{x}_v \cdot \mathbf{x}_{uu} &= [2\ ,\ 12] = \tfrac{1}{2} G_u = \tfrac{1}{2} g_{22,1} \\ \mathbf{x}_v \cdot \mathbf{x}_{vv} &= [2\ ,\ 22] = \tfrac{1}{2} G_v = \tfrac{1}{2} g_{22,2} \end{aligned}\right\} \ . \tag{2.122}$$

Now consider the derivatives of the equation defining F relative to u and v,

$$\mathbf{x}_u \cdot \mathbf{x}_v = G \ , \tag{2.123}$$

$$\mathbf{x}_u \cdot \mathbf{x}_{uv} + \mathbf{x}_v \cdot \mathbf{x}_{uu} = F_u \ , \ \mathbf{x}_u \cdot \mathbf{x}_{vv} + \mathbf{x}_v \cdot \mathbf{x}_{uv} = F_v \ , \tag{2.124}$$

or

$$[1\ ,\ 12] + [2\ ,\ 11] = g_{12,1} \ \ , \ [1\ ,\ 22] + [2\ ,\ 12] = g_{12,2} \ . \tag{2.125}$$

Using Eqs.(2.120) and (2.122), Eqs.(2.125) reduce to

$$[2\ ,\ 11] = g_{12,1} - \frac{1}{2} g_{11,2} \ , \ [2\ ,\ 12] = g_{12,2} - \frac{1}{2} g_{22,1} \ . \tag{2.126}$$

Since $[1\ ,\ 12] = [1\ ,\ 21]$ and $[2\ ,\ 12] = [2\ ,\ 21]$ all eight Christoffel symbols of the first kind have been determined.

In the index notation, differentiate Eq.(2.113) relative to u^b to obtain

$$[d\ ,\ bc] + [c\ ,\ bd] = g_{cd,b} \ . \tag{2.127}$$

In the case that $d = c$, the terms on the left are equal. Thus

$$[d\ ,\ bc] = \frac{1}{2} g_{cd,b} \ (\text{when } c = d). \tag{2.128}$$

Since the indices only take values 1 and 2, at least two of the three indices above are equal. Thus the case missed out in Eq.(2.128) is when $b = c \neq d$. Notice that when $b = d \neq c$ we can use the symmetry property of $[d\ ,\ bc]$ to write it as $[d\ ,\ cb]$. Now again the first and third indices will be equal and hence Eq.(2.128) will apply with b and c interchanged. In the case $b = c \neq d$, Eq.(2.128) applies for the second Christoffel symbol in Eq.(2.127) with $d \to b$, $b \to c$, $c \to d$, so that

$$[d\ ,\ bc] = g_{cd,b} - \frac{1}{2} g_{bc,d} \ (\text{when } b = c \neq d). \tag{2.129}$$

Now, when $b = c$ we can write

$$g_{cd,b} = \frac{1}{2} \left(g_{cd,b} - g_{dc,b} \right) \ . \tag{2.130}$$

Eqs.(2.128) and (2.130) can be combined in the single equation

$$[d\ ,\ bc] = \frac{1}{2} \left(g_{cd,b} + g_{bd,c} - g_{bc,d} \right) \ , \tag{2.131}$$

as the last two terms cancel when $c = d$ and Eq.(2.129) reduces to Eq.(2.131) in the remaining case $b = c \neq d$.

To evaluate the Christoffel symbols of the second kind we need to invert the partitioned row matrix, g_{cd}, in Eq.(2.118). We denote the inverse by g^{de} so that its product with g_{cd} gives the Kronecker delta (the identity in index notation)

$$g_{cd}g^{de} = \delta^e_c \ . \tag{2.132}$$

In partitioned matrix form it can be written as the column matrix

$$g^{de} = g^{ed} = \frac{1}{g}\begin{pmatrix} G \\ -F \\ - - - \\ -F \\ E \end{pmatrix} \tag{2.133}$$

where $g = \det(g_{ab})$, which is strictly positive (as shown earlier, see section 3, Eqs.(**??**)–(**??**)) for non-singular points. To simplify the notation we shall henceforth write g_{ab} and g^{ab} as square matrices.

Multiplying Eq.(2.118) by g^{de} we get

$$\begin{aligned} g^{de}\left[\ d\ ,\quad ab\ \right] &= g^{de}g_{cd}\left\{_a\ ^c\ _b\right\} \\ &= \delta^e_c\left\{_a\ ^c\ _b\right\} = \left\{_a\ ^e\ _b\right\} \ . \end{aligned} \tag{2.134}$$

Thus, multiplying Eq.(2.131) by g^{ad}, using Eq.(2.134) with the appropriate change of indices, we obtain

$$\left\{_b\ ^a\ _c\right\} = \frac{1}{2}g^{ad}\left(g_{bd,c} + g_{cd,b} - g_{bc,d}\right) \ . \tag{2.135}$$

Eqs.(2.117) with the symbol defined by Eqs.(2.135) are *Gauss' equations* written in the modern index notation.

So far we have obtained three of the required five equations. For the other two we will write the derivative of $\mathbf{N}(u^a)$ relative to u^a as a linear combination of the basis vectors

$$\mathbf{N}_a = \alpha_a\ ^b\ \mathbf{e}_b + \beta_a\mathbf{N} \ . \tag{2.136}$$

Differentiating the second of Eqs.(2.115) relative to u^a gives

$$\mathbf{N}\cdot\mathbf{N}_a = 0 \ . \tag{2.137}$$

Taking the scalar product of Eq.(2.136) with $\mathbf{N}$ using Eq.(2.137) gives

$$\beta_a = 0 \ . \tag{2.138}$$

Taking the scalar product of Eq.(2.136) with $\mathbf{e}_c$ using Eqs.(2.113) and (2.115) gives

$$\mathbf{N}_a\cdot\mathbf{e}_c = \alpha_a\ ^b\ g_{bc} \ . \tag{2.139}$$

Also, differentiating the first of Eqs.(2.115) relative to u^a gives

$$\mathbf{N}_a\cdot\mathbf{e}_c + \mathbf{N}\cdot\mathbf{e}_{ac} = 0 \ . \tag{2.140}$$

Using Eqs.(2.114), (2.139) and (2.140) we obtain

$$\alpha_a\ ^b\ g_{bc} = \widetilde{g}_{ac} \ . \tag{2.141}$$

Multiplying Eq.(2.141) with g^{cd}, using Eq.(2.132) and the index substitution role of the Kronecker delta, we evaluate the coefficients

$$\alpha_a{}^d = \widetilde{g}_{ac} g^{cd} \ . \tag{2.142}$$

Using Eqs.(2.138) and (2.142) we can re-write Eq.(2.136) as

$$\mathbf{N}_a = \widetilde{g}_{ac} g^{bc} \mathbf{e}_b \ . \tag{2.143}$$

These (two) equations (for $a = 1, 2$) are *Weingarten's equations*. The two sets of equations are jointly known as the *Gauss-Weingarten equations*. This nomenclature is useful to distinguish the Gauss equations here from those to be discussed in the next section.

2.8 The Gauss-Codazzi Equations and Gauss' Theorem

To determine manifestly invariant expressions for some invariant quantities, Gauss obtained some identities. These identities use the second derivatives of the basis vectors, and the fact that the order of the second and third derivatives of $\mathbf{x}$ does not matter. For example

$$(\mathbf{x}_{uu})_v = (\mathbf{x}_{uv})_u \ , \ \ (\mathbf{x}_{uv})_v = (\mathbf{x}_{vv})_u \ , \tag{2.144}$$

or equivalently

$$(\mathbf{e}_{11})_{,2} = (\mathbf{e}_{12})_{,1} \ , \ \ (\mathbf{e}_{12})_{,2} = (\mathbf{e}_{22})_{,1} \ , \tag{2.145}$$

where "," denotes partial differentiation relative to the corresponding coordinate.

Differentiating Eq.(2.117) with an appropriate choice of indices and using it and Eq. (2.118), we obtain

$$\left.\begin{aligned} \mathbf{e}_{11,2} &= (\{_1{}^c{}_1\}\,\mathbf{e}_c - \widetilde{g}_{11}\mathbf{N})_{,2} \\ &= \{_1{}^c{}_1\}_{,2}\,\mathbf{e}_c + \{_1{}^c{}_1\}\,\mathbf{e}_{c,2} - \widetilde{g}_{11,2}\mathbf{N} - \widetilde{g}_{11}\mathbf{N}_{,2} \\ &= \begin{array}{l} \{_1{}^c{}_1\}_{,2}\,\mathbf{e}_c + \{_1{}^c{}_1\}\{_c{}^d{}_2\}\,\mathbf{e}_d - \{_1{}^c{}_1\}\,\widetilde{g}_{c2}\mathbf{N} \\ -\widetilde{g}_{11,2}\mathbf{N} - \widetilde{g}_{11}\widetilde{g}_{2c}g^{bc}\mathbf{e}_b \ . \end{array} \end{aligned}\right\} \tag{2.146}$$

To simplify the expression by collecting together all the terms depending on each of the basis vectors, we need to re-label some of the "dummy indices" over which summation is implicit. For example, in the first term on the right

$$\{_1{}^c{}_1\}_{,2}\,\mathbf{e}_c = \{_1{}^1{}_1\}_{,2}\,\mathbf{e}_1 + \{_1{}^2{}_1\}_{,2}\,\mathbf{e}_2 = \{_1{}^d{}_1\}_{,2}\,\mathbf{e}_d \ . \tag{2.147}$$

Similarly, for the last term on the right in Eq.(2.121) the dummy index b can be replaced by d. Arbitrary re-labelling of dummy indices is permissible provided the new index chosen is not already in use. Having four indices the same (or even three for that matter) would cause confusion as to which index is summed up with which. For example

$$A_a B^a C_a D^a = A_1 B^1 C_1 D^1 + A_2 B^2 C_2 D^2 \ , \tag{2.148}$$

which is not the same as

$$A_a B^a C_b D^b = \left(A_1 B^1 + A_2 B^2\right)\left(C_1 D^1 + C_2 D^2\right) \ , \tag{2.149}$$

both of which are different from

$$A_a B^b C_b D^a = \left(A_1 D^1 + A_2 D^2\right)\left(B^1 C_1 + B^2 C_2\right) \ . \tag{2.150}$$

As another example, for the last term on the right in Eq.(2.121) the dummy index b can *not* be replaced by c as that would make four indices the same.

Using the re-labelling given in Eq.(2.122) and changing b to d in the last term on the right in Eq.(2.121), we can write Eq.(2.121) as

$$\begin{aligned} e_{11,2} &= \left[\left\{_1{}^d{}_1\right\}_{,2} + \left\{_1{}^c{}_1\right\}\left\{_c{}^d{}_2\right\} - \widetilde{g}_{11}\widetilde{g}_{2c}g^{cd}\right]\mathbf{e}_d \\ &\quad - \left[\widetilde{g}_{11,2} + \left\{_1{}^c{}_1\right\}\widetilde{g}_{c2}\right]\mathbf{N} \ . \end{aligned} \tag{2.151}$$

Following the same procedure to evaluate $e_{12,1}$ we have

$$\begin{aligned} e_{12,1} &= \left(\left\{_1{}^c{}_2\right\}\mathbf{e}_c - \widetilde{g}_{12}\mathbf{N}\right)_{,1} \\ &= \left[\left\{_1{}^d{}_2\right\}_{,1} + \left\{_1{}^c{}_2\right\}\left\{_c{}^d{}_1\right\} - \widetilde{g}_{12}\widetilde{g}_{1c}g^{cd}\right]\mathbf{e}_d \\ &\quad - \left[\widetilde{g}_{12,1} + \left\{_1{}^c{}_2\right\}\widetilde{g}_{c1}\right]\mathbf{N} \ . \end{aligned} \tag{2.152}$$

Comparing the coefficients of $\mathbf{N}$ in Eqs.(2.126) and (2.127) using Eq.(2.119) we obtain the *Codazzi equation*

$$\widetilde{g}_{11,2} + \left\{_1{}^c{}_1\right\}\widetilde{g}_{c2} = \widetilde{g}_{12,1} + \left\{_1{}^c{}_2\right\}\widetilde{g}_{c1} \ , \tag{2.153}$$

or in terms of the other notation

$$\widetilde{E}_v + \left\{_1{}^1{}_1\right\}\widetilde{F} + \left\{_1{}^2{}_1\right\}\widetilde{G} = \widetilde{F}_u + \left\{_1{}^1{}_2\right\}\widetilde{E} + \left\{_1{}^2{}_2\right\}\widetilde{F} \ . \tag{2.154}$$

Now, comparing the coefficients of $\mathbf{e}_d$ in Eqs.(2.151) and (2.152) using Eq.(2.144) and making appropriate transpositions

$$\begin{aligned} \left(\widetilde{g}_{11}\widetilde{g}_{2c} - \widetilde{g}_{12}\widetilde{g}_{1c}\right)g^{cd} &= -\left[\left\{_1{}^d{}_1\right\}_{,2} - \left\{_1{}^d{}_2\right\}_{,1} + \left\{_1{}^c{}_1\right\}\left\{_c{}^d{}_2\right\}\right. \\ &\quad \left. - \left\{_1{}^c{}_2\right\}\left\{_c{}^d{}_1\right\}\right] \ . \end{aligned} \tag{2.155}$$

The left side of this equation can be simplified by evaluating the implicit summation

$$\left(\widetilde{g}_{11}\widetilde{g}_{21} - \widetilde{g}_{12}\widetilde{g}_{11}\right)g^{1d} + \left(\widetilde{g}_{11}\widetilde{g}_{22} - \widetilde{g}_{12}\widetilde{g}_{12}\right)g^{2d} = \widetilde{g}g^{2d} \ , \tag{2.156}$$

where $\widetilde{g}$ is $\det\left(\widetilde{g}_{ab}\right)$. Putting in the explicit values for d, using Eqs.(2.108) and (2.131) in Eq.(2.155) we obtain the *Gauss equations*

$$\left.\begin{aligned} \frac{\widetilde{g}}{g} &= -\frac{1}{F}\left[\left\{_1{}^1{}_1\right\}_{,2} - \left\{_1{}^1{}_2\right\}_{,1} + \left\{_1{}^c{}_1\right\}\left\{_c{}^1{}_2\right\} - \left\{_1{}^c{}_2\right\}\left\{_c{}^1{}_1\right\}\right] \\ \frac{\widetilde{g}}{g} &= -\frac{1}{E}\left[\left\{_1{}^2{}_1\right\}_{,2} - \left\{_1{}^2{}_2\right\}_{,1} + \left\{_1{}^c{}_1\right\}\left\{_c{}^2{}_2\right\} - \left\{_1{}^c{}_2\right\}\left\{_c{}^2{}_1\right\}\right] \end{aligned}\right\} \ . \tag{2.157}$$

The other set of Codazzi and Gauss equations, given by the other identity of Eq.(2.119), are

$$\widetilde{g}_{12,2} + \left\{_1{}^c{}_2\right\}\widetilde{g}_{2c} = \widetilde{g}_{22,1} + \left\{_2{}^c{}_2\right\}\widetilde{g}_{1c} \ , \tag{2.158}$$

$$\left.\begin{aligned} \frac{\widetilde{g}}{g} &= -\frac{1}{G}\left[\left\{_2{}^1{}_2\right\}_{,1} - \left\{_1{}^1{}_2\right\}_{,2} + \left\{_2{}^c{}_2\right\}\left\{_c{}^1{}_1\right\} - \left\{_1{}^c{}_2\right\}\left\{_c{}^1{}_2\right\}\right] \\ \frac{\widetilde{g}}{g} &= -\frac{1}{F}\left[\left\{_2{}^2{}_2\right\}_{,1} - \left\{_1{}^2{}_2\right\}_{,2} + \left\{_2{}^c{}_2\right\}\left\{_c{}^2{}_1\right\} - \left\{_1{}^c{}_2\right\}\left\{_c{}^2{}_2\right\}\right] \end{aligned}\right\} \ . \tag{2.159}$$

The Codazzi equations given in Eqs.(2.128) and (2.133) can be written as the single equation

$$\widetilde{g}_{ab,c} + \left\{_a \ ^d \ _b\right\} \widetilde{g}_{cd} - \widetilde{g}_{bc,a} - \left\{_b \ ^d \ _c\right\} \widetilde{g}_{ad} = 0 \ . \tag{2.160}$$

This is easily verified. Notice that the left side of Eq.(2.135) is skew-symmetric under interchange of a and c, since g_{ab} and $\left\{_a \ ^d \ _b\right\}$ are symmetric under interchange of a and b. Thus the equation is trivially satisfied when $a = c$. If $a \neq c$ there are two cases, $b = a$ or $b = c$. Without loss of generality we can take $a = 1$, $c = 2$. (With $a = 2$, $c = 1$ the left side would just be the negative of the value obtained with $a = 1$, $c = 2$) Now $b = a$ gives Eq.(2.128) and $b = c$ gives Eq.(2.133).

For a further discussion of Gauss' equations it is necessary to consider the behavior of g and $\widetilde{g}$ under coordinate transformations. From Eqs.(2.36), (2.57) and (2.113) we see that

$$g_{\hat{a}\hat{b}} = \delta^a_{\hat{a}} \delta^b_{\hat{b}} g_{ab} \ , \tag{2.161}$$

whence

$$\hat{g} := \det\left(g_{\hat{a}\hat{b}}\right) = \Delta^2 g \ , \tag{2.162}$$

as the determinant of the product of matrices equals the product of the determinants. Also from Eqs.(2.57), (2.66) and (2.114) we obtain

$$\begin{aligned} \widetilde{g}_{\hat{a}\hat{b}} &= -\mathbf{e}_{\hat{a}\hat{b}} \cdot \widehat{\mathbf{N}} = \delta^a_{\hat{a}} \delta^b_{\hat{b}} \mathbf{e}_{ab} \cdot \mathbf{N} \\ &= \delta^a_{\hat{a}} \delta^b_{\hat{b}} \widetilde{g}_{ab} \ , \end{aligned} \tag{2.163}$$

which yields

$$\widehat{\widetilde{g}} := \det\left(\widetilde{g}_{\hat{a}\hat{b}}\right) = \Delta^2 \widetilde{g} \ . \tag{2.164}$$

Eqs.(2.162) and (2.130) show that the ratio $\widetilde{g}/g$ is invariant under coordinate transformations.

It was shown in section 3 that if the basis vectors are chosen along the principal directions the coordinate system is orthogonal, i.e. $F = 0 = \widetilde{F}$. From Eqs.(2.75) we see that in this case

$$\kappa_+ \kappa_- = \widetilde{E}\widetilde{G}/EG \ . \tag{2.165}$$

Also, in this case

$$\widetilde{g}/g = \widetilde{E}\widetilde{G}/EG \ . \tag{2.166}$$

Since the ratio given by the left side of Eq.(2.141) is invariant under coordinate transformations so is the product on the left side of Eq.(2.140). Thus the *Gaussian curvature*

$$K := \kappa_+ \kappa_- = \widetilde{g}/g \ , \tag{2.167}$$

is an invariant quantity. This invariance could have been obtained directly by inserting the general values for $\kappa_\pm$. However, the procedure adopted here is simpler and gives a flavor of the methods of deriving results using Tensors and Geometry.

Eqs.(2.132), (2.134)-(2.135) are collectively known as the *Gauss-Codazzi equations*. The Gauss equations are of special interest. Whereas Eq.(2.142) defines the Gaussian curvature in terms of the geometry *extrinsic* to the surface (given by the second fundamental form) as well as the geometry *intrinsic* to the

surface (given by the first fundamental form), Eqs.(2.132) and (2.134) give it *purely* in terms of the intrinsic geometry. In other words K can be computed without reference to the normal to the surface in terms of the two dimensional basis and the corresponding metric coefficients and their first and second partial derivatives. This fact is known as *Gauss' theorem*. It is the keystone of the geometry of curved spaces developed by Riemann.

Since the principal curvatures are separately invariant under coordinate transformations, not only their product but their average is also an invariant quantity. It is called the *mean curvature*

$$M = \kappa_+ + \kappa_- = \frac{\widetilde{E}G + E\widetilde{G} - 2F\widetilde{F}}{EG - F^2} \ . \tag{2.168}$$

It can be more easily evaluated by taking the scalar product of Eq.(2.118) with $\mathbf{e}^a$ (the dual basis vector) to give

$$\begin{aligned} M &= \mathbf{N}_a \cdot \mathbf{e}^a = \widetilde{g}_{ac} g^{bc} \mathbf{e}_b \cdot \mathbf{e}^a \\ &= \widetilde{g}_{ac} g^{bc} \delta^a_b = \widetilde{g}_{ac} g^{ac} \ . \end{aligned} \tag{2.169}$$

It is also called the *mean extrinsic curvature* and is useful when we consider a curved space embedded in a higher dimensional space.

The Gaussian curvature will be discussed in much greater detail in the next chapter and, therefore, specific examples will not be presented here. However, the reader is encouraged to work through the Gauss-Weingarten and Gauss-Codazzi equations for all the previous examples.

2.9 Exercises

1. It often happens that there is no analytic expression for the arc length parameter, s, in terms of the parameter u, in use. One, then has to go back to first principles and obtain all the geometric quantities in terms of u. Even if there is an analytic expression, $s(u)$, it may not be possible to invert it to obtain $u(s)$, so as to write $\mathbf{x}(u)$ as $\mathbf{x}(s)$. Further, one may choose a bad parameterization of the curve. Bearing this in mind, obtain the Frenet frame and the curvature and torsion for the following curves:
(a) a parabolic helix (so that it is constrained by $z = a^2(x^2 + y^2)$);
(b) a hyperbolic (of 2-sheets) helix constrained by $z^2 - (x^2 + y^2) = a^2$;
(c) a hyperbolic (of 1-sheet) helix constrained by $(x^2 + y^2) - z^2 = a^2$;
(d) a spheroidal helix $z^2/a^2 + (x^2 + y^2)/b^2 = 1$.
Hint: You can use the formula given in Eq.(2.26), and the Frenet-Serret formulae, to obtain the torsion, which turns out to be $\tau = (\mathbf{x}'.\mathbf{x}'' \times \mathbf{x}''')/|\mathbf{x}' \times \mathbf{x}''|^2$.

2. Construct the basis and hence obtain the first and second fundamental forms, the principal curvatures, the Gaussian and the mean curvatures of all surfaces mentioned above, and give the type of the surface. Further, if there exist asymptotic directions, determine them. Is there a globally parabolic space among them? If not, is there a locally parabolic space among them? If so, which one? If there is, determine the point(s) at which the parabolic nature appears.

3. For the spiral ramp $\mathbf{x}(u, v) = (av\cos u, av\sin u, bu) \qquad (1 < v < 2)$:
(a) Obtain all the quantities asked for above and determine its nature;
(b) Now think of two handrails of height c on the inner side and outer side of the ramp. Take the radius of the handrails to be δ and write $\mathbf{x}(u, v)$ for each of them. Obtain the difference of area of the two handrails in one full turn of the ramp. If the density of the steel used for the handrails is ρ, determine the shearing stress on the ramp on account of the handrails.

4. For each of the above, use the first fundamental form to derive the Riemann and Ricci tensor components and the Ricci scalar. Notice that while there is only one independent component of the Riemann tensor, there are different values of its components. List all those that are different by more than just a sign and are non-zero. Further, there are three different values of the Ricci tensor components, even though there can only be one independent component. Determine all of them. Verify that the Gaussian curvature is half the Ricci scalar for each of them.

Chapter 3

Tensors and Differential Geometry

The language of "tensors" is required to express GR. We have already developed the index notation of tensors to a fair extent and dealt with 2-dimensional tensors in a curved space without formally defining them. We need to generalize the concepts developed there to four dimensions for GR and Cosmology and to higher dimensions for understanding recent developments in attempts to unify gravity with other forces. As such we will deal with arbitrarily many dimensions here. Before proceeding to that discussion we will deal (in section1) with curves in flat n-dimensional space.

In GR, and especially in Cosmology, regarding spaces as embedded in higher dimensional flat spaces causes problems. (In Cosmology one is stuck with questions like: "what lies *outside* the Universe?", or "what was there *before* the Universe began?" Regarding spaces as curved in themselves obviates these problems.) As the number of coordinates equals the number of dimensions of the space, Cartesian coordinates will not apply in curved spaces, e.g. as the surface of a ball is 2-d we must describe it by only 2 coordinates, while using Cartesian coordinates would need 3. A more formal, abstract, definition of these spaces is needed. These abstract spaces are called *manifolds*. After defining the curved space generalization of vectors Penrose's "abstract index notation" is introduced, which helps to develop Tensor calculus. This will be used to provide a coordinate independent description of curvature and the curved space analogue of the straight line. Next, the relation between the geometrical and algebraic descriptions of symmetry is discussed. Finally, an introduction to the language of forms, tensor densities and the Weyl curvature tensor, which gives the gravitational fields where matter is not present is given.

3.1 Space Curves in Flat n-Dimensional Space

To be able to discuss curves in an n-dimensional space we need to set up some terminology. The position vector, $\mathbf{x}$ of point P, can be expressed in terms of its Cartesian components, x^i $(i = 1, \ldots, n)$, taking the basis vectors as implicit (since they are constant over the entire flat space). In a curved space this identification is not possible. It is this identification that leads to ambiguity in

the usual tensor notation and is avoided by Penrose's notation.

For a *straight line* $\mathbf{x}$ will be a linear function of a single variable

$$\mathbf{x}(u) = \mathbf{A}u + \mathbf{B} \text{ or } x^i(u) = A^i u + B^i \ . \tag{3.1}$$

For $n = 2$ this is the only flat sub-space that can be formed. For $n > 2$ a *plane* is given by a linear function of 2 variables

$$\mathbf{x}(u, v) = \mathbf{A}u + \mathbf{B}v + \mathbf{C} \text{ or } x^i(u, v) = A^i u + B^i v + C^i \ . \tag{3.2}$$

For $n = 3$ these are the only flat sub-spaces that can be formed. However, for $n > 3$ we could have linear functions of 3 (or more) parameters giving higher dimensional flat sub-spaces. These are called *hyperplanes*. In general an m-dimensional hyperplane ($m < n$) is given by

$$x^i(u^a) = A^i_a u^a + B^i, \ (a = 1, \ldots, n) \ . \tag{3.3}$$

Generally, if the dimensionality of the hyperplanes is not given it is assumed that $m = n - 1$.

A *space curve* is an arbitrary vector valued function, $\mathbf{x}$, of a single variable, say u, written as $\mathbf{x}(u)$ or $x^i(u)$. An m-dimensional sub-space ($m < n$) is given by $\mathbf{x}(u^a)$ or $x^i(u^a)$ $(a = 1, \ldots, n)$. If $m = n - 1$ this sub-space will be called a *hypersurface*. If $m = 2$ it will be called a *surface*. At present we will only deal with space curves.

We start by extending the three dimensional theory to four dimensions, generalizing the moving triad to a *moving tetrad*. This will describe the geometry from the point of view of an observer moving along the curve in a truly relativistic spirit. (It is of importance in Relativity infact.) We will then go on to derive the generalized Frenet-Serret formulae.

Consider two points on the curve, P and Q. As Q approaches P the two points define a tangent, which may be regarded as "an osculating line". This is a 1-dimensional space spanned by a single orthonormal basis vector,

$$\mathbf{t}(s) = \dot{\mathbf{x}}(s) = \frac{du}{ds}\mathbf{x}'(u) = \frac{\mathbf{x}'(u)}{ds/du} = \frac{\mathbf{x}'(u)}{|\mathbf{x}'(u)|}, \tag{3.4}$$

where s is the *affine* parameter, or the arc-length. We will write it as $\mathbf{e}_1(s)$, the first basis vector. If $n = 2$ there is a *unique* (up to sign) vector orthogonal to it, which we call $\mathbf{e}_2(s)$. The sign can be determined by requiring a counter-clockwise orientation. Since $\mathbf{e}_1(s)$ is normalized $\dot{\mathbf{e}}_1(s) \perp \mathbf{e}_1(s)$, as shown in the last chapter. Hence

$$\dot{\mathbf{e}}_1(s) \propto \mathbf{e}_2(s) \text{ or } \dot{\mathbf{e}}_1(s) = \kappa_1(s)\,\mathbf{e}_2(s) \ . \tag{3.5}$$

For $n = 3$ we had construct an osculating plane. Whereas for $n = 2$, $\dot{\mathbf{e}}_2(s) = -\kappa\mathbf{e}_1(s)$, this is not the case for $n = 3$ as there is a component of $\dot{\mathbf{e}}_2(s)$ which lies outside the osculating plane spanned by $\mathbf{e}_1$ and $\mathbf{e}_2$. This plane is defined by three points, P, Q and R, as Q and R approach P. The *third* vector is defined as being orthogonal to the other two by $\mathbf{e}_3 = \mathbf{e}_1 \wedge \mathbf{e}_2$.

For $n = 4$ the second basis vector is again defined through the osculating plane given by the three points on the curve. As in chapter 2, section 2, (henceforth written as section 2.2)

$$\dot{\mathbf{e}}_2(s) = -\kappa_1(s)\,\mathbf{e}_1(s) + \kappa_2(s)\,\mathbf{e}_2(s) \ . \tag{3.6}$$

However, now $\mathbf{e}_3$ is not given by the cross product of $\mathbf{e}_1(s)$ and $\mathbf{e}_2(s)$ as there are infinitely many vectors $\perp$ to $\mathbf{e}_1(s)$ and $\mathbf{e}_2(s)$. Therefore we construct the osculating hyperplane given by four points P, Q, R and S (provided that they are not co-planar) as Q, R and S approach P. Then $\mathbf{e}_3(s)$ is defined by $\mathbf{e}_1(s) \wedge \mathbf{e}_2(s)$ which is now spanned by the orthonormal basis $(\mathbf{e}_1(s), \mathbf{e}_2(s), \mathbf{e}_3(s))$. Ofcourse, now $\mathbf{e}_4$ is defined to be $\perp$ $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$ so that $(\mathbf{e}_1(s), \mathbf{e}_2(s), \mathbf{e}_3(s), \mathbf{e}_4(s))$ from the complete orthonormal basis for the 4-dimensional space, called the moving tetrad.

Now consider the normality condition of $\mathbf{e}_3(s)$. As before differentiating it gives the requirement that $\dot{\mathbf{e}}_3(s) \perp \mathbf{e}_3(s)$. Similarly, we can differentiate the orthogonality conditions for $\mathbf{e}_3(s)$ with $\mathbf{e}_1(s)$ and $\mathbf{e}_2(s)$ to obtain

$$\begin{aligned} \dot{\mathbf{e}}_1(s) \cdot \mathbf{e}_3(s) + \mathbf{e}_1(s) \cdot \dot{\mathbf{e}}_3(s) = 0, \\ \dot{\mathbf{e}}_2(s) \cdot \mathbf{e}_3(s) + \mathbf{e}_2(s) \cdot \dot{\mathbf{e}}_3(s) = 0 \ . \end{aligned} \tag{3.7}$$

Using Eq.(3.5) with the forms of Eqs.(3.7) gives $\dot{\mathbf{e}}_3(s) \perp \mathbf{e}_1(s)$. Using Eq.(3.6) gives the coefficient of $\mathbf{e}_2(s)$ in the expression of $\dot{\mathbf{e}}_3(s)$. Writing $\dot{\mathbf{e}}_3(s)$ as an expansion in terms of the moving tetrad, bearing in mind that the coefficients of $\mathbf{e}_1(s)$ and $\mathbf{e}_3(s)$ are zero, we have

$$\dot{\mathbf{e}}_3(s) = -\kappa_2(s)\,\mathbf{e}_2(s) + \kappa_3(s)\,\mathbf{e}_4(s) \ , \tag{3.8}$$

where $\kappa_3(s)$ is called the *third curvature*. The normality condition for $\mathbf{e}_4(s)$ gives $\dot{\mathbf{e}}_4(s) \perp \mathbf{e}_4(s)$. Differentiating the orthogonality condition gives

$$\begin{aligned} \dot{\mathbf{e}}_1(s) \cdot \mathbf{e}_4(s) + \mathbf{e}_1(s) \cdot \dot{\mathbf{e}}_4(s) = 0, \\ \dot{\mathbf{e}}_2(s) \cdot \mathbf{e}_4(s) + \mathbf{e}_2(s) \cdot \dot{\mathbf{e}}_4(s) = 0, \\ \dot{\mathbf{e}}_3(s) \cdot \mathbf{e}_4(s) + \mathbf{e}_3(s) \cdot \dot{\mathbf{e}}_4(s) = 0 \ . \end{aligned} \tag{3.9}$$

Eqs.(3.5) and (3.6) with the first two of the above equations give $\dot{\mathbf{e}}_4(s) \perp \mathbf{e}_1(s)$, $\mathbf{e}_2(s)$. Eq.(3.8) with the last of the above equations gives the coefficient of $\mathbf{e}_3(s)$ in the expansion of $\dot{\mathbf{e}}_4(s)$. Thus

$$\dot{\mathbf{e}}_4(s) = -\kappa_3(s)\,\mathbf{e}_3(s) \ . \tag{3.10}$$

Eqs.(3.5), (3.6), (3.8) and (3.10) are the 4-dimensional Frenet-Serret formulae.

The procedure for the generalizing to n-dimensional spaces should now be clear. Two points (in the limit) give the osculating line spanned by $\mathbf{e}_1(s)$. Three points give the osculating plane spanned by $(\mathbf{e}_1(s), \mathbf{e}_2(s))$, four points the osculating 3-hyperplane spanned by $(\mathbf{e}_1(s), \mathbf{e}_2(s), \mathbf{e}_3(s))$ and so on till n points give the osculating $(n-1)$ hyperplane which uniquely defines $\mathbf{e}_{n-1}(s)$. Finally $\mathbf{e}_n(s)$ is taken to be perpendicular to this hyperplane. The entire *moving enad* $(\mathbf{e}_1(s), \ldots, \mathbf{e}_n(s))$ is an orthonormal basis for the n-dimensional space. As before, the derivative of each basis vector has components along the one before it and the one after, i.e.

$$\dot{\mathbf{e}}_m(s) = -\kappa_{m-1}(s)\,\mathbf{e}_{m-1}(s) + \kappa_{m+1}(s)\,\mathbf{e}_{m+1}(s) \ , \tag{3.11}$$

except for first and last basis vectors which have only one basis vector, either

after or before, each. The generalized Frenet-Serret formulae are, then

$$\begin{pmatrix} \dot{\mathbf{e}}_1(s) \\ \dot{\mathbf{e}}_2(s) \\ \dot{\mathbf{e}}_3(s) \\ \vdots \\ \dot{\mathbf{e}}_{n-1}(s) \\ \dot{\mathbf{e}}_n(s) \end{pmatrix} = \begin{pmatrix} 0 & -\kappa_1(s) & 0 & \dots & 0 & 0 \\ -\kappa_1(s) & 0 & \kappa_2(s) & & 0 & 0 \\ 0 & -\kappa_2(s) & 0 & \kappa_3(s)\cdot\cdot & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & -\kappa_{n-2}(s) & 0 & \kappa_{n-1}(s) \\ 0 & 0 & 0 & \dots & -\kappa_{n-1}(s) & 0 \end{pmatrix} \begin{pmatrix} \mathbf{e}_1(s) \\ \mathbf{e}_2(s) \\ \mathbf{e}_3(s) \\ \vdots \\ \mathbf{e}_{n-1}(s) \\ \mathbf{e}_n(s) \end{pmatrix} \tag{3.12}$$

The reason for calling the "torsion", defined in (section 2.2) the "second curvature" should also be clear now. In the general case there will be $(n-1)$ curvatures. Numbering them is obviously much more convenient than finding new names for each. Notice that if $\kappa_m(s) = 0$ for some $m < n$ then $\kappa_p(s)$ cannot be defined for $p > m$ and the osculating m-dimensional hyperplane will be the last that can be defined. The reason is that more than $m+1$ points will now always lie in the same m-dimensional hyperplane. Thus the entire curve will always lie in it.

3.2 Manifolds

When dealing with curved spaces it is necessary to provide a more rigorous, formal definition of the space. To explain why, let me start with a riddle. "A man on a bear-hunt sees a bear sleeping 1 km due East, but the shot would not be good from that angle, so he runs 1 km due North, points due South and shoots the bear. What colour was the bear?" The answer is "White", since the shot could only come from the North pole, where every direction is South and the intuition of plane geometry ceases to be valid. Why do we need to go beyond Gauss' geometry? Because for our purposes we need to deal with n-dimensional bears, in curved spaces that have no orthogonal direction to them.

The problems that arise are best explained in terms of the simplest possible example of a curved space, namely a circle. Though it is a 1-dimensional space it is not possible to provide one and only one coordinate to each point consistently in such a way that calculus can be used. Assigning coordinates means that a one-to-one mapping be constructed from a circle, $\mathbb{S}^1$, into the set of real numbers, $\mathbb{R}$. An obvious procedure would be to assign the angle in radians, θ, starting from some arbitrary origin on the circle, S, to each point P. This θ lies between 0 and 2π. The problem arises when we ask whether the end points of the range of θ, 0 and 2π , lie inside or outside the allowed range. If they are excluded, i.e. $\theta \in (0, 2\pi)$, then S is excluded from the circle. If they are included, i.e. $\theta \in [0, 2\pi]$, there are two numbers assigned to S, 0 and 2π, which is not allowed. We could use the semi-open interval in which one point is excluded, $(0, 2\pi]$ or $[0, 2\pi)$, which would provided a one-to-one mapping. However continuity would be lost as $\theta \to 2\pi$, $P \to S$. In the limit $\theta = 2\pi$, $P = S$ but $2\pi \neq 0$. Thus the image of $\mathbb{S}^1$ in $\mathbb{R}$ is discontinuous at the point S while $\mathbb{S}^1$ is not.

The entire real line, $\mathbb{R}$, cannot provide a satisfactory coordinate system for $\mathbb{S}^1$ either. Consider the geometrical mapping from $\mathbb{S}^1$ to $\mathbb{R}$ shown in Fig. 3.1. Clearly $S^1 \setminus \{N\}$ has a one-to-one correspondence with $\mathbb{R}$ and N has two images, $-\infty$ and $+\infty$. Deforming $\mathbb{R}$, or equivalently the mapping, does not alter the above conclusion. Again it is not possible to find a one-to-one continuous mapping from $\mathbb{S}^1$ to $\mathbb{R}$ The same statement holds true for mapping a sphere, $\mathbb{S}^2$, into $\mathbb{R}^2$ $(= \mathbb{R} \times \mathbb{R})$. The usual procedure for dealing with a circle is to embed

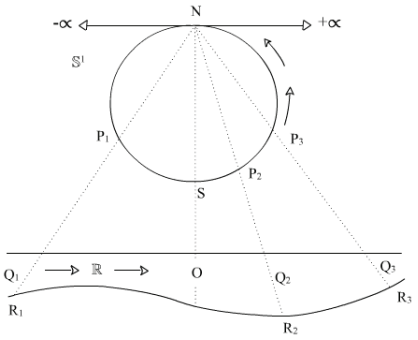


Figure 3.1: Geometrically mapping the unit circle, $\mathbb{S}^1$, onto the real line $\mathbb{R}$. Join N to any point $P_i \in \mathbb{S}^1$ to get a point $Q_i \in \mathbb{R}$. As $P_i \to N$ from the right $Q_i \to \infty$. As $P_i \to N$ from the left $Q_i \to -\infty$. Thus, in the limit, N corresponds to both $+\infty$ and $-\infty$. Distorting the real line continuously so that there remains only one image for each $P_i \in \mathbb{S}^1$, does not alter the conclusion. All that is done is to replace Q by R.

it in a (flat) 2 -dimensional Euclidean space and use Cartesian coordinates. This is an *extrinsic* description. Here we will develop the *intrinsic* description that will be used in Relativity and Cosmology.

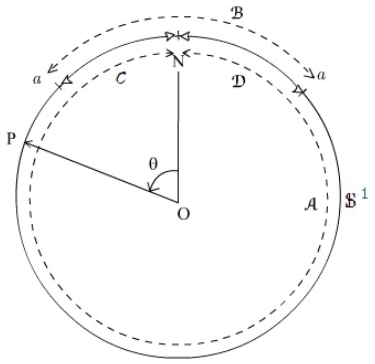


Figure 3.2: The unit circle $\mathbb{S}^1$ covered by two open sets $\mathcal{A} = \mathbb{S}^1 \setminus \{N\}$ and $\mathcal{B} = (-a, a)$. $\mathcal{A} \cap \mathcal{B} = \mathcal{C} \cup \mathcal{D}$ where $\mathcal{B} \setminus (\mathcal{C} \cup \mathcal{D}) = \{N\}$. In $\mathcal{A}$ the coordinates of a point $P \in \mathcal{A}$ is given by the angle θ (in radians) while in $\mathcal{B}$ it is given by the distance, negative to the left or positive to the right, from N.

The procedure adopted here is to cover the circle with two coordinate systems which overlap, see Fig. 3.2. In both regions continuity is maintained. In the overlapping regions one coordinate description can be translated into the other, i.e. there exist coordinate transformations between the coordinate systems. Take the two regions to be the open sets $\mathcal{A} = \mathbb{S}^1 \setminus \{N\}$ and $\mathcal{B}$, a set containing N with the end points (removed) equidistant from N. The reason open sets are required is so as to be able to use calculus at all points in the space. If a set were closed, there would be points at which limits could not be taken from both

sides. Define coordinates on $\mathcal{A}$ by

$$f : \mathcal{A} \to (0, 2\pi) \text{ , given by } f(P) = \theta,\ P \in \mathcal{A},\ \theta \in (0, 2\pi)\ . \tag{3.13}$$

The coordinates for $\mathcal{B}$ are defined by

$$\begin{aligned} &g : \mathcal{B} \to (-a, a) \text{ , given by } g(P) = \phi,\ P \in \mathcal{B},\ \phi \in (-a, a)\ , \\ &g(N) = 0\ . \end{aligned} \tag{3.14}$$

There are two regions of overlap, $\mathcal{C}$ and $\mathcal{D}$, as shown in Fig. 3.2,

$$\mathcal{A} \cap \mathcal{B} = \mathcal{C} \cup \mathcal{D},\ \mathcal{A} \cup \mathcal{B} = \mathbb{S}^1,\ \mathcal{C} \cap \mathcal{D} = \phi\ . \tag{3.15}$$

In $\mathcal{B}$, $f(P) = g(P)$, if it is a unit circle (otherwise there would be a re-scaling by the radius of the circle). In $\mathcal{D}$, $f(P) = 2\pi + g(P)$. These are the required coordinate transformations.

The spaces we will deal with are essentially generalizations of curved 2-d surfaces on which calculus is applicable. As such the space must be a continuum everywhere and not have too many points anywhere. Also the space must be connected everywhere and there must be unique limit points. Further, it must be possible to provide the same number of coordinates for each region of the space, as in the above example. Spaces satisfying these requirements are called *manifolds*. Instead of first defining all the terms to be used and then proceeding to the definition of the manifold, I will reverse the order so as to maintain motivation for all the Topology needed.

A manifold of dimension n is a separable, connected, Hausdorff space with a homeomorphism from each element of its open cover into $\mathbb{R}^n$.

A space is said to be *separable* if it has a countable dense subset. Countability was explained in section 1.5**.** A subset is said to be *dense* if its closure is the original set. The *closure* of a set is the smallest closed set containing it, e.g. $[0, 1]$ is the closure of $(0, 1)$. This statement is written as $\overline{[0, 1]} = \overline{(0, 1)}$. The set of rational numbers, $\mathbb{Q}$, is a dense subset of $\mathbb{R}$, i.e. $\overline{\mathbb{Q}} = \mathbb{R}$. Since $\mathbb{Q}$ is countable, $\mathbb{R}$ is separable but $\mathbb{Q}$ is not. This condition ensures that the space is at least a continuum and there are no accumulation points.

Not surprisingly, a space is said to be *connected* if it is not disconnected. It is said to be *disconnected* if there exist two sets A and B, whose union is the whole space but which are disjoint, i.e. $A \cap B = \emptyset$, such that the closure of either is disjoint with the other, i.e. $\overline{A} \cap B = A \cap \overline{B} = \emptyset$. It is *not* necessary that $\overline{A} \cap \overline{B} = \emptyset$. The set of rationals is disconnected as we can define

$$A = \left\{ q |\, q \in \mathbb{Q}\,, q < \sqrt{2} \right\}, B = \left\{ q |\, q \in \mathbb{Q}\,, q > \sqrt{2} \right\} \tag{3.16}$$

Now $\sqrt{2} \notin \mathbb{Q}$. Thus $A \cup B = \mathbb{Q}$. The closures of these sets are

$$\overline{A} = \left\{ r |\, r \in \mathbb{R}\,, r \leq \sqrt{2} \right\}, \overline{B} = \left\{ r |\, r \in \mathbb{R}\,, r \geq \sqrt{2} \right\} \tag{3.17}$$

Thus $\overline{A} \cap B = A \cap \overline{B} = \phi$. Even though $\overline{A} \cap \overline{B} = \left\{\sqrt{2}\right\}$, we see that $\mathbb{Q}$ is disconnected. Again, it is clear that $\mathbb{R}$ is *connected.*

A space is said to be *Hausdorff* if two distinct points possess disjoint neighborhoods.. A *neighborhood* of a point is a set containing an open set containing the point. This applies for example to $\mathbb{R}$, $\mathbb{R}^2$, etc. or $\mathbb{S}^1$, $\mathbb{S}^2$, etc. In fact

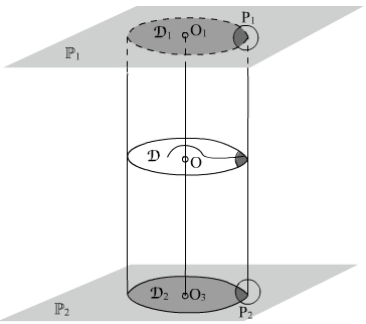


Figure 3.3: An example of a non-Hausdorff space. Two open unit discs $\mathfrak{D}_1$ and $\mathfrak{D}_2$ are cut out of two planes $\mathbb{P}_1$ and $\mathbb{P}_2$ and replaced by a single open disc $\mathfrak{D}$. Two points on the edges of the two discs cut out, $P_1 \in \mathbb{P}_1$ and $P_2 \in \mathbb{P}_2$ are considered, each given by $(1,0)$ in the corresponding space. The neighbourhoods of both points intersect in the disc $\mathfrak{D}$. For a curve in $\mathfrak{D}$ which approaches $(1,0)$ there will be two limit points P_1 and P_2 neither in $\mathfrak{D}$. As such it would not be possible to use calculus in such a space.

it is difficult to visualize a space which is not Hausdorff. To fully grasp what Hausdorff implies it is necessary to see an example of a non-Hausdorff space. Consider two planes, see Fig. 3.3,

$$\mathbb{P}_1 = \{\,(x_1,y_1)|\,x_1,y_1 \in \mathbb{R}\,\}\ ,\ \mathbb{P}_2 = \{\,(x_2,y_2)|\,x_2,y_2 \in \mathbb{R}\,\}\ , \tag{3.18}$$

with their respective open discs

$$\left.\begin{array}{l}\mathfrak{D}_1 = \left\{\,(x_1,y_1)|\,x_1,y_1 \in \mathbb{R}\,,\ x_1^2+y_1^2<1\right\}\ ,\\ \mathfrak{D}_2 = \left\{\,(x_2,y_2)|\,x_2,y_2 \in \mathbb{R}\,,\ x_2^2+y_2^2<1\right\}\ ,\end{array}\right\} \tag{3.19}$$

replaced by a single open disc

$$\mathfrak{D} = \left\{\,(x,y)|\,x,y \in\ R,\ x^2+y^2<1\right\}\ . \tag{3.20}$$

This space is $(\mathbb{P}_1 \setminus \mathfrak{D}_1) \cup \mathfrak{D} \cup (\mathbb{P}_2 \setminus \mathfrak{D}_2)$. Consider the two points $P_1 = (1,0) \in \mathbb{P}_1$, $P_2 = (1,0) \in \mathbb{P}_2$. Now $P_1 \neq P_2$, but every open set containing P_1 intersects every open set containing P_2, the intersection lying in the open disc $\mathfrak{D}$. Thus, moving along a curve in $\mathfrak{D}$ that goes infinitesimally close to $(1,0)$ there will be two distinct limit points. Hence calculus could not be used on this space. This is why we require the space to be Hausdorff.

A *homeomorphism* is a one-to-one invertible, continuous, mapping. It will be used to assign coordinates to points on the manifold. If the mappings are differentiable they are called *diffeomorphisms*. If the mappings are diffeomorphisms the manifold is said to be differentiable. If they are n times differentiable it is said to be $\mathcal{C}^n$. (Thus a continuous manifold is $\mathcal{C}^0$.) If the homeomorphisms are infinitely differentiable it is $\mathcal{C}^\infty$ and if they are analytic it is called $\mathcal{C}^p$. A function is said to be analytic if Taylor's theorem applies to it. An example of a function that is infinitely differentiable but non-analytic is

$$\left.\begin{array}{ll}f(x) = \exp\left[-a^2/\left(a^2-x^2\right)\right] & |x|<a\\ = 0 & |x| \geq a\end{array}\right\}\ , \tag{3.21}$$

which is sketched in Fig. 3.4. Clearly, Taylor's theorem is not applicable everywhere. Consider a point $x = -a - \epsilon$. Thus $f(x)$, $f'(x)$, ... are all zero. However, $f(-a + \epsilon)$ is non-zero. At $x = \pm a$, $f(x) = 0$. The differentiation introduces a singular factor into the derivative, but the exponential factor goes faster to zero at this point than the extra factor goes to zero. Thus $f'(x) = 0$, and similarly the higher derivatives, there as well, and hence $f(x)$ is infinitely differentiable there, and at all other points, so it is infinitely differentiable.

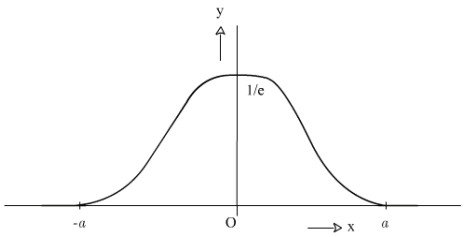


Figure 3.4: A curve that is continuous and in fact infinitely differentiable, without being analytic. It is zero for $|x| > a$ and joins the x-axis perfectly smoothly at $x = \pm a$. Its value at $x = 0$ is $1/e$.

An *open cover* is a set of open sets whose union is the whole space. It is non-unique. In our example of $\mathbb{S}^1$, $\{\mathcal{A}, \mathcal{B}\}$ is an open cover. $\{\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}\}$ would be another open cover. It is required that there be well defined coordinate transformations from one coordinate system to another in the intersection of the two regions. The manifold is said to be compact if there exists a finite open cover, i.e. the open cover has a finite number of elements. It is said to be *para-compact*, or *locally compact*, if the open cover is countable or has a locally finite refinement.

Let the open cover of the manifold, $\mathcal{M}_n$, be denoted by $\{\mathcal{U}_i\}_{i\in \mathrm{I}}$, where I is some index set. Then $\bigcup_{i\in \mathrm{I}} \mathcal{U}_i = \mathcal{M}_n$ and for every $\mathcal{U}_i$ there exists some $\mathcal{U}_j$ such that $\mathcal{U}_i \cap \mathcal{U}_j \neq \emptyset$. Here $\mathcal{U}_i$ is called a *coordinate patch*, $f_i : \mathcal{U}_i \to R^n$ a *coordinatization*, $(\mathcal{U}_i, f_i)$ a *coordinate chart* and $\{(\mathcal{U}_i, f_i)\}_{i\in \mathrm{I}}$ an *atlas*. R^n is called the *coordinate system*. In the example of the circle, $\mathcal{M}_n = S^1$, $n = 1$, $\mathrm{I} = \{1, 2\}$, $\mathcal{U}_1 = \mathcal{A}$, $\mathcal{U}_2 = \mathcal{B}$, $f_1 = f$, $f_2 = g$. On $\mathcal{U}_1 \cap \mathcal{U}_2 = \mathcal{C} \cup \mathcal{D}$ there are two coordinate transformations. On $\mathcal{C}$, $f_1 = f_2$, and on $\mathcal{D}$, $f_1 = 2\pi + f_2$. The atlas is $\{(\mathcal{U}_1, f_1), (\mathcal{U}_2, f_2)\}$ or $\{(\mathcal{U}_i, f_i)\}_{i=1.2}$.

In general let $f_i : \mathcal{U}_i \to R^n$ so that $f_i(P) = x^a$ $(a = 1, ..., n)$ for all $P \in \mathcal{U}_i$ and $f_j : \mathcal{U}_j \to \hat{\mathbb{R}}^n$ given by $f_j(P) = \hat{x}^a$ for all $P \in \mathcal{U}_j$ (see Fig. 3.5). The coordinate transformation in $\mathcal{U}_i \cap \mathcal{U}_j$ will be given by $\left(f_j f_i^{-1}\right) : f_i(P) \to (P) \to f_j(P)$ and the inverse transformation by $\left(f_i f_j^{-1}\right) : f_j(P) \to f_i(P)$. In terms of the earlier notation the former corresponds to $\delta_a^{\hat{a}}$ and the latter to $\delta_{\hat{a}}^a$.

Notice that as developed so far coordinates have been presented as having only mathematical significance, to assign n numbers to a point. Many people take this to mean that there is no physical significance to the choice of coordinates. This is simply not true. In GR the change of coordinates can be due to a change of *point of view*; i.e. change of the position of the observer; change of a *frame of reference*, i.e. change of motion of the observer, or a change of *coordinate system*, i.e. the definition of how to assign numbers, such as by one distance and all others as angles, or all distances.

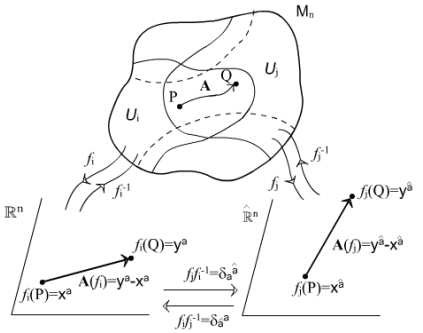


Figure 3.5: A manifold $\mathcal{M}_n$ is covered (here) by four open sets. Two (dotted ones) are not named and the other two (full lines) are called $\mathcal{U}_i$ and $\mathcal{U}_j$. Two points $P, Q \in \mathcal{U}_i \cap \mathcal{U}_j$. With each open set there is a homeomorphism to some $\mathbb{R}^n$. Here two are shown $f_i : \mathcal{M}_n \to \mathbb{R}^n$. Their inverse act in reverse. $\mathbf{A} : P \to Q$ on the manifold. The images of $\mathbf{A}$, $\mathbf{A}(f)$ in the coordinate systems, map the images of P to the images of Q. Thinking of these mappings as trains from the central station to different districts and back, from $\mathbb{R}^n$ one can "catch" an inverse homeomorphism, to the manifold, "change" at $\mathcal{U}_i \cap \mathcal{U}_j$ to a homeomorphism, f_j, going to $\widehat{\mathbb{R}}^n$. The composition $f_j f_i^{-1}$ takes $f_i(P)$ to $f_j(P)$. Generally the composition is the coordinate transformation $\delta^a_{\hat{a}}$, which corresponds to a direct train from one district to the other. The reverse procedure gives $f_i f_j^{-1}$ corresponding to $\delta^a_{\hat{a}}$.

3.3 Vectors in Curved Spaces

The concepts and properties of vectors in flat space do not generally carry over to curved spaces as may be seen by a simple example. Consider a person at the (Earth's) equator who goes due North, then due East, then due South and then due West, each time travelling $5,000$ km. Using the equality of parallel vectors of equal magnitude, that is the basis of vector analysis in flat space, we would conclude that the person ends up at the starting position. However, remember that the Earth has a circumference of about $40,000$ km and the curvature of the Earth changes this expectation. Though the first and the third laps both cover $\pi/4 = .25\pi$ radians, like the last lap of the journey covers, the second lap going East along the $45°$ N circle of latitude (see Fig. 3.6), has a circumference of $28,284$ km. The angle subtended by this arc is about 0.35π radians, leaving an angular gap of about 0.1π at the equator. This amounts to about $2,071$ km East of the starting point! Thus, vectors of equal magnitude that appear parallel cannot be naively regarded as equal. Only in the tangent space to the curved space would the usual flat space concepts hold. Generally the procedure for dealing with vectors in curved spaces will be to define everything in terms of the tangent space. Going back to the bear-hunt, the three vertices of the triangle involved, *each subtend a right angle*, and so the sum of the three angles of the triangles is not two, but three right angles. (This is the most thoroughly right-angled triangle there can be!)

The curved space analogue of a vector is called a *derivation*. (The significance of this name will become clear in section 5 of this chapter.) It is defined as a

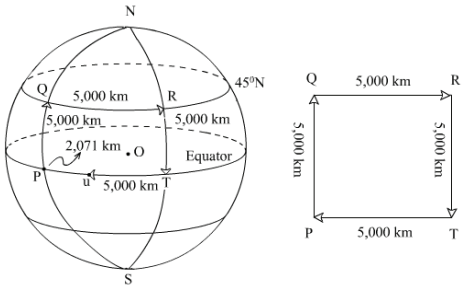


Figure 3.6: (a) Someone starts at point P on the equator and travels $5,000$ km due North to point Q on the circle of $45°$ N latitude; then due East along the circle of latitude for $5,000$ km to point R; then due South for $5,000$ km to point T on the Equator, and finally due West along the Equator to point U, which is $5,000$ km from T and $2,071$ km from P. (b) With flat space vectors we would have expected to reach P and not U.

mapping, $\mathbf{A}$, on the manifold that moves a point P to another point Q in the same coordinate patch, see Fig. 3.6. The image of $\mathbf{A}$ in the coordinate system will be written as $\mathbf{A}(f)$ instead of $f(\mathbf{A})$, for reasons that will become clear later. This image is simply the n components of the vector in the coordinate system, A^a. If $P, Q \in \mathcal{U}_i$, $f_i(P) = x^a$ and $f_i(Q) = y^a$, then $A^\alpha = y^a - x^a$ as usual. The action of $\mathbf{A}(f_i)$ on $f_i(P)$ to give $f_i(Q)$ is clearly by addition. Notice that if one of the coordinates is an angle the corresponding component of the vector is also an angle and not a distance as is clear from Fig.3.7 (see SR also). Thus we can not add components, or their squares, without multiplying them by appropriate factors to make the whole dimensionally invariant.

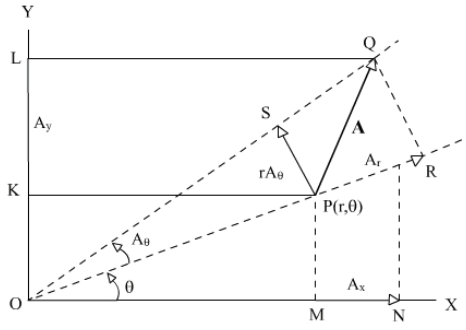


Figure 3.7: The Cartesian and polar components of a vector $\mathbf{A}$. Notice that the polar component, A^θ, is an angle while $\overrightarrow{PS}$ is rA^θ in the θ-direction at P. A^θ is the difference between the angle of Q and the angle of P at O, made with the X-axis.

Linear combinations of derivations can be defined by requiring that the image of the linear combination be the linear combination of the images under the coordinatization. Let $\mathbf{B} : P \to R$ be another derivation and $\lambda \in \mathbb{R}$. Then

$$\mathbf{C} = \mathbf{A} + \lambda\mathbf{B} \Leftrightarrow C^a = A^a + \lambda B^a \ . \tag{3.22}$$

A set, $\mathbf{V}$, is called a *vector space* over a field, $\mathbb{F}$ if it is an abelian group and linear combinations of elements of the set, using the elements of the field to give scalar multiplication, lie in the set. A field is defined in terms of a ring. A *ring* is a set with two binary operations defined on it, $(S, +, \cdot)$ such that it is an abelian group under " $+$ ", a semigroup under " $\cdot$ " and the distributive law holds: $a \cdot (b + c) = a \cdot b + a \cdot c$, $(a + b) \cdot c = a \cdot c + b \cdot c$; $\forall\ a, b, c \in S$. A ring such that on removing some elements from S we get a group under " $\cdot$ " is called a *field*. The set of all derivations, $\mathcal{D}$, at a point, P, forms a linear vector space over the field $\mathbb{R}$. This vector space corresponds to a tangent space at the point P.

A *dual derivation*, $\mathbf{X}$, is defined as a mapping from $\mathcal{D}$ to $\mathbb{R}$, and $\mathbf{X}(\mathbf{A})$ is written as $\mathbf{X} \cdot \mathbf{A}$ and identified with $\mathbf{A} \cdot \mathbf{X}$, for all $\mathbf{A} \in \mathcal{D}$. The image of $\mathbf{X}$ under f will be written as X_a. Thus $\mathbf{X}(\mathbf{A}) = \mathbf{X} \cdot \mathbf{A} = \mathbf{A} \cdot \mathbf{X} = X_a A^a = A^a X_a$, is the analogue of the usual "dot product" of vectors except that $\mathbf{A}$ and $\mathbf{X}$ belong to different spaces. The set of dual derivations, $\mathcal{D}^*$, forms a linear vector space over $\mathbb{R}$, the linear combinations being given by

$$\mathbf{Z} = \mathbf{X} + \lambda \mathbf{Y} \Leftrightarrow Z_a A^a = X_a A^a + \lambda Y_a A^a \ \forall\ \mathbf{A} \in \mathcal{D}\ . \tag{3.23}$$

We can define operators that convert derivations to dual derivations; derivations and dual derivations to other derivations and dual derivations; or dual derivations to derivations. Examples of these operators are $\mathbf{AB}$, $\mathbf{AX}$ (or $\mathbf{XA}$) and $\mathbf{XY}$. We can also construct linear combinations of these operators to give new linear vector spaces $\mathcal{D} \otimes \mathcal{D}$, $\mathcal{D} \otimes \mathcal{D}^* = \mathcal{D}^* \otimes \mathcal{D}$ and $\mathcal{D}^* \otimes \mathcal{D}^*$ respectively, where "$\otimes$" is the *tensor product*, which is a linearity preserving Cartesian product. These operators can be further generalized by taking k copies of $\mathcal{D}$ and l of $\mathcal{D}^*$ to define the linear vector space (at P) $\mathbf{V}_l^k = \mathcal{D} \otimes ... \otimes \mathcal{D} \otimes \mathcal{D}^* \otimes ... \mathcal{D}^*$.

An element of $\mathbf{V}_l^k$ is called a *tensor of valence* $[^k_l]$ and rank $(k + l)$, with the *dimension*, n, of the manifold. Let all the points $Q, R, ...$, obtained by applying derivations to P, belong to $\mathcal{U}_i$. Then the image of $\mathbf{T} \in \mathbf{V}_l^k$ under f_i is the set of images of $\mathbf{T}$ in the coordinate system,

$$\mathbf{T}(f_i(P)) = T^{a...c}{}_{d...f}\ (\mathbf{x})\ , \tag{3.24}$$

where $a....c$ are k indices, $d...f$ are l indices and $\mathbf{x}$ denotes that the components of the tensor are given in the x^a-coordinate system. In Cartesian coordinates, where applicable, the tensor can again be identified with its components as the basis vectors are constant over the entire space. This identification is not possible in general. If all the points lie in $\mathcal{U}_i \cap \mathcal{U}_j$ and $f_i(P) = x^{\hat{a}}$, the components can be transformed to the new coordinate system by

$$T^{\hat{a}...\hat{c}}{}_{\hat{d}...\hat{f}}\ (\hat{\mathbf{x}}) = \delta^{\hat{a}}_a ... \delta^{\hat{c}}_c \delta^d_{\hat{d}} ... \delta^f_{\hat{f}} T^{a...c}{}_{d...f}\ (\mathbf{x})\ . \tag{3.25}$$

The $\delta^{\hat{a}}_a$ and $\delta^a_{\hat{a}}$ are defined here as in the previous section. It will be seen in section 5 that they turn out to be the quantities given by generalizing Eqs.(2.58) and (2.59) in chapter 2 to n-dimensions.

By definition, tensors of the same valence can be added and linear combinations can be formed. Tensors of valence $[^k_l]$ and $[^p_q]$ can be multiplied to give a tensor of valence $[^{k+p}_{l+q}]$. One or more of the $(l + q)$ dual derivations can act on a corresponding number of the $(k + p)$ derivations. If a derivation and dual derivation "annihilate each other" to give a real number the process is called

contraction. It reduces the valence of the tensor by $\left[\begin{smallmatrix}1\\1\end{smallmatrix}\right]$ giving a tensor of valence $\left[\begin{smallmatrix}k-1\\l-1\end{smallmatrix}\right]$. In component form, contraction is given by applying the Kronecker delta. If $\mathbf{S}$ is an appropriately contracted form of $\mathbf{T}$,

$$S^{a\ldots\ldots c}{}_{d\ldots\ldots f} = \delta^e_b T^{a\ldots b\ldots c}{}_{d\ldots e\ldots f} = T^{a\ldots b\ldots c}{}_{d\ldots b\ldots f} \ . \tag{3.26}$$

The advantage of the index notation is clear from here. If we want to actually compute some expression involving tensors explicitly, we need to keep track of which indices are contracted. For complicated calculations it is not possible to do so without writing the components out explicitly. For example $A^aB^bX_bY_c \neq A^bB^aX_cY_b$. These would normally be distinguished by writing them as $\mathbf{A}\,(\mathbf{B}\cdot\mathbf{X})\,\mathbf{Y}$ and $(\mathbf{A}\cdot\mathbf{Y})\,\mathbf{B}\mathbf{X}$. However, if the order of writing the vectors had to be maintained (for example because $\mathbf{B}$ was an operator), there would already be problems with the notation. Further, if $\mathbf{A}\in\mathbf{V}^2_0$ and $\mathbf{B}\in\mathbf{V}^0_2$, $\mathbf{A}\cdot\mathbf{B}$ would be very ambiguous. Its components could be $A^{ab}B_{cb}$, $A^{ab}B_{ab}$, or $A^{ab}B_{ba}$. The first are components of a tensor of valence $\left[\begin{smallmatrix}1\\1\end{smallmatrix}\right]$ while the latter two quantities are scalars. Thus we seem to be forced to choose between the rigorous notation that is most inconvenient for calculations, or the convenient but non-rigorous component index notation.

Would it not be great to "have one's cake and eat it too"? In other words, could we not get a notation that is both rigorous and easy to use? The great thing with Mathematics is that, unlike physical systems, there are no limitations to what can be done. In Mathematics one can just define what one wants. For example, a big stumbling block in the development of algebra by the Indians and the Muslims was that no solution existed for an equation like $x^2+1=0$. However, Cardano simply *defined* the solutions to be $\pm\iota$. *Then* the solutions existed! You may feel that too much emphasis has been placed on "mere notation", but remember that the history of Mathematics is a history of the development of notation. The introduction of the notation of calculus opened up new vistas in Mathematical Analysis. The notation due to Penrose makes rigorous GR accessible to students at an earlier stage. It has been emphasized but not overemphasized. The next section presents Penrose's prescription for having our cake and eating it too.

3.4 Penrose's Abstract Index Notation

Abstract indices distinguish between different copies of the space of derivations or dual derivations by providing them with labels such that contraction can be denoted by repeated labels. Thus the pair $(\mathcal{D},a)$ will be written as $\mathcal{T}^{\underline{a}}$ and $\mathbf{A}\in\mathcal{D}$ as $A^{\underline{a}}\in\mathcal{T}^{\underline{a}}$. Other copies will be given different abstract index labels, $\mathcal{T}^{\underline{b}}$, $\mathcal{T}^{\underline{c}}$, ... Similarly $(\mathcal{D}^*,a)=\mathcal{T}_{\underline{a}}$, $(\mathcal{D}^*,b)=\mathcal{T}_{\underline{b}}$, ... and $\mathbf{X}\in\mathcal{D}^*$ becomes $X_{\underline{a}}\in\mathcal{T}_{\underline{a}}$. Further, for tensors of valence $\left[\begin{smallmatrix}2\\0\end{smallmatrix}\right]$, $\left[\begin{smallmatrix}1\\1\end{smallmatrix}\right]$ or $\left[\begin{smallmatrix}0\\2\end{smallmatrix}\right]$ we have $\mathcal{T}^{\underline{ab}}$, $\mathcal{T}^{\underline{a}}{}_{\underline{b}}$, $\mathcal{T}_{\underline{ab}}$. Generally, if $\mathbf{T}\in\mathbf{V}^k_l$ then $T^{\underline{a},\ldots,\underline{c}}{}_{\underline{d},\ldots,\underline{f}}\in\mathcal{T}^{\underline{a},\ldots,\underline{c}}{}_{\underline{d},\ldots,\underline{f}}$ where $\underline{a},\ldots,\underline{c}$ are k labels and $\underline{d},\ldots,\underline{f}$ are l labels. These labels do not take values $1,\ldots,n$. They may be thought of as labels on jam bottles which merely give information about contents of the bottle without dipping into it. To extend the analogy regard the derivations as jams and the dual derivatives as pickles. Notice that $\mathcal{T}^{\underline{a}}\neq\mathcal{T}^{\underline{b}}$ i.e. $A^{\underline{a}}\neq A^{\underline{b}}$ even though it is the same derivation. The isomorphism between the two copies of derivations is given by the *index substitution* operator, $\delta^{\underline{b}}_{\underline{a}}$,

$$\delta^{\underline{b}}_{\underline{a}}:\mathcal{T}^{\underline{a}}\to\mathcal{T}^{\underline{b}} \text{ given by } \delta^{\underline{b}}_{\underline{a}}A^{\underline{a}}=A^{\underline{b}} \ . \tag{3.27}$$

This operator is *not* the Kronecker delta but is an identity operator between copies of derivations. In terms of our analogy it is a label changing machine in the jam factory which, say, changes the name of the marketing company without changing the substance.

The $\delta^{\underline{b}}_{\underline{a}}$ may be used to contract a tensor. For example if $\mathbf{A} \in \mathbf{V}^1_1$

$$\delta^{\underline{b}}_{\underline{a}} A^{\underline{a}}_{\underline{b}} = B \in \mathcal{T} \ , \tag{3.28}$$

the set of scalar functions at P. The summation convention carries over to abstract indices, so that only two indices, one up and one down, are to be repeated. Clearly $\mathcal{T}$ forms a linear vector space. Let $A^{\underline{a}}_{\underline{b}} = C^{\underline{a}} X_{\underline{b}}$. Then

$$\mathrm{B} = \delta^{\underline{b}}_{\underline{a}} C^{\underline{a}} X_{\underline{b}} = C^{\underline{b}} X_{\underline{b}} = \mathbf{C} \cdot \mathbf{X} = C^{\underline{a}} X_{\underline{a}} \ . \tag{3.29}$$

Clearly, for consistency $\delta^{\underline{b}}_{\underline{a}}$ gives an isomorphism from $\mathcal{T}_{\underline{b}}$ to $\mathcal{T}_{\underline{a}}$.

$$\delta^{\underline{b}}_{\underline{a}} : \mathcal{T}_{\underline{b}} \to \mathcal{T}_{\underline{a}} \text{ given by } \delta^{\underline{b}}_{\underline{a}} X_{\underline{b}} = X_{\underline{a}} \ . \tag{3.30}$$

In our analogy, the jam label changing machine can be used in reverse as a pickle label changing machine. In Eq.(3.28), B is the trace of $\mathbf{A}$. Since the dot product, or scalar product, is invariant under coordinate transformations the trace of a matrix is the same in all frames. Generally, if $\mathbf{S} \in \mathbf{V}^{k-1}_{l-1}$, is the *contraction* of $\mathbf{T} \in \mathbf{V}^k_l$

$$\delta^{\underline{e}}_{\underline{b}} : T^{\ \underline{a}\ldots\underline{b}\ldots\underline{c}}_{\ \underline{d}\ldots\underline{e}\ldots\underline{f}} = S^{\ \underline{a}\ldots x\ldots\underline{c}}_{\ \underline{d}\ldots x\ldots\underline{f}} \ , \tag{3.31}$$

where x is the usual component index summed over as a dummy index. Often the contracted tensor is denoted by the same symbol with one upper and one lower index missing. This is convenient if the number of indices is not too large but becomes confusing in a general formula, like Eq.(3.21). The elements of $\mathcal{T}^{\underline{a}}$ are called *contravariant vectors* and of $\mathcal{T}_{\underline{a}}$ *covariant vectors*. Elements of $\mathcal{T}^{\underline{a}\ldots\underline{c}}$, for two or more indices, are called *contravariant tensors*, of $\mathcal{T}_{\underline{a}\ldots\underline{c}}$ *covariant tensors* and of $\mathcal{S}^{\ \underline{a}\ldots\underline{c}}_{\ \underline{d}\ldots\underline{f}}$, with at least one upper and one lower index, *mixed tensors*. Vectors are tensors of rank 1 and scalars are tensors of rank 0.

In this notation f_i will be different for different objects being coordinatized. For the point it will be denoted as before. For the contravariant vector it will be

$$\delta^{a}_{\underline{a}} : \mathcal{T}^{\underline{a}} \to \mathbb{R}^n \text{ given by } \delta^{a}_{\underline{a}} A^{\underline{a}} = A^a \ , \tag{3.32}$$

and for the covariant vector by

$$\delta^{\underline{a}}_{a} : \mathcal{T}_{\underline{a}} \to \mathbb{R}^n \text{ given by } \delta^{\underline{a}}_{a} X_{\underline{a}} = X_a \ , \tag{3.33}$$

so that $\delta^{a}_{\underline{a}}$ and $\delta^{\underline{a}}_{a}$ are the analogues of the basis vectors $\mathbf{e}_a$ and $\mathbf{e}^a$. As the scalar product on the manifold is the same as in the coordinate system,

$$A^{\underline{a}} X_{\underline{a}} \ = \ A^a X_a = \left(\delta^{a}_{\underline{a}} A^{\underline{a}}\right)\left(\delta^{\underline{b}}_{a} X_{\underline{b}}\right) = A^{\underline{a}}\left(\delta^{a}_{\underline{a}} \delta^{\underline{b}}_{a}\right) X_{\underline{b}} \ . \tag{3.34}$$

Comparing this equation with Eq.(3.19) gives

$$\delta^{a}_{\underline{a}} \delta^{\underline{b}}_{a} = \delta^{\underline{b}}_{a} \delta^{a}_{\underline{a}} = \delta^{\underline{b}}_{\underline{a}} \ . \tag{3.35}$$

i.e. the $\delta^a_{\underline{a}}$ and $\delta^{\underline{a}}_a$ are inverses of each other. Hence

$$\delta^a_{\underline{a}}\delta^{\underline{a}}_b = \delta^{\underline{a}}_b\delta^a_{\underline{a}} = \delta^a_b \ , \tag{3.36}$$

so that the vectors can be written in terms of their components as

$$A^{\underline{a}} = A^a\delta^{\underline{a}}_a \ ; \ X_{\underline{a}} = X_a\delta^a_{\underline{a}} \ . \tag{3.37}$$

Generally, the coordinatization of a tensor is given by

$$\delta^a_{\underline{a}}...\delta^c_{\underline{c}}\delta^{\underline{d}}_d...\delta^{\underline{f}}_f T^{\ \underline{a}...\underline{c}}_{\ \ \underline{d}...\underline{f}} = T^{\ a...c}_{\ \ d...f}\ (\mathbf{x}) \tag{3.38}$$

and the tensor can be expressed in terms of its components by

$$T^{\ \underline{a}...\underline{c}}_{\ \ \underline{d}...\underline{f}} = \delta^{\underline{a}}_a...\delta^{\underline{c}}_c\delta^{\underline{d}}_d...\delta^{\underline{f}}_f T^{\ a...c}_{\ \ d...f}\ (\mathbf{x}) \ . \tag{3.39}$$

We take the tensors to be defined at a point, P, in the intersection of two coordinate patches with $f_i(P) = x^a$, $f_j(P) = x^{\hat{a}}$. Since f_i^{-1} corresponds to $\delta^{\underline{a}}_a$ and f_j to $\delta^{\hat{a}}_{\underline{a}}$ for contravariant vectors, the coordinate transformation $f_jf_i^{-1}$ for them is

$$\delta^{\underline{a}}_a\delta^{\hat{a}}_{\underline{a}} = \delta^{\hat{a}}_{\underline{a}}\delta^{\underline{a}}_a = \delta^{\hat{a}}_a \ . \tag{3.40}$$

Similarly, for covariant vectors the coordinate transformation is

$$\delta^a_{\underline{a}}\delta^{\underline{a}}_{\hat{a}} = \delta^{\underline{a}}_{\hat{a}}\delta^a_{\underline{a}} = \delta^a_{\hat{a}} \ , \tag{3.41}$$

i.e. the scalar product of the basis vector in one coordinate system with its dual in the other, gives the coordinate transformation between the two systems.

If an index in a "δ" contracts with one in a tensor, the net result is to replace the index in the tensor with the other index in the "δ". Though all four types of "δ"s perform this function, they are four geometrically distinct objects. Two of these, $\delta^{\underline{a}}_{\underline{b}}$ and δ^a_b are identity operators, the former from one tangent space copy to another and the latter the unit matrix. In our analogy, this is a machine to replace jam labels by other jam labels and pickle labels by other pickle labels. The third type are the basis vectors/coordinatizations $\delta^{\underline{a}}_a$ and $\delta^a_{\underline{a}}$, while the fourth are the coordinate transformations $\delta^{\hat{a}}_a$ and $\delta^a_{\hat{a}}$. A particular choice of basis vectors is the moving enad of section 1, which is often denoted by $\eta^{\underline{a}}_a$ or $\eta^a_{\underline{a}}$. In flat space the latter is obtained from the former by direct identification in Cartesian coordinates. This is not possible in curved spaces. Eq.(3.15) for coordinate transformation of the components of a tensor is often taken as the definition of a tensor. This can cause confusion. A tensor is an *invariant quantity*. Its components are covariant and contravariant. Defining tensors in terms of their components hides their geometric significance. The distinction between tensors and their components is necessary to understand the Principle of General Covariance. Physical laws are defined on the manifold while their actual testing is done through measurements made in a particular coordinate frame, by some observer. Changing from one observer to the other amounts to transforming coordinates. This must leave the physical laws (on the manifold) unchanged while transforming the measured quantities.

Very convenient calculational tools, which make expressions more compact and transparent, are the *symmetrizer* and *skew* brackets. We can construct

symmetric and skew symmetric tensors from an arbitrary tensor, $A_{\underline{ab}}$, denoted by

$$\left.\begin{array}{l} A_{(\underline{ab})} = \frac{1}{2}\left(A_{\underline{ab}} + A_{\underline{ba}}\right) = A_{(\underline{ba})} \\ A_{[\underline{ab}]} = \frac{1}{2}\left(A_{\underline{ab}} - A_{\underline{ba}}\right) = -A_{[\underline{ba}]} \ . \end{array}\right\} . \tag{3.42}$$

Thus $A_{\underline{ab}}$ can be broken into totally symmetric and totally skew parts,

$$A_{\underline{ab}} = A_{(\underline{ab})} + A_{[\underline{ab}]} \ . \tag{3.43}$$

For covariant tensors of rank k we can define totally symmetric and totally skew tensors

$$\left.\begin{array}{l} A_{(\underline{a}\ldots\underline{c})} = \frac{1}{k!}\left(A_{\underline{a}\ldots\underline{c}} + \text{ all permutations of } \underline{a}\ldots\underline{c}\right) \\ A_{[\underline{a}\ldots\underline{c}]} = \frac{1}{k!}\left(A_{\underline{a}\ldots\underline{c}} + \text{ all even} - \text{all odd permutations}\right) \end{array}\right\} , \tag{3.44}$$

and similarly for contravariant tensors. Generally, some indices could be symmetrized and/or skewed while leaving other indices as they are. The indices within either bracket to be left as they are will be put between bars "||". For example

$$\begin{aligned} A^{\underline{a}}{}_{\underline{b}[\underline{cde}]} = \frac{1}{3!}\Big(&A^{\underline{a}}{}_{\underline{bcde}} + A^{\underline{a}}{}_{\underline{bdec}} + A^{\underline{a}}{}_{\underline{becd}} - A^{\underline{a}}{}_{\underline{bced}} \\ &- A^{\underline{a}}{}_{\underline{bdce}} - A^{\underline{a}}{}_{\underline{bedc}}\Big) \ , \end{aligned} \tag{3.45}$$

$$\begin{aligned} A_{(\underline{a}[\underline{b}]|\underline{c}|\underline{d}]} &= \frac{1}{2!}\left(A_{(\underline{ab})\underline{cd}} - A_{(\underline{ad})\underline{cb}}\right) \\ &= \frac{1}{2!2!}\left(A_{\underline{abcd}} + A_{\underline{adbc}} - A_{\underline{dabc}} - A_{\underline{dacb}}\right) \ . \end{aligned} \tag{3.46}$$

A fully covariant or contravariant tensor of rank greater than the dimension can not be fully skew, as it would then be zero. It is not easy to break higher rank tensors into parts where all indices are either symmetrized or skewed. For example for a tensor of rank 3 the extension of Eq.(3.33) is

$$A_{\underline{abc}} = A_{(abc)} + \frac{1}{3}A_{(\underline{a}[\underline{b})\underline{c}]} + \frac{1}{3}A_{[\underline{a}(\underline{b}]\underline{c})} + \frac{1}{3}A_{(\underline{ab})\underline{c}} + \frac{1}{3}A_{[\underline{ab}]\underline{c}} + A_{[\underline{abc}]} \ . \tag{3.47}$$

This is called a break-up into the *irreducible parts*.

3.5 The Metric Tensor and Covariant Differentiation

No magnitude for vectors has been defined so far. In pure mathematical terminology $\mathcal{D}$ and $\mathcal{D}^*$ are not *normed spaces*, even through a concept of magnitude is associated with vectors. If we could define a product of two elements of $\mathcal{D}$ (or of $\mathcal{D}^*$) to give a real number (called an *inner product*) we could use it to define the *norm* or "magnitude" of a vector. Since a scalar product is defined between elements of $\mathcal{D}$ and $\mathcal{D}^*$, we can define an inner product by associating a unique covariant vector with a contravariant vector and vice versa. We define the *metric tensor* and its inverse by

$$\left.\begin{array}{l} \mathbf{g} : \mathcal{D} \to \mathcal{D}^* \text{ given by } \mathbf{g}(\mathbf{A}) = \mathbf{X} \ , \\ \mathbf{g}^{-1} : \mathcal{D} \to \mathcal{D}^* \text{ given by } \mathbf{g}^{-1}(\mathbf{X}) = \mathbf{A} \ . \end{array}\right\} \tag{3.48}$$

In the abstract index notation we replace "**X**" by "**A**", relying on the index position to indicate that it is a covariant vector, so as to make the unique association explicit,

$$\left.\begin{array}{l} g_{\underline{ab}} : \mathcal{T}^{\underline{a}} \to \mathcal{T}_{\underline{b}} \text{ given by } g_{\underline{ab}} A^{\underline{a}} = A_{\underline{b}} \ , \\ g^{\underline{ab}} : \mathcal{T}_{\underline{a}} \to \mathcal{T}^{\underline{b}} \text{ given by } g^{\underline{ab}} A_{\underline{a}} = A^{\underline{b}} \ . \end{array}\right\} \tag{3.49}$$

Thus indices are raised and lowered by $g^{\underline{ab}}$ and $g_{\underline{ab}}$ in this notation.

The metric tensor must be symmetric, $g_{\underline{ab}} = g_{\underline{ba}}$, so that the "distance" from the initial point, P, to the final point, Q, should be the same as from Q to P. If the metric tensor is positive definite (see section 2.3 and specially Eqs.(2.37) and (2.38)) the space is said to be *Riemannian* in general and *Euclidean* if flat. If the metric tensor is not positive definite the space is said to be *Lobachevskian*, or *pseudo-Riemannian*, in general and *Minkowskian* if flat. An additional requirement for the space to be Riemannian is that the *triangular inequality* hold. Defining

$$A^2 = g_{\underline{ab}} A^{\underline{a}} A^{\underline{b}} \tag{3.50}$$

as the "square of the magnitude of **A**", the triangular inequality is

$$A + B \geq C \ , \tag{3.51}$$

where $C^{\underline{a}} = A^{\underline{a}} + B^{\underline{a}}$ and the equality holds only when all three vectors are proportional. This equation is another way of stating the well known Euclidean theorem that the sum of the lengths of two sides of a triangle is greater than the length of the third side. This requirement can not be expected to hold in Lobachevskian spaces where A^2 etc., can be negative. As Riemannian geometry is better developed it is fortunate that many of the results are applicable to Lobachevskian spaces as well.

We can use Eqs.(3.17) and (3.39) to express the fact that $g^{\underline{ab}}$ is the inverse of $g_{\underline{ab}}$ in abstract index notation

$$A^{\underline{c}} = g^{\underline{bc}} A_{\underline{b}} = g^{\underline{bc}} g_{\underline{ab}} A^{\underline{a}} = \delta^{\underline{c}}_{\underline{a}} A^{\underline{a}}$$

or

$$g_{\underline{ab}} g^{\underline{bc}} = g^{\underline{bc}} g_{\underline{ab}} = \delta^{\underline{c}}_{\underline{a}} \ . \tag{3.52}$$

In other words the contraction of the metric tensor with its inverse is the identity.

The coordinatization of the metric tensor is the n-dimensional generalization of the metric coefficients of section 2.3. The difference between an *essential* and *coordinate* singularly is that at the former $\mathbf{g}$ and/or $\mathbf{g}^{-1}$ cease to exist, while at the latter g_{ab} or g^{ab} become infinite somewhere but the tensors $g_{\underline{ab}}$ and $g^{\underline{ab}}$ continue to exist. For an infinitesimal displacement vector, dx^a, the *metric* is defined by Eq.(2.44), with $a, b = 1, ..., n$. A manifold with a metric tensor defined on it $(\mathcal{M}_n, \mathbf{g})$ is called a *metric space*. A metric space is obviously an inner product space, which in turn is a normed space. The converse statements are not valid.

The *affine connection*, ∇ is defined as a mapping from $\mathbf{V}^k_l$ to $\mathbf{V}^k_{l+1}$, written in abstract index notation as

$$\nabla_{\underline{p}} : \mathcal{T}^{\ \underline{a}\cdots\underline{c}}_{\ \underline{d}\cdots\underline{f}} \to \mathcal{T}^{\ \underline{a}\cdots\underline{c}}_{\ \underline{d}\cdots\underline{fp}} \ , \tag{3.53}$$

which acts on a scalar function like a partial derivative, in the sense that

$$\delta^{p}_{\underline{p}}\nabla_{\underline{p}} = \frac{\partial f}{\partial x^{p}} := f_{,p} \tag{3.54}$$

and satisfies the Leibniz rule, so that for appropriate rank tensors $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, it gives

$$\nabla\left(\mathbf{A}+\mathbf{BC}\right) = \nabla\mathbf{A} + \left(\nabla\mathbf{B}\right)\mathbf{C} + \mathbf{B}\left(\nabla\mathbf{C}\right) \ . \tag{3.55}$$

In particular, if $\mathbf{B}$ is a scalar, B, and $\mathbf{A}$ and $\mathbf{C}$ are equal valence tensors we have the usual linear combination given by $\mathbf{A}+B\mathbf{C}$.

It is natural to require that

$$\nabla_{\underline{a}}A^{\underline{b}} = 0 \Leftrightarrow \nabla_{\underline{a}}A^{\underline{c}} = 0 \Leftrightarrow \nabla_{\underline{a}}A_{\underline{c}} = 0 \ . \tag{3.56}$$

Now the first equivalence gives

$$\nabla_{\underline{a}}A^{\underline{c}} = \nabla_{\underline{a}}\left(\delta^{\underline{c}}_{\underline{b}}A^{\underline{b}}\right) = \left(\nabla_{\underline{a}}\delta^{\underline{c}}_{\underline{b}}\right)A^{\underline{b}} + \delta^{\underline{c}}_{\underline{b}}\left(\nabla_{\underline{a}}A^{\underline{b}}\right) = 0 \ .$$

Therefore

$$\nabla_{\underline{a}}\delta^{\underline{c}}_{\underline{b}} = 0 \ . \tag{3.57}$$

Similarly

$$\nabla_{\underline{a}}A_{\underline{c}} = \nabla_{\underline{a}}\left(g_{\underline{bc}}A^{\underline{b}}\right) = \left(\nabla_{\underline{a}}g_{\underline{bc}}\right)A^{\underline{b}} + g_{\underline{bc}}\left(\nabla_{\underline{a}}A^{\underline{b}}\right) = 0 \ .$$

Hence

$$\nabla_{\underline{a}}g_{\underline{bc}} = 0 \ . \tag{3.58}$$

As $\delta^{\underline{c}}_{\underline{b}}$ is the identity, Eq.(3.47) holds automatically. However, Eq.(3.48) connects the choice of affine connection with the choice of metric tensor, or vice-versa. Neither of them is unique.

Consider the coordinatization of the tensor, $\mathbf{B}$, obtained by acting on a contravariant vector, $\mathbf{A}$, with the affine connection. This is called the *covariant derivative* of $\mathbf{A}$,

$$\begin{aligned} B^{a}_{\ b} := A^{a}_{\ ;b} &= \delta^{a}_{\underline{a}}\delta^{\underline{b}}_{b}\nabla_{\underline{b}}A^{\underline{a}} = \delta^{a}_{\underline{a}}\delta^{\underline{b}}_{b}\nabla_{\underline{b}}\left(\delta^{\underline{a}}_{c}A^{c}\right) \\ &= \left(\delta^{a}_{\underline{a}}\delta^{\underline{b}}_{b}\nabla_{\underline{b}}\delta^{\underline{a}}_{c}\right)A^{c} + \delta^{a}_{\underline{a}}\delta^{\underline{a}}_{c}\left(\delta^{\underline{b}}_{\underline{b}}\nabla_{b}A^{c}\right) \ . \end{aligned} \tag{3.59}$$

Since A^{c} is a set of n scalar functions Eqs.(3.25) and (3.44) can be used to reduce the second term to $A^{a}_{\ ,b}$. The expression in brackets in the first term is the derivatives of all the basis vectors contracted with all the dual basis vectors, analogous to the Christoffel symbols appearing in the Gauss (-Weingarten) equations, (2.135), called the *connection symbols*

$$\Gamma^{a}_{\ bc} = \delta^{a}_{\underline{a}}\delta^{\underline{b}}_{b}\nabla_{\underline{b}}\delta^{\underline{a}}_{c} \ . \tag{3.60}$$

We can thus write the covariant derivative of A^{a} as

$$A^{a}_{\ ;b} = A^{a}_{\ ,b} + \Gamma^{a}_{\ bc}A^{c} \ . \tag{3.61}$$

We can write Eq.(3.50) in the form of the Gauss(-Weingarten) equations by taking the inverse coordinatizations of $\delta^{a}_{\underline{a}}$ and $\delta^{\underline{b}}_{b}$ so as to express the derivatives

of the basis vectors in terms of the connection symbols. Contracting Eq.(3.50) with $\delta^c_{\underline{a}}$ and $\delta^b_{\underline{d}}$, using Eqs. (3.25) and (3.47) we get

$$\nabla_{\underline{d}}\delta^{\underline{c}}_{c} = \Gamma^{a}_{bc}\delta^{b}_{\underline{d}}\delta^{\underline{c}}_{a} \ . \tag{3.62}$$

The covariant derivative of a covariant vector, $\mathbf{X}$, is evaluated by forming the scalar product $\mathbf{A}\cdot\mathbf{X}$ and using Eqs.(3.47) and (3.51).

$$\begin{aligned}(A^aX_a)_{;b} &= (A^aX_a)_{,b} = A^a{}_{,b}X_a + A^aX_{a,b} \\ &= A^a{}_{;b}X_a + A^aX_{a;b} = A^a{}_{,b}X_a + \Gamma^a_{bc}A^cX_a + A^aX_{a;b} \ .\end{aligned} \tag{3.63}$$

In the second last term we can interchange the dummy indices "a" and "c".

$$A^aX_{a;b} = A^a\left(X_{a,b} - \Gamma^c_{ba}X_c\right) \ . \tag{3.64}$$

This equation holds true for all choices of A^a. In particular, it holds true if we choose $A^1 = 1$, $A^a = 0$ otherwise. Similarly for $A^2 = 1$, $A^a = 0$ otherwise and so on. Hence each component separately satisfies the equation and A^a can be dropped. Thus

$$X_{a;b} = X_{a,b} - \Gamma^c_{ba}X_c \ . \tag{3.65}$$

The difference between the affine connection and the partial derivative is that the former acts on the basis vectors as well as the components while the latter acts only on the components. Since the covariant derivative is defined on the manifold while the partial derivative is not, the connection symbols are not components of a tensor. This point will be demonstrated rigorously shortly. Before doing so we need to evaluate the coordinate transformation.

Consider the affine connection acting on the coordinatization at a point P. By virtue of Eq.(3.44)

$$\delta^b_{\underline{b}}\nabla_{\underline{b}}f\left(P\right) = f\left(P\right)_{,b} = x^a_{\ ,b} = \delta^a_b \ . \tag{3.66}$$

Comparing Eq.(3.66) with Eq.(3.25) we see that

$$\nabla_{\underline{b}}f\left(P\right) = \nabla_{\underline{b}}x^a = \delta^a_{\underline{b}} \ , \tag{3.67}$$

i.e. the affine connection acting on the coordinatization gives the covariant basis vector. Thus using Eq.(3.57), we can express the derivation, $\mathbf{A}$, acting on the coordinatization as

$$\begin{aligned}\mathbf{A}\left[f\left(P\right)\right] &= A^a = A^{\underline{a}}\delta^a_{\underline{a}} \ , \\ &= A^{\underline{a}}\nabla_{\underline{a}}x^a = A^{\underline{a}}\nabla_{\underline{a}}\left[f\left(P\right)\right] := A^a\nabla_a\left[f\left(P\right)\right] \ .\end{aligned} \tag{3.68}$$

Thus the derivative acts like a derivative operator − which is why it is called a derivation and why we write $\mathbf{A}(f)$ instead of $f(\mathbf{A})$. Notice that "$\nabla_{\underline{a}}$" acts like a basis vector, if we drop the explicit action on the coordinatization. Now, from Eq.(3.40), we have

$$\begin{aligned}\delta^a_{\hat{a}} &= \delta^{\underline{a}}_{\hat{a}}\delta^a_{\underline{a}} = \delta^{\underline{a}}_{\hat{a}}\nabla_{\underline{a}}f\left(P\right) \\ &= \frac{\partial f\left(P\right)}{\partial x^{\hat{a}}} = \frac{\partial x^a}{\partial x^{\hat{a}}} \ .\end{aligned} \tag{3.69}$$

Thus the coordinate transformation is of the form given by Eq.(2.58).

Let us evaluate the connection symbols in $\hat{\mathbb{R}}^n$. By definition

$$\begin{aligned}\Gamma^{\hat{a}}_{\hat{b}\hat{c}} &= \delta^{\hat{a}}_{\underline{a}}\delta^{\underline{b}}_{\hat{b}}\nabla_{\underline{b}}\delta^{\underline{a}}_{\hat{c}} = \delta^{\hat{a}}_{a}\delta^{a}_{\underline{a}}\delta^{\underline{b}}_{\hat{b}}\delta^{\underline{b}}_{\underline{b}}\nabla_{\underline{b}}\left(\delta^{\underline{a}}_{\underline{c}}\delta^{c}_{\hat{c}}\right)\\ &= \delta^{\hat{a}}_{a}\delta^{\underline{b}}_{\hat{b}}\delta^{c}_{\hat{c}}\left(\delta^{a}_{\underline{a}}\delta^{\underline{b}}_{\underline{b}}\nabla_{\underline{b}}\delta^{\underline{a}}_{\underline{c}}\right) + \delta^{\hat{a}}_{a}\delta^{a}_{\underline{a}}\delta^{\underline{a}}_{\underline{c}}\left(\delta^{\underline{b}}_{\hat{b}}\nabla_{\underline{b}}\delta^{c}_{\hat{c}}\right)\\ &= \delta^{\hat{a}}_{a}\delta^{\underline{b}}_{\hat{b}}\delta^{c}_{\hat{c}}\Gamma^{a}_{bc} + \frac{\partial x^{\hat{a}}}{\partial x^{c}}\frac{\partial^2 x^{c}}{\partial x^{\hat{c}}\partial x^{\hat{b}}}\ .\end{aligned} \tag{3.70}$$

If Γ^a_{bc} had been components of a tensor the second term would have been zero. Thus, for linear transformations (in which the second term is zero), these symbols behave as tensors. Generally, however, the second term is non-zero and these symbols do not behave as tensors, i.e. they are not components of a tensor.

We can construct a tensor quantity from Γ^a_{bc}. Consider

$$T^a_{bc} = 2\Gamma^a_{[bc]} := \Gamma^a_{bc} - \Gamma^a_{cb}\ . \tag{3.71}$$

Using Eq.(3.60) we have

$$T^{\hat{a}}_{\hat{b}\hat{c}} = \delta^{\hat{a}}_{a}\delta^{b}_{\hat{b}}\delta^{c}_{\hat{c}}T^{a}_{bc} + \frac{\partial x^{\hat{a}}}{\partial x^{c}}\frac{\partial^2 x^{c}}{\partial x^{[\hat{c}}\partial x^{\hat{b}]}}\ . \tag{3.72}$$

The last term is zero as it gives the difference between the second derivative of x^c which respect to ($x^{\hat{c}}$ and $x^{\hat{b}}$) and ($x^{\hat{b}}$ and $x^{\hat{c}}$). Since the order of differentiation is irrelevant for analytic functions the two must cancel. Since the T^a_{bc} transform as components of a tensor we can use the inverse coordinatization to define the *torsion tensor*, $T^{\underline{a}}_{\underline{b}\underline{c}}$. This tensor has nothing to do with the "torsion" of section 2.2. This is another reason why the term "second curvature" is preferred to "torsion". Cartan and Einstein tried to develop a theory in which this tensor provided the "spin" part of the stress-energy tensor (explained in the next chapter). The theory was unsuccessful. In standard GR it is assumed that $T^{\underline{a}}_{\underline{b}\underline{c}} = 0$. Hence in all coordinate systems the connection symbols are symmetric in such spaces (called *torsion-free* spaces).

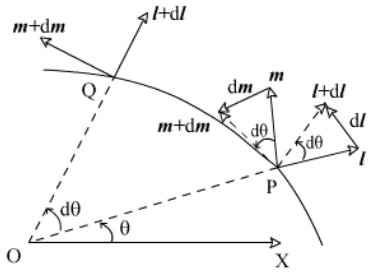


Figure 3.8: A point $P(r,\theta)$ and a neighbouring point $Q(r+dr,\theta+d\theta)$ lie on the orbit of a particle. The origin of the central force holding the particle in its orbit is O at $r=0$. The radial and polar unit vectors at P are $\mathbf{l}$ and $\mathbf{m}$ and at Q are $\mathbf{l}+d\mathbf{l}$ and $\mathbf{m}+d\mathbf{m}$. Taking the vectors at Q back to P we get $d\mathbf{l}$ and $d\mathbf{m}$. Since $\mathbf{l}$ is a unit vector, in the limit $d\theta\to 0$, $d\mathbf{l}\to\mathbf{m}d\theta$, $d\mathbf{m}\to-\mathbf{l}d\theta$.

The common use of Cartesian coordinates causes the naive intuition to regard basis vectors as constant. It is, therefore, useful to consider a familiar physical example in which basis vectors are differentiated. This occurs in the derivation of the equations of motion under a central force using plane polar coordinates, see Fig. 3.8. Writing the radial and polar unit vectors as $\mathbf{l}$ and $\mathbf{m}$

$$\frac{d^2\mathbf{r}}{dt^2} = \frac{d}{dt}\left[\frac{d}{dt}(r\mathbf{l})\right] = \frac{d}{dt}\left(\dot{r}\mathbf{l} + r\frac{d\mathbf{l}}{dt}\right) . \tag{3.73}$$

As is clear from Fig. 3.8

$$\frac{\partial\mathbf{l}}{\partial\theta} = \mathbf{m} , \ \frac{\partial\mathbf{m}}{\partial\theta} = -\mathbf{l} . \tag{3.74}$$

Therefore

$$\frac{d\mathbf{l}}{dt} = \frac{d\theta}{dt}\frac{\partial\mathbf{l}}{\partial\theta} = \dot{\theta}\mathbf{m} , \frac{d\mathbf{m}}{dt} = \frac{d\theta}{dt}\frac{\partial\mathbf{m}}{\partial\theta} = -\dot{\theta}\mathbf{l} . \tag{3.75}$$

Hence

$$\begin{aligned} \frac{d^2\mathbf{r}}{dt^2} &= \ddot{r}\mathbf{l} + \dot{r}\dot{\theta}\mathbf{m} + \dot{r}\dot{\theta}\mathbf{m} + r\ddot{\theta}\mathbf{m} - r\dot{\theta}^2\mathbf{l} \\ &= \left(\ddot{r} - r\dot{\theta}^2\right)\mathbf{l} + \left(r\ddot{\theta} + 2\dot{r}\dot{\theta}\right)\mathbf{m} . \end{aligned} \tag{3.76}$$

Since the central force has no polar component the coefficient of $\mathbf{m}$ is zero. The second term is equated to the magnitude of the force. Notice that in Eqs.(3.64) the derivatives of the basis vectors have been used. Contracting the equations with $\mathbf{l}$ and $\mathbf{m}$ gives

$$\Gamma^1_{21} = 0 \ , \ \Gamma^1_{22} = 1 \ , \ \Gamma^2_{21} = -1 \ , \ \Gamma^2_{22} = 0 . \tag{3.77}$$

These are not the connection symbols for polar coordinates as the polar basis vector, $\mathbf{m}$, is a vector of unit *length*. In polar coordinates the polar component of a vector is an *angle*.

The covariant derivative of $\mathbf{A} \in \mathbf{V}^k_l$ can be evaluated by expanding the tensor in terms of its basis vectors and its components, using the fact that

$$\nabla_{\underline{a}}\delta^{\underline{b}}_{\underline{b}} = -\Gamma^b_{ca}\delta^a_{\underline{a}}\delta^c_{\underline{b}} , \tag{3.78}$$

as may be seen by using Eq.(3.25) in reverse in Eq.(3.47) and using Eq.(3.52) with appropriately changed indices. Thus

$$\begin{aligned} A^{\ a\ldots c}_{\ d\ldots f;p} =& \delta^a_{\underline{a}}\ldots\delta^c_{\underline{c}}\delta^{\underline{d}}_{\underline{d}}\ldots\delta^{\underline{f}}_{\underline{f}}\delta^{\underline{p}}_{\underline{p}}\nabla_{\underline{p}}A^{\ \underline{a}\ldots\underline{c}}_{\ \underline{d}\ldots\underline{f}} \\ =& A^{\ a\ldots c}_{\ d\ldots f,p} + \Gamma^a_{px}A^{\ x\ldots c}_{\ d\ldots f} + \ldots + \Gamma^c_{px}A^{\ a\ldots x}_{\ d\ldots f} \\ & - \Gamma^x_{pd}A^{\ a\ldots c}_{\ x\ldots f} - \ldots - \Gamma^x_{pf}A^{\ a\ldots c}_{\ d\ldots x} . \end{aligned} \tag{3.79}$$

We can use Eq.(3.48) to evaluate the connection symbol in terms of the metric coefficients and their derivatives. Using Eq.(3.69) for the metric tensor we have

$$g_{bc;d} = 0 = g_{bc,d} - \Gamma^x_{db}g_{xc} - g_{bx}\Gamma^x_{dc} , \tag{3.80}$$

$$g_{bd;c} = 0 = g_{bd,c} - \Gamma^x_{cb}g_{xd} - g_{bx}\Gamma^x_{cd} , \tag{3.81}$$

$$g_{cd;b} = 0 = g_{cd,b} - \Gamma^x_{bc}g_{xd} - g_{cx}\Gamma^x_{bd} . \tag{3.82}$$

Using the symmetry properties of g_{ab} and Γ^a_{bc} (in a torsion-free space), subtracting Eq.(3.70) from the sum of Eqs.(3.71) and (3.72) we get

$$0 = g_{bd,c} + g_{cd,b} - g_{bc,d} - 2\Gamma^x_{bc} g_{xd} \ . \tag{3.83}$$

Multiplying this equation through by $\frac{1}{2}g^{ad}$ and transposing yields

$$\Gamma^a_{bc} = \frac{1}{2} g^{ad} \left(g_{bd,c} + g_{cd,b} - g_{bc,d} \right) \ . \tag{3.84}$$

Comparing Eq.(3.74) with Eq.(2.135) we see that the connection symbol is identical with the Christoffel symbol, $\{_b{}^a{}_c\}$. (A word of caution here. This identity holds as we have used a coordinate basis in a torsion-free space throughout, but would not hold in general, see [15] for details.) Henceforth we shall use $\{_b{}^a{}_c\}$ instead of Γ^a_{bc}.

The number of Christoffel symbol components is n^3. Due to the symmetry property,for each upper index there are $n\left(n-1\right)/2$ symbols which are equal to the other $n\left(n-1\right)/2$ symbols with $b \neq c$. Thus there are $n\left(n-1\right)/2 + n = n\left(n+1\right)/2$ independent symbols for each upper index. The total number is, then, $n^2\left(n+1\right)/2$. For 2 dimensions this is 6, for 3 it is 18 and for 4 it is 40. Any procedure that can reduce the number of independent components and/or reduce the computation involved in each component, would be extremely useful. This can be done for diagonal metric tensors. In that case if $a = b$, since $d = a$ for a non-zero component in Eq.(3.74), the middle and last terms cancel. If $a \neq b = c$, the first and second terms are zero. If $a \neq b \neq c \neq a$ all terms are zero. Hence for diagonal metric tensors, we have,

$$\left.\begin{array}{ll} \{_b{}^a{}_c\} & = \frac{1}{2} g^{ad} g_{bd,c} = \{_c{}^a{}_b\} \quad \text{if } a = b \ , \\ & = -\frac{1}{2} g^{ad} g_{bc,d} \quad\quad\ \text{if } a \neq b = c \ , \\ & = 0 \quad\quad\quad\quad\quad\quad\ \text{if } a \neq b \neq c \neq a \ . \end{array}\right\} \tag{3.85}$$

This procedure reduces the complexity of calculation as summation is eliminated ($d = a$ always), and the number of components when all three indices are different. In n dimensions there are $n(n-1)(n-2)$ ways of choosing 3 distinct indices. Symmetry in the lower two indices leaves $n(n-1)(n-2)/2$ components that are zero. For $n = 2$ this is 0, for $n = 3$ it is 3 and for $n = 4$ it is 12. Thus, for a diagonal metric tensor, there are 6, 15 and 28 independent components to be computed in 2, 3, and 4 dimensions. Unfortunately, physical requirements often force use of a non-diagonal metric tensor, even though it could have been diagonalised in principle. The problem is that if we define a diagonalising procedure at each point, it can happen that following a closed path, there can be two distinct diagonalised metrics at one point, which would be absurd.

To make the rather abstract discussion so far more concrete, it is necessary to consider some simple examples.

Example 1 : Euclidean 2-d space, (a plane in polar coordinates). We use $x^a = (x, y)$, $x^{\hat{a}} = (r, \theta)$. Then $g_{ab} = 1$ if $a = b$ and $g_{ab} = 0$ if $a \neq b$. Also

$$\delta^a_{\hat{a}} = \begin{pmatrix} \partial x/\partial r & \partial x/\partial\theta \\ \partial y/\partial r & \partial y/\partial\theta \end{pmatrix} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix} \ ,$$

as $x = r\cos\theta$, $y = r\sin\theta$. Conversely $r = \sqrt{x^2 + y^2}$, $\theta = \tan^{-1} y/x$.

Therefore

$$\delta_a^{\hat{a}} = \begin{pmatrix} \partial r/\partial x & \partial r/\partial y \\ \partial\theta/\partial x & \partial\theta/\partial y \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta/r & \cos\theta/r \end{pmatrix} .$$

Clearly, $\delta_a^{\hat{a}}$ is the inverse of $\delta_{\hat{a}}^a$ and the determinant, $\Delta = r$, and hence the coordinate transformation is singular at $r = 0$.

Consider a contravariant vector with Cartesian components A^x, A^y and polar components A^r, A^θ. Then by usual matrix multiplication

$$\begin{aligned} \begin{pmatrix} A^r \\ A^\theta \end{pmatrix} &= A^a = \delta_{\hat{a}}^a A^{\hat{a}} \\ &= \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta/r & \cos\theta/r \end{pmatrix} \begin{pmatrix} A^x \\ A^y \end{pmatrix} \\ &= \begin{pmatrix} (A^x\cos\theta + A^y\sin\theta) \\ (A^y\cos\theta - A^x\sin\theta)/r \end{pmatrix} . \end{aligned}$$

Similarly, for a covariant vector with components B_a or $B_{\hat{a}}$,

$$\begin{aligned} \begin{pmatrix} B_r & B_\theta \end{pmatrix} &= \begin{pmatrix} B_x & B_y \end{pmatrix} \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix} \\ &= \begin{pmatrix} (B_x\cos\theta + B_y\sin\theta) , & r\,(B_y\cos\theta - B_x\sin\theta) \end{pmatrix} . \end{aligned}$$

We will write the metric tensor as a square matrix for convenience. In doing so we have to give up the usual matrix multiplication procedure. We have

$$\begin{aligned} g_{\hat{1}\hat{1}} &= \delta_{\hat{1}}^1\delta_{\hat{1}}^1 g_{11} + \delta_{\hat{1}}^1\delta_{\hat{1}}^2 g_{12} + \delta_{\hat{1}}^2\delta_{\hat{1}}^1 g_{21} + \delta_{\hat{1}}^2\delta_{\hat{1}}^2 g_{22} \\ &= \left(\delta_{\hat{1}}^1\right)^2 + \left(\delta_{\hat{1}}^2\right)^2 = \cos^2\theta + \sin^2\theta = 1 , \\ g_{\hat{1}\hat{2}} &= \delta_{\hat{1}}^1\delta_{\hat{2}}^1 + \delta_{\hat{1}}^2\delta_{\hat{2}}^2 = -r\sin\theta\cos\theta + r\sin\theta\cos\theta = 0 = g_{\hat{2}\hat{1}} , \\ g_{\hat{2}\hat{2}} &= \left(\delta_{\hat{2}}^1\right)^2 + \left(\delta_{\hat{2}}^2\right)^2 = r^2\sin^2\theta + r^2\cos^2\theta = r^2 . \end{aligned}$$

Therefore

$$g_{\hat{a}\hat{b}} = \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix} . \tag{3.86}$$

The procedure may be visualized as multiplying the columns of $\delta_{\hat{a}}^a$ with themselves, for the diagonal elements, and each other for the off-diagonal elements. For a diagonal matrix the inverse is obtained by inverting the diagonal elements. Therefore

$$g^{\hat{a}\hat{b}} = \begin{pmatrix} 1 & 0 \\ 0 & 1/r^2 \end{pmatrix} .$$

We now treat the x^a as plane polar coordinates.. The Christoffel symbols are obtained from Eq.(3.84), using the fact that $g_{22,1} = 2r$, $g_{ab,c} = 0$ otherwise. Thus if there are two "1s" in the symbols we must have $g_{11,c}$ which is zero. Also if there are "2s" we have $g_{22,2}$ which is zero.

$$\left.\begin{aligned} &\left\{_2\ ^1\ _2\right\} = -\frac{1}{2}g^{11}g_{22,1} = -r , \ \left\{_1\ ^2\ _2\right\} = \frac{1}{2}g^{22}g_{22,1} = \left\{_2\ ^2\ _1\right\} . \\ &\left\{_1\ ^1\ _1\right\} = \left\{_1\ ^1\ _2\right\} = \left\{_2\ ^1\ _1\right\} = \left\{_2\ ^2\ _2\right\} = 0 . \end{aligned}\right\} \tag{3.87}$$

Thus the covariant derivative of A^a is

$$\begin{aligned}
A^r{}_{;r} &= A^r{}_{,r} + \left\{ {1 \atop 1\ a} \right\} A^a = A^r{}_{,r} \ , \\
A^r{}_{;\theta} &= A^r{}_{,\theta} + \left\{ {2 \atop 1\ a} \right\} A^a = A^r{}_{,\theta} + \left\{ {2 \atop 1\ 2} \right\} A^\theta = A^r{}_{,\theta} - rA^\theta \ , \\
A^\theta{}_{;r} &= A^\theta{}_{,r} + \left\{ {1 \atop 2\ a} \right\} A^a = A^\theta{}_{,r} + \left\{ {1 \atop 2\ 2} \right\} A^r = A^\theta{}_{,r} - \frac{1}{r} A^\theta \ , \\
A^\theta{}_{;\theta} &= A^\theta{}_{,\theta} + \left\{ {2 \atop 2\ a} \right\} A^a = A^\theta{}_{,\theta} + \left\{ {2 \atop 2\ 2} \right\} A^r = A^\theta{}_{,\theta} - \frac{1}{r} A^r \ .
\end{aligned}$$

Example 2 : The surface of a sphere of radius a. We know from Eq. (2.81) that

$$g_{ab} = \begin{pmatrix} a^2 & 0 \\ 0 & a^2 \sin^2\theta \end{pmatrix} \quad , \quad g^{ab} = \begin{pmatrix} 1/a^2 & 0 \\ 0 & 1/a^2 \sin^2\theta \end{pmatrix} \ . \tag{3.88}$$

where $x^1 = \theta$, $x^2 = \phi$. Thus $g_{22,1} = 2a^2 \sin\theta\cos\theta$, $g_{ab,c} = 0$ otherwise. Therefore

$$\left.\begin{aligned}
&\left\{ {1 \atop 1\ 1} \right\} = \left\{ {1 \atop 1\ 2} \right\} = \left\{ {2 \atop 1\ 1} \right\} = \left\{ {2 \atop 2\ 2} \right\} = 0 \ . \\
&\left\{ {2 \atop 1\ 2} \right\} = -\sin\theta\cos\theta \ , \ \left\{ {1 \atop 2\ 1} \right\} = \cot\theta = \left\{ {2 \atop 2\ 1} \right\} \ .
\end{aligned}\right\} \tag{3.89}$$

Consider a contravariant vector with components A^θ, A^ϕ.

Therefore

$$\begin{aligned}
A^\theta{}_{;\theta} &= A^\theta{}_{,\theta} + \left\{ {1 \atop 1\ a} \right\} A^a = A^\theta{}_{,\theta} \ , \\
A^\theta{}_{;\phi} &= A^\theta{}_{,\phi} + \left\{ {2 \atop 1\ a} \right\} A^a = A^\theta{}_{,\phi} - \sin\theta\cos\theta A^\phi \ , \\
A^\phi{}_{;\theta} &= A^\phi{}_{,\theta} + \left\{ {1 \atop 2\ a} \right\} A^a = A^\phi{}_{,\theta} + \cot\theta A^\phi \ , \\
A^\phi{}_{;\phi} &= A^\phi{}_{,\phi} + \left\{ {2 \atop 2\ a} \right\} A^a = A^\phi{}_{,\phi} + \cot\theta A^\theta \ .
\end{aligned}$$

Example 3 : Euclidean 3-dimensional space in (spherical) polar coordinates. (Again the transformations are given in SR. They will not be repeated here for this case but are left as an exercise for the reader.) We take the metric tensor as given

$$g_{ab} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2 \sin^2\theta \end{pmatrix} \quad , \quad g^{ab} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/r^2 & 0 \\ 0 & 0 & 1/r^2 \sin^2\theta \end{pmatrix} \ . \tag{3.90}$$

There are clearly coordinate singularities at $r = 0$, $\theta = 0$ and π. These all lie on the z-axis. The former is a point on it while the latter are the complete axis.

For the Christoffel symbols notice that

$$\begin{aligned}
g_{22,1} &= 2r \ , \ g_{33,1} = 2r\sin^2\theta \ ; \ g_{33,2} = 2r^2 \sin\theta\cos\theta \ , \\
g_{ab,c} &= 0 \text{ otherwise.}
\end{aligned}$$

Since the metric tensor is diagonal

$$\left\{ {2 \atop 1\ 3} \right\} = \left\{ {3 \atop 1\ 2} \right\} = \left\{ {1 \atop 2\ 3} \right\} = \left\{ {3 \atop 2\ 1} \right\} = \left\{ {1 \atop 3\ 2} \right\} = \left\{ {2 \atop 3\ 1} \right\} = 0 \ .$$

Also, from the above the following are zero:

$$\begin{aligned}
&\left\{ {1 \atop 1\ 1} \right\}, \ \left\{ {1 \atop 1\ 2} \right\}, \ \left\{ {2 \atop 1\ 1} \right\}, \ \left\{ {1 \atop 1\ 3} \right\}, \ \left\{ {3 \atop 1\ 1} \right\}, \\
&\left\{ {2 \atop 2\ 2} \right\}, \ \left\{ {2 \atop 2\ 3} \right\}, \ \left\{ {3 \atop 2\ 2} \right\}, \ \left\{ {3 \atop 3\ 3} \right\} \ .
\end{aligned}$$

The following are non-zero:

$$\left.\begin{array}{l}\left\{_2\ ^1\ _2\right\} = -\frac{1}{2}g^{11}g_{22,1} = -r\ , \\ \left\{_1\ ^2\ _2\right\} = \frac{1}{2}g^{22}g_{22,1} = 1/r = \left\{_2\ ^2\ _1\right\}\ , \\ \left\{_3\ ^2\ _3\right\} = -\frac{1}{2}g^{22}g_{33,2} = -\sin\theta\cos\theta\ , \\ \left\{_1\ ^3\ _3\right\} = \frac{1}{2}g^{33}g_{33,1} = 1/r = \left\{_3\ ^3\ _1\right\}\ , \\ \left\{_2\ ^3\ _3\right\} = -\frac{1}{2}g^{33}g_{33,2} = \cot\theta = \left\{_2\ ^3\ _3\right\}\ .\end{array}\right\} \tag{3.91}$$

Notice the fact that the metric tensor here reduces to that in Example 1 if the third dimension is removed and, more interestingly, reduces to that in Example 2, if we put $r = a$ and remove the first dimension. Also notice the similarity and reduction of the Christoffel symbols. This procedure, of obtaining sub-spaces with desired symmetries, is called *projection*.

3.6 The Curvature Tensors and Scalars

The generalization of Gauss' invariant intrinsic curvature of a surface to higher dimensional spaces is obtained by carrying a basis vector along two different directions in opposite order and taking the difference of the two results. This is essentially what was done in the example given at the start of section 3. Mathematically it is done using the operator

$$\Delta_{\underline{ab}} := 2\nabla_{[\underline{b}}\nabla_{\underline{a}]}\ . \tag{3.92}$$

Consider this operator acting on a scalar function

$$\begin{aligned}\Delta_{\underline{ab}}f &= 2\nabla_{[\underline{b}}\left(\delta^a_{\underline{a}]}f_{,a}\right) \\ &= 2\delta^b_{[\underline{b}}\delta^a_{\underline{a}]}f_{,ab} - 2\delta^b_{[\underline{b}}\delta^c_{\underline{a}]}\left\{_c\ ^a\ _b\right\}f_{,a}\ ,\end{aligned} \tag{3.93}$$

where we have used Eq.(3.68). Now

$$2\delta^b_{[\underline{b}}\delta^a_{\underline{a}]} = \delta^b_{\underline{b}}\delta^a_{\underline{a}} - \delta^b_{\underline{a}}\delta^a_{\underline{b}} = \delta^b_{\underline{b}}\delta^a_{\underline{a}} - \delta^a_{\underline{b}}\delta^b_{\underline{a}} = 2\delta^{[b}_{\underline{b}}\delta^{a]}_{\underline{a}}\ . \tag{3.94}$$

Using this result in Eq.(3.83) and shifting the skew to the other dummy indices

$$\Delta_{\underline{ab}}f = 2\delta^b_{\underline{b}}\delta^a_{\underline{a}}f_{,[ab]} - 2\delta^b_{\underline{b}}\delta^c_{\underline{a}}\left\{_{[b}\ ^a\ _{c]}\right\}f_{,a}\ . \tag{3.95}$$

Any object that is symmetric and skew is zero. Thus the first term is zero anyhow. The second term would be the torsion tensor times some basis vectors and the gradient of f. Since we have taken the torsion tensor to be zero

$$\Delta_{\underline{ab}}f = T^{\underline{c}}_{\underline{ab}}\nabla_{\underline{c}}f = 0\ . \tag{3.96}$$

The *Riemann* (or *Riemann-Christoffel*) *curvature tensor* is defined by

$$R^{\underline{a}}_{\ \underline{bcd}} := \delta^b_{\underline{b}}\Delta_{\underline{dc}}\delta^{\underline{a}}_b = 2\delta^b_{\underline{b}}\nabla_{[\underline{c}}\nabla_{\underline{d}]}\delta^{\underline{a}}_b\ . \tag{3.97}$$

Using Eq.(3.52) we get

$$\left.\begin{aligned}R^{\underline{a}}_{\ \underline{bcd}} &:= \delta^b_{\underline{b}}\nabla_{[\underline{c}}\left(\delta^d_{\underline{d}]}\delta^{\underline{a}}_a\left\{_b\ ^a\ _d\right\}\right) \\ &= 2\delta^b_{\underline{b}}\Big[-\delta^c_{[\underline{c}}\delta^e_{\underline{d}]}\delta^{\underline{a}}_a\left\{_c\ ^d\ _e\right\}\left\{_b\ ^a\ _d\right\} + \delta^c_{[\underline{c}}\delta^d_{\underline{d}]}\Big(\delta^{\underline{a}}_e\left\{_c\ ^e\ _a\right\}\left\{_b\ ^a\ _d\right\} \\ &\quad + \delta^{\underline{a}}_a\left\{_b\ ^a\ _d\right\}_{,c}\Big)\Big]\ .\end{aligned}\right\} \tag{3.98}$$

Using Eq.(3.84) the first term is zero on account of the above argument. Interchanging the dummy indices "a" and "e" in the second term and again using Eq.(3.84) we finally obtain

$$\begin{aligned} R^{\underline{a}}_{\underline{bcd}} &:= \delta^{\underline{a}}_{a}\delta^{b}_{\underline{b}}\delta^{c}_{\underline{c}}\delta^{d}_{\underline{d}}R^{a}{}_{bcd} \\ &= 2\delta^{\underline{a}}_{a}\delta^{b}_{\underline{b}}\delta^{c}_{\underline{c}}\delta^{d}_{\underline{d}}\left(\left\{_{b}{}^{a}{}_{[d}\right\}_{,c]} + \left\{_{e}{}^{a}{}_{[c}\right\}\left\{_{d]}{}^{e}{}_{b}\right\}\right) \ . \end{aligned} \tag{3.99}$$

This tensor and its trace

$$\begin{aligned} R_{\underline{bd}} &:= R^{\underline{a}}_{\underline{bad}} \\ &= \delta^{b}_{\underline{b}}\delta^{d}_{\underline{d}}2\left(\left\{_{b}{}^{a}{}_{[d}\right\}_{,a]} + \left\{_{e}{}^{a}{}_{[a}\right\}\left\{_{d]}{}^{e}{}_{b}\right\}\right) = \delta^{b}_{\underline{b}}\delta^{d}_{\underline{d}}R_{bd} \ , \end{aligned} \tag{3.100}$$

called the *Ricci tensor*, and its trace

$$R = g^{\underline{bd}}R_{\underline{bd}} = g^{bd}R_{bd} \ , \tag{3.101}$$

called the *Ricci scalar*, are of prime importance in Relativity. The Ricci scalar is the generalization of the Gaussian curvature and is related to it by a factor that only depends on the dimension of the space.

We write the expanded form of the curvature tensor for purposes of explicit calculation of the components

$$R^{a}{}_{bcd} = \left\{_{b}{}^{a}{}_{d}\right\}_{,c} - \left\{_{b}{}^{a}{}_{c}\right\}_{,d} + \left\{_{e}{}^{a}{}_{c}\right\}\left\{_{b}{}^{e}{}_{d}\right\} - \left\{_{e}{}^{a}{}_{d}\right\}\left\{_{b}{}^{e}{}_{c}\right\} \ , \tag{3.102}$$

$$R_{bd} = \left\{_{b}{}^{a}{}_{d}\right\}_{,a} - \left\{_{b}{}^{a}{}_{a}\right\}_{,d} + \left\{_{e}{}^{a}{}_{a}\right\}\left\{_{b}{}^{e}{}_{d}\right\} - \left\{_{e}{}^{a}{}_{d}\right\}\left\{_{b}{}^{e}{}_{a}\right\} \ . \tag{3.103}$$

There are two Christoffel symbols in the Ricci tensor that can be further reduced, even for a general metric tensor. (As explained earlier, every reduction of the Christoffel symbols is very useful.)

$$\begin{aligned} \left\{_{b}{}^{a}{}_{a}\right\} &= \frac{1}{2}g^{ad}\left(g_{bd,a} + g_{ad,b} - g_{ba,d}\right) \\ &= \frac{1}{2}g^{ad}g_{ab,b} + g^{ad}g_{b[d,a]} = \frac{1}{2}g^{ad}g_{ad,b} \ , \end{aligned} \tag{3.104}$$

as g^{ad} is symmetric while $g_{b[d,a]}$ is skew in "a" and "d". Now, from the definition of the inverse of a matrix it can be written as the matrix of cofactors divided by the determinant. Thus

$$\left\{_{b}{}^{a}{}_{a}\right\} = \frac{1}{2}\sum_{a,d=1}^{n}\mathrm{cofac}\left(g_{ad}\right)g_{ad,b} = \frac{1}{2g}g_{,b} = \left(\ln\sqrt{|g|}\right)_{,b} \ . \tag{3.105}$$

Hence the Ricci tensor can be written as

$$R_{bd} = \left\{_{b}{}^{a}{}_{d}\right\}_{,a} - \left(\ln\sqrt{|g|}\right)_{,bd} + \left(\ln\sqrt{|g|}\right)_{,e}\left\{_{b}{}^{e}{}_{d}\right\} - \left\{_{e}{}^{a}{}_{d}\right\}\left\{_{b}{}^{e}{}_{a}\right\} \ . \tag{3.106}$$

Since $\Delta_{\underline{dc}}\delta^{\underline{a}}_{\underline{b}} = 0$ we can split the $\delta^{\underline{a}}_{\underline{b}}$ as $\delta^{\underline{a}}_{b}\delta^{b}_{\underline{b}}$ to obtain an expression for the action of $\Delta_{\underline{dc}}$ on a covariant basis vector. It is clear that it is the negative of the result with the contravariant vector. From Eq.(3.97)

$$\delta^{\underline{a}}_{b}\Delta_{\underline{dc}}\delta^{b}_{\underline{b}} = -R^{\underline{a}}_{\underline{bcd}} \ . \tag{3.107}$$

This result can be used to evaluate the action of $\Delta_{\underline{pq}}$ on a general tensor, $\mathbf{T} \in \mathbf{V}_l^k$. The tensor can be expanded in terms of its basis vectors and components. Since the "$\Delta_{\underline{pq}}$" will not act on the scalar components and its action on basis vectors is given by Eqs.(3.87) and (97) we have the Ricci identities

$$\left.\begin{aligned}\Delta_{\underline{pq}} T^{\underline{a}\cdots\underline{c}}{}_{\underline{d}\cdots\underline{f}} =& R^{\underline{a}}{}_{\underline{x}\underline{pq}} T^{\underline{x}\cdots\underline{c}}{}_{\underline{d}\cdots\underline{f}} + \ldots + R^{\underline{c}}{}_{\underline{x}\underline{pq}} T^{\underline{a}\cdots\underline{x}}{}_{\underline{d}\cdots\underline{f}} \\ & - R^{\underline{x}}{}_{\underline{d}\underline{pq}} T^{\underline{a}\cdots\underline{c}}{}_{\underline{x}\cdots\underline{f}} - R^{\underline{x}}{}_{\underline{f}\underline{pq}} T^{\underline{a}\cdots\underline{c}}{}_{\underline{d}\cdots\underline{x}} \ .\end{aligned}\right\} \tag{3.108}$$

In particular, for vectors

$$\Delta_{\underline{pq}} A^{\underline{a}} = R^{\underline{a}}{}_{\underline{b}\underline{pq}} A^{\underline{b}} \ , \ \Delta_{\underline{pq}} X_{\underline{b}} = -R^{\underline{a}}{}_{\underline{b}\underline{pq}} X_{\underline{a}} \ . \tag{3.109}$$

There are n^2 components of $g_{\underline{ab}}$, n^2 components of $R_{\underline{ab}}$ and n^4 components of $R^{\underline{a}}_{\underline{bcd}}$. The symmetry property of $g_{\underline{ab}}$ reduces the components to $n(n+1)/2$ independent components. Now it is clear from Eq.(3.96) that $R_{\underline{ab}}$ is also symmetric. Since the calculational process here is very tedious, the reduction to $n\left(n+1\right)/2$ components is very important. In 3 dimensions we need to work out 6 instead of 9 components and in 4 dimensions 10 instead of 16. For $R^{\underline{a}}_{\underline{bcd}}$ in 2 dimensions there are 16, in 3 dimensions 81 and in 4 dimensions 256 components. It is obviously very necessary to look for symmetry properties of the Riemann tensor so that we reduce the number of components to be explicitly calculated.

From its definition in Eq.(3.87) it is obvious that the Riemann tensor is skew in its last two indices. Using Eq.(3.57) rewrite Eq.(3.97) as

$$R^{\underline{a}}{}_{\underline{bcd}} = -2\delta^{\underline{a}}_{\underline{b}} \nabla_{[\underline{c}} \nabla_{\underline{d}]} \left(\nabla_{\underline{b}} x^{b}\right) \ . \tag{3.110}$$

To find the next symmetry it is necessary to first prove a simple result regarding skew brackets. Notice that

$$\begin{aligned} A_{[[ab]c]} &= \frac{1}{2!}\left(A_{[abc]} - A_{[bac]}\right) = \frac{1}{2!}\left(A_{[abc]} - A_{[acb]}\right) \\ &= A_{[a[bc]]} = A_{[abc]} \ . \end{aligned} \tag{3.111}$$

Generally, in fact, skew brackets inside skew brackets can be removed and shifted at will. The same applies for symmetrizer brackets inside symmetrizer brackets. Using this identity we have

$$\begin{aligned} R^{\underline{a}}{}_{[\underline{bcd}]} &= -2\delta^{\underline{a}}_{\underline{b}} \nabla_{[[\underline{c}} \nabla_{\underline{d}]} \left(\nabla_{\underline{b}]} x^{b}\right) \\ &= -2\delta^{\underline{a}}_{\underline{b}} \nabla_{[\underline{c}} \Delta_{\underline{bd}]} x^{b} = 0 \ , \end{aligned} \tag{3.112}$$

since $\Delta_{\underline{bd}}$ annihilates scalars. Another symmetry property is obtained from the Ricci identities for the metric tensor

$$\begin{aligned} 0 &= \Delta_{\underline{ac}} g_{\underline{ab}} = -R^{\underline{x}}{}_{\underline{acd}} g_{\underline{xb}} - R^{\underline{x}}{}_{\underline{bcd}} g_{\underline{ax}} \\ &= -R_{\ \underline{bacd}} - R_{\underline{abcd}} \ . \end{aligned} \tag{3.113}$$

$$R_{\underline{abcd}} = -R_{\underline{bacd}} = -R_{\underline{abdc}} \ . \tag{3.114}$$

To prove the last symmetry property, using Eq.(3.102), we write

$$\begin{aligned} 2R_{\underline{abcd}} &= R_{\underline{abcd}} + R_{\underline{badc}} \\ &= -R_{\underline{acdb}} - R_{\underline{adbc}} - R_{\underline{bcad}} - R_{\underline{bdca}} \\ &= -R_{\underline{cabd}} - R_{\underline{cbda}} - R_{\underline{dacb}} - R_{\underline{dbac}} \\ &= R_{\underline{cdab}} + R_{\underline{dcba}} = 2R_{\underline{cdab}} \ , \end{aligned}$$

or

$$R_{\underline{abcd}} = R_{\underline{cdab}} \ . \tag{3.115}$$

The symmetry properties given by Eq.(3.94) reduce the number of independent, non-zero components to $[n\,(n-1)\,/2]^2$. Due to Eq.(3.105) this number is further reduced. Regarding the first and second pairs of indices as compound indices, say A and B with the range $n\,(n-1)\,/2$ each, R_{abcd} can be regarded as the symmetric second rank tensor, R_{AB}, of dimension $n(n-1)/2$. Thus, the number of independent non-zero components is now

$$[n(n-1)/2][1+n(n-1/2)] = n(n-1)\left(n^2-n+2\right)/8 \ .$$

Eq.(3.102) does not give any new constraint for $n < 4$. Generally, for $n \geq 4$, the number of extra constraints are ${}^nC_4 = n!/\left[(n-4)!4!\right]$. Therefore the total number of independent non-zero components is

$$\frac{n\,(n-1)\left(n^2-n+2\right)}{8} - \frac{n\,(n-1)\,(n-2)\,(n-3)}{4!} = \frac{n^2\left(n^2-1\right)}{12} \ . \tag{3.116}$$

For $n = 2$ this is 1. Thus there is only one independent component of the Riemann tensor and hence of the Ricci tensor. (Hence for $n = 2$ the formula for the number of independent components of the Ricci tensor is invalid.) For $n = 3$ Eq.(3.106) gives 6 and for $n = 4$ it gives 20. For $n = 5$, notice that R_{ab} has 15 and R_{abcd} has 50 independent components! Even with the symmetries the number of independent components blows up rapidly for larger n.

There are additional (differential) identities satisfied by the Riemann tensor. These must be taken into account when trying to assign a physical role to the Riemann, or Ricci, tensors. These identities will put constraints on the physical equations these tensors are required to satisfy. They are obtained by using the Ricci identities for a second rank tensor in the form

$$2B^a{}_{b;[d;c]} = R^a{}_{pcd}B^p{}_b - R^p{}_{bcd}B^a{}_p \ . \tag{3.117}$$

Skewing this equation over the three indices b, c and d and using Eq.(3.102) the second term disappears and we obtain

$$2B^a{}_{[b;d;c]} = R^a{}_{p[cd}B^p{}_{b]} \ . \tag{3.118}$$

Consider the case when $B^a{}_b = A^a{}_{;b}$. Then Eq.(3.108) becomes

$$2A^a{}_{;[b;d;c]} = R^a{}_{p[cd}A^p{}_{;b]} = R^a{}_{p[bc}A^p{}_{;d]} \ , \tag{3.119}$$

as cyclic permutations of an odd number of elements are even, and even permutations of indices inside a skew bracket leave the expression un-altered. Now using Eqs.(3.102) and (3.99)

$$\begin{aligned} 2A^a{}_{;[b;d;c]} &= 2A^a{}_{;[c;b;d]} = 2A^a{}_{;[d;c;b]} \\ &= \left(R^a{}_{p[bc}A^p\right)_{;d]} \\ &= R^a{}_{p[bc;d]}A^p + R^a{}_{p[bc}A^p{}_{;d]} \ . \end{aligned} \tag{3.120}$$

Comparing Eqs.(3.109) and (3.110) we see that

$$R^a{}_{p[bc;d]} = 0 = \frac{1}{3}\left(R^a{}_{pbc;d} + R^a{}_{pcd;b} + R^a{}_{pdb;c}\right) \ , \tag{3.121}$$

as the equations hold for all A^a. These equations are known as the *Bianchi identities*. According to Pias [12] Einstein's ignorance of these equations was responsible for his many failed attempts at finding a generalization of the Special Theory of Relativity.

The Riemann tensor is useful for determining whether a singularity is essential or coordinate. If the curvature becomes infinite the singularity is essential. The problem is that $R^a{}_{bcd}$, given by Eq.(3.92), is expressed in coordinate terms. Thus the malaise of the coordinate system will infect the Riemann tensor components. However, scalar quantities are invariant under coordinate transformations. Constructing scalars from the Riemann tensor we could check if they become infinite somewhere. Of course, infinitely many scalars could be constructed from the Riemann tensor. However, symmetry considerations can be used to show that there are only finitely many independent scalars. All others can be expressed in terms of these. Thus we only need to construct the simplest scalars. Some of these are

$$\mathcal{R}_1 := g^{ab}R_{ab} = R \ , \tag{3.122}$$

$$\mathcal{R}_2 := R^{ab}{}_{cd}R^{cd}{}_{ab} \ , \tag{3.123}$$

$$\mathcal{R}_3 := R^{ab}{}_{cd}R^{cd}{}_{ef}R^{ef}{}_{ab} \ , \tag{3.124}$$

$$\mathcal{R}_4 := R^{ab}R_{ab} \ , ... \ . \tag{3.125}$$

If the independent *curvature invariants*, defined above, are all finite the singularity (if any) will be coordinate. If any of them is infinite the singularity will be essential.

Example 1 : Euclidean 2-dimensional space in plane polar coordinates. Using Eqs.(3.87) we have, from Eq.(3.102)

$$\begin{aligned} R^1_{212} &= \left\{_2{}^1{}_2\right\}_{,1} - \left\{_2{}^1{}_1\right\}_{,2} + \left\{_a{}^1{}_1\right\}\left\{_2{}^a{}_2\right\} - \left\{_a{}^1{}_2\right\}\left\{_1{}^a{}_2\right\} \\ &= (-r)_{,1} - \left\{_1{}^1{}_2\right\}\left\{_1{}^1{}_2\right\} - \left\{_2{}^1{}_2\right\}\left\{_1{}^2{}_2\right\} \\ &= -1 - (-r)\,(1/r) = 0 \ . \end{aligned}$$

Therefore

$$\begin{aligned} R_{1212} &= g_{11}R^1_{212} + g_{12}R^2_{212} = R^1_{212} = 0 \ . \\ &= -R_{2112} = -R_{1221} = R_{2121} \ . \\ R_{1111} &= R_{1112} = R_{1121} = R_{1122} = R_{1211} = R_{2111} \\ &= R_{2211} = R_{2212} = R_{2221} = R_{2222} = R_{1222} = R_{2122} = 0 \ . \end{aligned}$$

Since there is only one independent component here, which is zero, the Ricci tensor and Ricci scalar are zero. Thus the space is flat, as we know. Hence the singularity at $r = 0$ is a coordinate singularity.

Example 2 : The surface of a sphere of radius a. From Eqs.(3.89) and (3.102)

$$\begin{aligned} R^1_{212} &= \left\{_2{}^1{}_2\right\}_{,1} - \left\{_2{}^1{}_1\right\}_{,2} + \left\{_a{}^1{}_1\right\}\left\{_2{}^a{}_2\right\} - \left\{_a{}^1{}_2\right\}\left\{_1{}^a{}_2\right\} \\ &= (-\sin\theta\cos\theta)_{,1} - (-\sin\theta\cos\theta)\,(\cot\theta) \\ &= \sin^2\theta - \cos^2\theta + \cos^2\theta = \sin^2\theta \ . \end{aligned}$$

Therefore

$$R_{1212} = -R_{1221} = -R_{2112} = R_{2121} = a^2\sin^2\theta \ , \ R_{abcd} = 0 \text{ otherwise.}$$

Using Eq.(3.96) with $g = a^4 \sin^2\theta$, we have

$$\begin{aligned}
R_{11} &= \{_1\ ^a\ _1\}_{,a} - \left[\ln\left(a^2\sin\theta\right)\right]_{,11} + \left[\ln\left(a^2\sin\theta\right)\right]_{,a}\{_1\ ^a\ _1\} \\
&\quad - \{_b\ ^a\ _1\}\{_a\ ^b\ _1\} \\
&= -(\cot\theta)_{,1} - \{_2\ ^2\ _1\}^2 = \csc^2\theta - \cot^2\theta = 1\ , \\
R_{12} &= \{_1\ ^a\ _2\}_{,a} - \left[\ln\left(a^2\sin\theta\right)\right]_{,12} + \left[\ln\left(a^2\sin\theta\right)\right]_{,a}\{_1\ ^a\ _2\} \\
&\quad - \{_b\ ^a\ _1\}\{_a\ ^b\ _2\} \\
&= \{_1\ ^1\ _2\}_{,1} + \{_1\ ^2\ _2\}_{,2} - (\cot\theta)_{,2} + \left[\ln\left(a^2\sin\theta\right)\right]_{,1}\{_1\ ^1\ _2\} \\
&\quad + \left[\ln\left(a^2\sin\theta\right)\right]_{,2}\{_1\ ^2\ _2\} - \{_1\ ^1\ _1\}\{_1\ ^1\ _2\} - \{_2\ ^1\ _1\}\{_1\ ^2\ _2\} \\
&\quad - \{_1\ ^2\ _1\}\{_2\ ^1\ _2\} - \{_2\ ^2\ _1\}\{_2\ ^2\ _2\} = 0 = R_{21}\ , \\
R_{22} &= \{_2\ ^a\ _2\}_{,a} - \left[\ln\left(a^2\sin\theta\right)\right]_{,22} + \left[\ln\left(a^2\sin\theta\right)\right]_{,a}\{_2\ ^a\ _2\} \\
&\quad - \{_b\ ^a\ _2\}\{_a\ ^b\ _2\} \\
&= \{_2\ ^1\ _2\}_{,1} + \cot\theta\{_2\ ^1\ _2\} - \{_2\ ^1\ _2\}\{_1\ ^2\ _2\} - \{_1\ ^2\ _2\}\{_2\ ^1\ _2\} \\
&= \sin^2\theta - \cos^2\theta - \cos^2\theta - 2(-\sin\theta\cos\theta)\cot\theta = \sin^2\theta\ .
\end{aligned}$$

We could also have obtained the Ricci tensor components from the Riemann tensor components as

$$\begin{aligned}
R_{11} &= R^1{}_{111} + R^2{}_{121} = g^{22}R_{\ 2121} = a^2\sin^2\theta/a^2\sin^2\theta = 1\ , \\
R_{12} &= R^1{}_{112} + R^2{}_{122} = 0 = R_{21}\ , \\
R_{22} &= R^1{}_{212} + R^2{}_{222} = \sin^2\theta\ .
\end{aligned}$$

Therefore

$$R = g^{11}R_{11} + g^{12}R_{12} + g^{21}R_{21} + g^{22}R_{22} = 1/a^2 + \sin^2\theta/a^2\sin^2\theta = 2/a^2\ .$$

Notice that $R = 2K$. Since the curvature scalar is finite the singularities at $\theta = 0, \pi$ are coordinate singularities. Suppose we had a 3-sphere instead. We would now have 3 angles, say (θ, ϕ, ψ). Choosing $\psi = 0$ would yield the 2-sphere we just worked out. Since there is perfect symmetry between the angles, we could choose any one to be zero. Thus the total curvature would be $3\times 2 = 3!$. Proceeding the same way, we see that for an n-sphere, $R = n!K$.

Example 3 : Euclidean 3-dimensional space in spherical polar coordinates. Using Eqs.(3.91) and (3.102) for the Riemann tensor and Eq.(3.96) for the Ricci

tensor with $g = r^4 \sin^2\theta$, we have

$$\begin{aligned}
R^1_{212} &= \left\{_2\ ^1\ _2\right\}_{,1} - \left\{_2\ ^1\ _1\right\}_{,2} + \left\{_a\ ^1\ _1\right\}\left\{_2\ ^a\ _2\right\} - \left\{_a\ ^1\ _2\right\}\left\{_1\ ^a\ _2\right\} \\
&= (-r)_{,1} - (-r)\,(1/r) = -1 + 1 = 0\ , \\
R^1_{213} &= \left\{_2\ ^1\ _3\right\}_{,1} - \left\{_2\ ^1\ _1\right\}_{,3} + \left\{_a\ ^1\ _1\right\}\left\{_2\ ^a\ _3\right\} - \left\{_a\ ^1\ _3\right\}\left\{_1\ ^a\ _2\right\} \\
&= -\left\{_1\ ^1\ _3\right\}\left\{_1\ ^1\ _2\right\} - \left\{_2\ ^1\ _3\right\}\left\{_1\ ^2\ _2\right\} - \left\{_3\ ^1\ _3\right\}\left\{_1\ ^3\ _2\right\} = 0\ , \\
R^1_{223} &= \left\{_2\ ^1\ _3\right\}_{,1} - \left\{_2\ ^1\ _1\right\}_{,3} + \left\{_a\ ^1\ _2\right\}\left\{_2\ ^a\ _3\right\} - \left\{_a\ ^1\ _3\right\}\left\{_2\ ^a\ _2\right\} \\
&= \left\{_2\ ^1\ _2\right\}\left\{_2\ ^2\ _3\right\} + \left\{_3\ ^1\ _2\right\}\left\{_2\ ^3\ _3\right\} - \left\{_1\ ^1\ _3\right\}\left\{_2\ ^1\ _2\right\} \\
&\quad - \left\{_3\ ^1\ _3\right\}\left\{_2\ ^3\ _2\right\} = 0\ , \\
R^1_{313} &= \left\{_3\ ^1\ _3\right\}_{,1} - \left\{_3\ ^1\ _1\right\}_{,3} + \left\{_a\ ^1\ _1\right\}\left\{_3\ ^a\ _3\right\} - \left\{_a\ ^1\ _3\right\}\left\{_1\ ^a\ _3\right\} \\
&= \left(-r\sin^2\theta\right)_{,1} - \left\{_3\ ^1\ _3\right\}\left\{_1\ ^3\ _3\right\} = -\sin^2\theta - \left(-r\sin^2\theta\right)/r = 0\ , \\
R^1_{323} &= \left\{_3\ ^1\ _3\right\}_{,2} - \left\{_3\ ^1\ _2\right\}_{,3} + \left\{_a\ ^1\ _2\right\}\left\{_3\ ^a\ _3\right\} - \left\{_a\ ^1\ _3\right\}\left\{_2\ ^a\ _3\right\} \\
&= \left(-r\sin^2\theta\right)_{,2} + \left\{_2\ ^1\ _2\right\}\left\{_3\ ^2\ _3\right\} - \left\{_3\ ^1\ _3\right\}\left\{_2\ ^3\ _3\right\} \\
&= -2r\sin\theta\cos\theta - r\,(-\sin\theta\cos\theta) - \left(-r\sin^2\theta\right)\cot\theta \\
&= -2r\sin\theta\cos\theta + r\sin\theta\cos\theta + r\sin\theta\cos\theta = 0\ , \\
R^2_{323} &= \left\{_3\ ^2\ _3\right\}_{,2} - \left\{_3\ ^2\ _2\right\}_{,3} + \left\{_a\ ^2\ _2\right\}\left\{_3\ ^a\ _3\right\} - \left\{_a\ ^2\ _3\right\}\left\{_2\ ^a\ _3\right\} \\
&= (-\sin\theta\cos\theta)_{,2} + \left\{_1\ ^2\ _2\right\}\left\{_3\ ^1\ _3\right\} - \left\{_3\ ^2\ _3\right\}\left\{_2\ ^3\ _3\right\} \\
&= \sin^2\theta - \cos^2\theta + \left(-r\sin^2\theta\right)/r - (-\sin\theta\cos\theta)\cot\theta \\
&= \sin^2\theta - \cos^2\theta - \sin^2\theta + \cos^2\theta = 0\ .
\end{aligned}$$

We have obtained all 6 components that are independent. Notice that we started with the lowest value of the indices to the left and increased indices to the right in such a way as to avoid getting trivially zero elements, e.g. $R^1{}_{211}$. Also we did not work out $R^1{}_{312}$ as Eqs.(3.112) yield $R_{1312} + R_{1123} + R_{1231} = 0$. Therefore, it is not independent.

$$\begin{aligned}
R_{11} &= \left\{_1\ ^a\ _1\right\}_{,a} - \left[\ln\left(r^2\sin\theta\right)\right]_{,11} + \left[\ln\left(r^2\sin\theta\right)\right]_{,a}\left\{_1\ ^a\ _1\right\} \\
&\quad - \left\{_b\ ^a\ _1\right\}\left\{_a\ ^b\ _1\right\} \\
&= -(2/r)_{,1} - \left\{_2\ ^2\ _1\right\}^2 - \left\{_3\ ^3\ _1\right\}^2 = 2/r^2 - 1/r^2 - 1/r^2 = 0\ , \\
R_{12} &= \left\{_1\ ^a\ _2\right\}_{,a} - \left[\ln\left(r^2\sin\theta\right)\right]_{,12} + \left[\ln\left(r^2\sin\theta\right)\right]_{,a}\left\{_1\ ^a\ _2\right\} \\
&\quad - \left\{_b\ ^a\ _1\right\}\left\{_a\ ^b\ _2\right\}\ .
\end{aligned}$$

Evaluating only the non-zero Christoffel symbols, we get

$$R_{12} = \cot\theta/r - \cot\theta/r = 0\ .$$

Similarly,

$$\begin{aligned}
R_{13} &= \{_1\ ^a\ _3\}_{,a} - \left[\ln\left(r^2\sin\theta\right)\right]_{,13} + \left[\ln\left(r^2\sin\theta\right)\right]_{,a}\{_1\ ^a\ _3\} \\
&\quad - \{_b\ ^a\ _1\}\{_a\ ^b\ _3\} \\
&= \left[\ln\left(r^2\sin\theta\right)\right]_{,a}\{_1\ ^a\ _3\} - \{_b\ ^a\ _1\}\{_a\ ^b\ _3\} = 0\ , \\
R_{22} &= \{_2\ ^a\ _1\}_{,a} - \left[\ln\left(r^2\sin\theta\right)\right]_{,22} + \left[\ln\left(r^2\sin\theta\right)\right]_{,a}\{_2\ ^a\ _2\} \\
&\quad - \{_b\ ^a\ _2\}\{_a\ ^b\ _2\} \\
&= \{_2\ ^1\ _2\}_{,1} - (\cot\theta)_{,2} + \left[\ln\left(r^2\sin\theta\right)\right]_{,1}\{_2\ ^1\ _2\} - 2\{_2\ ^1\ _2\}\{_1\ ^2\ _2\} \\
&\quad - \{_3\ ^3\ _2\}^2 = 0\ , \\
R_{23} &= \{_3\ ^a\ _3\}_{,a} - \left[\ln\left(r^2\sin\theta\right)\right]_{,23} + \left[\ln\left(r^2\sin\theta\right)\right]_{,a}\{_2\ ^a\ _3\} \\
&\quad - \{_b\ ^a\ _2\}\{_a\ ^b\ _3\} = 0\ , \\
R_{33} &= \{_3\ ^a\ _3\}_{,a} - \left[\ln\left(r^2\sin\theta\right)\right]_{,33} + \left[\ln\left(r^2\sin\theta\right)\right]_{,a}\{_3\ ^a\ _3\} \\
&\quad - \{_b\ ^a\ _3\}\{_a\ ^b\ _3\} \\
&= -\sin^2\theta + \left(\sin^2\theta - \cos^2\theta\right) - 2\sin^2\theta - \cos^2\theta + 2\sin^2\theta \\
&\quad + 2\cos^2\theta = 0\ .
\end{aligned}$$

Of course, since $R^a_{\ bcd} = 0$, we already knew that $R_{ab} = 0$ and hence $R = 0$. Since the space is flat this was known anyhow. The purpose of doing the calculations explicitly was to demonstrate the procedure used in full detail. Clearly, since $\mathcal{R}_1$, $\mathcal{R}_2$ are zero there can be no essential singularity. Hence the singularities at $r = 0$, $\theta = 0$, π are all coordinate singularities.

3.7 Curves in Manifolds

The generalized theory of Relativity attempts to do dynamics relativistically. As such it has to generalize the Newtonian laws of dynamics. They may be collectively stated as "Particles move in a straight line at constant speed unless an external force acts upon them, in which case they suffer an acceleration, leading to a change in momentum proportional to the force". The generalisation is two-fold. On the one hand we now have a 4-dimensional *spacetime* instead of a 3-dimensional space and 1-dimensional time. Whereas in Newtonian Physics one thinks of things *happening* — a continual process of *becoming* — in spacetime things just *are* — an eternal state of *being*. The former is the Newtonian "particle" world-view while the latter is the Maxwellian, "field", world-view. (It is worth noting that in the Newtonian view the whole history is *implicit* in the equations, in the Maxwellian view it is *explicit*. However, in the latter the knowledge of the entire space at an instant is required.)Here we are concerned with the other generalisation that the spacetime need not be flat. As such there may be no such thing as a "straight line", like for example, there is no straight line available on the surface of the Earth. The question is, then, "What curve takes the place of the straight line?" To be able to answer this question we need a rigorous definition, and consequent discussion, of curves in manifolds.

A *curve*, γ, is a mapping from the unit interval, $\mathrm{I} = [0,\ 1]$, into the manifold, $\gamma : \mathrm{I} \to \mathcal{M}_n$, see Fig. 3.9. $\gamma(0)$ is called the *initial point* and $\gamma(1)$ the *final*, or *end*, *point* of the curve. The curve is parameterized by some $\lambda \in \mathbb{R}$. If

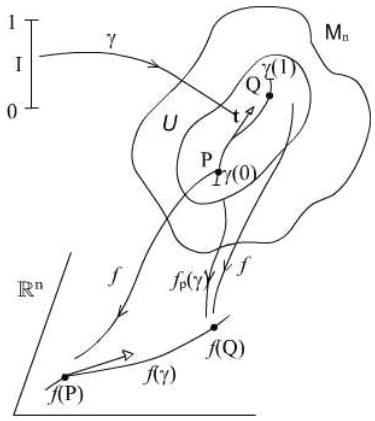


Figure 3.9: A curve $\gamma : \mathrm{I} \overset{\text{into}}{\to} \mathcal{M}_n$ in a coordinate chart $(\mathcal{U}, f)$ and there is a coordinatization, $f : P \to f(P)$ and $f : Q \to f(Q)$. A new function $f_P(\lambda)$ is defined by the requirement that if Q is at the parameter value λ, $f_P(\gamma) = f(Q)$. The tangent to the image of curve, $f(\gamma)$, is the image of the tangent to the curve, $\mathbf{t}(f)$.

$\gamma(0) = \gamma(1)$, γ is a *closed curve*. If $\gamma(\lambda_1) = \gamma(\lambda_2)$ for some $\lambda_2 \neq \lambda_1$ the curve is *self-intersecting*. To be able to use calculus on the manifold away from just the one point, P, we need to construct a differentiable function of λ such that in the coordinate system $x^a(\lambda) = f(P_\lambda)$ describes the curve. Let $P_{\lambda_0} = P$, $P_\lambda = Q$ and $\lambda \neq \lambda_0$. We define $f_P(\lambda) = x^a(\lambda)$ by

$$f_P(\lambda_0) = f(P) = x^a(\lambda_0) \ , \ f_P(\lambda) = f(Q) = x^a(\lambda) \ . \tag{3.126}$$

Henceforth we shall drop the subscript P. Our procedure will be to use calculus in the coordinate system and pull the result back onto the manifold.

The tangent vector to the curve, in the manifold is the inverse image of the tangent vector to the image of the curve in the coordinate system, see Fig. 3.9. Thus

$$\mathbf{t}(f) = \left.\frac{dx^a}{d\lambda}\right|_{\lambda=\lambda_0} = \lim_{\lambda\to\lambda_0} \frac{f(\lambda) - f(\lambda_0)}{\lambda - \lambda_0} = t^a \ . \tag{3.127}$$

Thus, acting on a scalar function it is the derivation

$$\mathbf{t} = t^a \nabla_a = d/d\lambda \ . \tag{3.128}$$

If $\mathbf{t}$ is the unit tangent vector λ will be the affine parameter, s,

$$t^a t^b g_{ab} = \dot{x}^a \dot{x}^b g_{ab} = \frac{dx^a}{ds}\frac{dx^b}{ds} g_{ab} = 1 \ . \tag{3.129}$$

Therefore

$$ds^2 = g_{ab} dx^a dx^b \tag{3.130}$$

is the arc-length, or the metric, in keeping with Eq.(2.44).

We want to obtain the derivative of a tensor quantity on the manifold, so that it is independent of the coordinate system (and hence of the choice of observer). The problem with this attempt is that the manifold is just a collection of *points*

and does not involve numbers. Since tensor quantities on the manifold are "compounds" of operators moving points to points on the tangent space, *there are no numbers*. However, we can only use calculus with numbers. That is why we required that the manifold can be locally mapped to collections of numbers in the first place. So, we are on the horns of a dilemma: on the one hand we need to use basis vectors; but on the other hand we want to dispense with them while working on the manifold. There are, then, two ways to go. One is to ignore the effect of the coordinatization on the tensor and just apply the derivation to the tensor. This is called the *intrinsic derivative* and denoted by $D_{\mathbf{t}}$, so that $D_{\mathbf{t}}\mathbf{A}(f) = \mathbf{t}[\mathbf{A}(f)]$. In other words we apply the derivation to the image of the tensor in the coordinate system. Written in component indices we use the partial derivative for the components of the tensor and then incorporate the derivative of the basis vectors by adding in terms involving the Christoffel symbols. Taking the coordinatization as implicit, in the abstract index notation

$$D_{\underline{t}} A^{\underline{a}\cdots\underline{c}}{}_{\underline{d}\cdots\underline{f}} = t^{\underline{p}} \nabla_{\underline{p}} A^{\underline{a}\cdots\underline{c}}{}_{\underline{d}\cdots\underline{f}} \ , \tag{3.131}$$

with components $t^p A^{a\cdots c}{}_{d\cdots f;p}$. The other way is to pull out the effect of the coordinatization of the tensor and compute the effect of the derivation on the tensor *in the manifold*. This is called the *Lie derivative*. Acting on a scalar it must clearly be the same as the intrinsic derivative. Further, since $\mathbf{p}(f)$ is a scalar and the Lie derivative must satisfy the Leibnitz rule, we have

$$\begin{aligned} \mathbf{t}[\mathbf{p}(f)] &= \mathcal{L}_{\mathbf{t}}[\mathbf{p}(f)] = (\mathcal{L}_{\mathbf{t}}\mathbf{p})(f) + \mathbf{p}(\mathcal{L}_{\mathbf{t}} f) \\ &= (\mathcal{L}_{\mathbf{t}}\mathbf{p})(f) + \mathbf{p}[\mathbf{t}(f)] \ . \end{aligned} \tag{3.132}$$

Therefore

$$(\mathcal{L}_{\mathbf{t}}\mathbf{p})(f) = \mathbf{t}[\mathbf{p}(f)] - \mathbf{p}[\mathbf{t}(f)] \tag{3.133}$$

is itself a derivation. Writing $\mathcal{L}_{\mathbf{t}}\mathbf{p}$ as $\mathbf{q}$ in the abstract index notation

$$\begin{aligned} \mathbf{q}(f) &= q^a \nabla_a f = t^b \nabla_b [p^a \nabla_a f] - p^b \nabla_b [t^a \nabla_a f] \\ &= \left(t^b \nabla_b p^a - p^b \nabla_b t^a\right) \nabla_a f + t^b p^a \Delta_{ab} f \ . \end{aligned}$$

Hence, in a torsion-free space

$$q^{\underline{a}} := (\mathcal{L}_{\mathbf{t}}\mathbf{p})^{\underline{a}} = t^{\underline{b}} \nabla_{\underline{b}} p^{\underline{a}} - p^{\underline{b}} \nabla_{\underline{b}} t^{\underline{a}} \ . \tag{3.134}$$

The corresponding component form is

$$\begin{aligned} q^a := (\mathcal{L}_{\mathbf{t}}\mathbf{p})^a &= t^b p^a_{\ ;b} - p^b t^a_{\ ;b} \\ &= t^b p^a_{\ ,b} + \{_b{}^a{}_c\} t^b p^c - p^b t^a_{\ ,b} - \{_b{}^a{}_c\} p^b t^c \\ &= t^b p^a_{\ ,b} - p^b t^a_{\ ,b} \ . \end{aligned} \tag{3.135}$$

To be able to write down the expression for the Lie derivative of a tensor we need to evaluate the Lie derivative of a covariant vector. As usual, consider

$$\begin{aligned} \mathcal{L}_{\mathbf{t}}(p^a r_a) &= (\mathcal{L}_{\mathbf{t}}\mathbf{p})^a r_a + p^a (\mathcal{L}_{\mathbf{t}}\mathbf{r})_a = t^b p^a_{\ ,b} r_a - p^b t^a_{\ ,b} r_a + p^a (\mathcal{L}_{\mathbf{t}}\mathbf{r})_a \\ &= t^b p^a_{\ ,b} r_a + p^a t^b r_{a,b} \ . \end{aligned}$$

Thus

$$(\mathcal{L}_{\mathbf{t}}\mathbf{r})_a = t^b r_{a,b} + r_b t^b_{\ ,a} \ , \tag{3.136}$$

having interchanged indices "a" and "b" in one term so that p^a can be taken as common. Thus, for a general tensor with components $A^{a\ldots c}{}_{d\ldots f}$

$$\mathcal{L}_{\mathbf{t}}\mathbf{A} = \left.\frac{d\mathbf{A}}{d\lambda}\right|_{\lambda=\lambda_0} , \tag{3.137}$$

we have the components

$$\left.\begin{aligned}(\mathcal{L}_{\mathbf{t}}A)^{a\ldots c}{}_{d\ldots f} =& t^p A^{a\ldots c}{}_{d\ldots f,p} - A^{p\ldots c}{}_{d\ldots f}\, t^a_{,p} - \ldots - A^{a\ldots p}{}_{d\ldots f}\, t^c_{,p} \\ &+ A^{a\ldots c}{}_{p\ldots f}\, t^p_{,d} + \ldots + A^{a\ldots c}{}_{d\ldots p}\, t^p_{,f} \ .\end{aligned}\right\} \tag{3.138}$$

Notice that there is no need to compute Christoffel symbols in working out the Lie derivative as there is for the intrinsic derivative. *This* is what makes working on the manifold much easier than in the coordinate system. To see the significance of this fact more clearly, think of the spheroid as $x^2/a^2 + y^2/a^2 + z^2/b^2 = 1$, and use the usual parameterization for the sphere using spherical coordinates, but using the 'radius' as a for x and y, but b for z. When working out Lie derivatives *the sphere and the spheroid will seem the same in terms of* θ *and* ϕ. It is only in the (x, y, z)-coordinate system that the distortion is seen. The intrinsic derivative will look very different. The difference becomes horrendous for the ellipsoid, where the off-diagonal metric coefficient is non-zero, as we now have to evaluate the Christoffel symbols without using the simplifying formula (3.85). It would be useful to actually compute the metric tensor for the ellipsoid and try to work out an intrinsic derivative there for some choices of $\mathbf{t}$ and $\mathbf{p}$. The Lie derivative on the ellipsoid will be the same as for the sphere.

So far our tensors have been defined at one point on the manifold. We would now like to use Taylor's theorem to evaluate them at other points along some curve with unit tangent vector $\mathbf{t}$. To be able to do so we need to state the theorem for functions of several variables. Let us start in a flat space of two dimensions. Thus we consider $f(a+h, b+k)$ and first look at the Taylor expansion in the y-direction at the higher value of x.

$$\begin{aligned} f\left(a+h, b+k\right) &= f\left(a+h, b\right) + k f_y\left(a+h, b\right) + \frac{1}{2!}k^2 f_{yy}\left(a+h, b\right) \\ &\quad + \frac{1}{3!}k^3 f_{yyy}\left(a+h, b\right) + \ldots\ldots \end{aligned}$$

Now use the Taylor series in the x-direction for each term above. Then

$$\begin{aligned} f(a+h,b+k) &= f(a,b)+hf_x(a,b)+\frac{1}{2!}h^2f_{xx}(a,b)+\frac{1}{3!}h^3f_{xxx}(a,b)+... \\ &\quad +kf_y(a,b)+hkf_{xy}(a,b)+\frac{1}{2!}h^2kf_{xxy}(a,b)+... \\ &\quad +\frac{1}{2!}k^2f_{yy}(a,b)+\frac{1}{2!}hk^2f_{xyy}(a,b)+... \\ &\quad +\frac{1}{3!}k^3f_{yyy}(a,b)+... \\ &= f(a,b)+\left(h\frac{\partial}{\partial x}+k\frac{\partial}{\partial y}\right)f(a,b) \\ &\quad +\frac{1}{2!}\left(h\frac{\partial}{\partial x}+k\frac{\partial}{\partial y}\right)^2 f(a,b) \\ &\quad +\frac{1}{3!}\left(h\frac{\partial}{\partial x}+k\frac{\partial}{\partial y}\right)^3 f(a,b)+... \\ &= \left[\exp\left(h\frac{\partial}{\partial x}+k\frac{\partial}{\partial y}\right)\right]f(x,y)\Big|_{x=a,y=b} . \end{aligned} \tag{3.139}$$

Generally, for a scalar function of n variables, displaced from the position a^i (at $\lambda=\lambda_0$) by an amount t^i(at $\lambda=\lambda_0$), we have

$$f\left(a^i+t^i\right)=\left[\exp(t^i\frac{\partial}{\partial x^i})f\left(x^j\right)\right]\Big|_{x^j=a^j} . \tag{3.140}$$

Since there are two inequivalent derivative operators (in general) there will be two distinct results on applying Taylor's theorem. The two coincide when acting on a scalar function. If we apply Taylor's theorem to the coordinatized tensor we will displace the tensor parallely in the coordinate system. This is called *parallel transport*

$$\mathbf{A}_{\parallel}(\mathbf{a}+\mathbf{t})=\left[\exp D_{\mathbf{t}}\right]\mathbf{A}(\mathbf{x})\big|_{\mathbf{x}=\mathbf{a}} . \tag{3.141}$$

If, instead, we displace along the curve on the manifold we must use the Lie derivatve. This is called *Lie transport*

$$\mathbf{A}_{\mathcal{L}}(\mathbf{a}+\mathbf{t})=\left[\exp \mathcal{L}_{\mathbf{t}}\right]\mathbf{A}(\mathbf{x})\big|_{\mathbf{x}=\mathbf{a}} . \tag{3.142}$$

The difference between the two is depicted in Fig. 3.10. Clearly Lie transport is more relevant for the geometry of the manifold in itself, without reference to the coordinate system.

In a flat space the straight line is the curve which maintains the same direction. The curved space generalisation is the curve which retains the same tangent vector. This is the "straightest available line". Thus we require that the tangent vector be parallely transported along the curve. (Notice that, by definition, every curve has a tangent vector that is Lie transported along the curve.)

$$\mathbf{t}_{\parallel}(\mathbf{a}+\mathbf{t})=\left[\exp D_{\mathbf{t}}\right]\mathbf{t}(\mathbf{x})\big|_{\mathbf{x}=\mathbf{a}}=\mathbf{t}(\mathbf{a}) . \tag{3.143}$$

For this equation to hold for all $\mathbf{a}$ on the curve we must have

$$D_{\mathbf{t}}\mathbf{t}=\mathbf{t}(\mathbf{t})=0 , \tag{3.144}$$

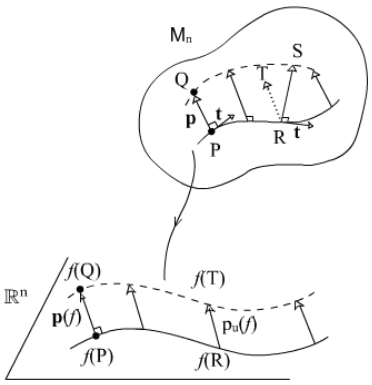


Figure 3.10: The difference between Lie and parallel transport of a derivation $\mathbf{p}$, originally defined by $\mathbf{p} : P \to Q$. Taking the image in a coordinate system and then shifting $\mathbf{p}_{\|}(f)$ by the inverse coordinatization gives $\mathbf{p}_{\|}$. Supposing $\mathbf{p}$ is defined to be $\perp \mathbf{t}$ the tangent to the curve on the manifold, $\mathbf{p}_{\mathcal{L}}$ will be $\perp \mathbf{t}$ there. The dotted vector $\overrightarrow{RT}$ gives $\mathbf{p}_{\|}$ and $\overrightarrow{RS}$ gives $\mathbf{p}_{\mathcal{L}}$.

which, in Penrose's notation and the component index notation is

$$t^{\underline{b}}\nabla_{\underline{b}}t^{\underline{a}} = 0 \quad \text{or} \quad t^b t^a{}_{;b} = 0 \ . \tag{3.145}$$

This is called the *geodesic equation* and its solution is called a *geodesic*. Taking $\mathbf{t}$ to be a unit vector, $t^\alpha = \dot{x}^\alpha$. Thus we can write Eq.(3.146) as

$$\begin{aligned} t^b t^a{}_{;b} &= t^b t^a{}_{,b} + \{_b{}^a{}_c\}\, t^b t^c \\ &= \ddot{x}^a + \{_b{}^a{}_c\}\, \dot{x}^b \dot{x}^c = 0 \ . \end{aligned} \tag{3.146}$$

In flat space the shortest path between two points is a straight line. Let us determine the shortest path between two points using the EL equations. The arc length between the points P and Q is given by

$$\begin{aligned} s_{PQ} &= \int_P^Q ds = \int_P^Q 1 ds \\ &= \int_P^Q g_{ab}\left(x^c\right) \dot{x}^\alpha \dot{x}^b ds = \int_P^Q \mathcal{L}\left[x^a, \dot{x}^\alpha\right] ds \ , \end{aligned} \tag{3.147}$$

from Eq.(3.120). It may seem strange to treat the constant "1" as the Lagrangian, $\mathcal{L}$. The point is that it can be a *functional* but not a *function*. The variation of one part can exactly cancel the variation of the other part and still allow for non-trivial functions $g_{ab}(x^c)$. The EL equations, given by Eq.(1.10), for the above Lagrangian are

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial x^c} &= g_{ab,c}\dot{x}^a \dot{x}^b \ , \\ \frac{\partial \mathcal{L}}{\partial \dot{x}^c} &= g_{cb}\dot{x}^b + g_{ac}\dot{x}^a \ . \end{aligned}$$

Therefore, as the dependence on x^c is only through the metric tensor

$$\begin{aligned}\frac{d}{ds}\left(\frac{\partial\mathcal{L}}{\partial\dot{x}^c}\right) &= \left(\frac{dx^d}{ds}\frac{\partial}{\partial x^d}g_{ac}\right)\dot{x}^a + \left(\frac{dx^d}{ds}\frac{\partial}{\partial x^d}g_{cb}\dot{x}^b\right) \\ &\quad + g_{ac}\frac{d\dot{x}^a}{ds} + g_{cb}\frac{d\dot{x}^b}{ds} \\ &= \left(g_{ac,d}\dot{x}^a + g_{cb,d}\dot{x}^b\right)\dot{x}^d + g_{dc}\ddot{x}^d + g_{cd}\ddot{x}^d \ . \end{aligned} \tag{3.148}$$

In the first term in the bracket replace the dummy index "d" by "b" and in the second term by "a". Thus

$$2g_{cd}\ddot{x}^d + \left(g_{ac,b} + g_{bc,a} - g_{ab,c}\right)\dot{x}^a\dot{x}^b = 0 \ . \tag{3.149}$$

Inverting the metric we obtain the geodesic equation. Thus the straightest available path is the shortest available path in a curved space as well. Not every "straightest path" need be "shortest" however. For example, moving away from a point on the Earth, by the straightest path would bring one back to the point by the long way around. (While the path-length would be a local minimum it would not be a global minimum.)

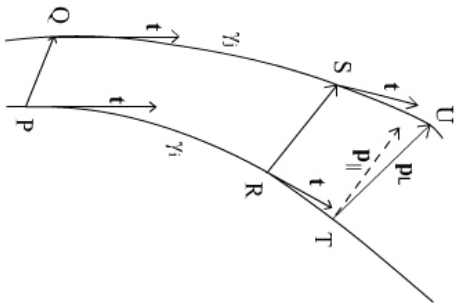


Figure 3.11: Two of a family of geodesics, γ_i and γ_j, having unit tangent $\mathbf{t}$, are connected by a separation vector $\mathbf{p}$. Acting on a point $P \in \gamma_i$, $\mathbf{p}$ gives Q on γ_j. Later, acting on $R \in \gamma_i$, $\mathbf{p}(R) = S$. Lie transport would take $T \in \gamma_i$ to $U \in \gamma_j$, while parallel transport would take it to V, i.e. $\mathbf{p}_{\mathcal{L}}(T) = U$, $p_{\parallel}(T) = V$, so that $\mathbf{p}_{\mathcal{L}} = \overrightarrow{TU}$, $p_{\parallel} = \overrightarrow{TV}$. Geodesic deviation, or acceleration, measures the difference between these two for given $\mathbf{t}$ and initial $\mathbf{p}$.

Consider two of a family of geodesics with tangent vectors $\mathbf{t}$. Hence a vector $\mathbf{p}$ that connects the geodesics will be Lie transported along the curve. Thus, analogous to the derivation of Eq.(3.135) from Eq.(3.134), we would require that $\mathcal{L}_{\mathbf{t}}\mathbf{p} = 0$. Therefore

$$t^d p^a_{;d} = p^d t^a_{;d} \ . \tag{3.150}$$

We define the *acceleration vector*, $\mathcal{A}$, which gives the *geodesic deviation*(see Fig.3.11), by

$$\begin{aligned}\mathcal{A}^{\underline{a}} &= \ddot{p}^{\underline{a}} = \frac{d^2 p^{\underline{a}}}{ds^2} \\ &= \left[t\left\{\mathbf{t}\left(\mathbf{p}\right)\right\}\right]^{\underline{a}} = t^{\underline{c}}\nabla_{\underline{c}}\left[t^{\underline{d}}\nabla_{\underline{d}}p^{\underline{a}}\right] \ . \end{aligned} \tag{3.151}$$

This can be written in component form as

$$\begin{aligned}
\mathcal{A}^a &= t^c \left[t^d p^a{}_{;d}\right]_{;c} \\
&= t^c \left[p^d t^a{}_{;d}\right]_{;c} && \text{[using Eq.(3.141)]} \\
&= t^c p^d{}_{;c} t^a{}_{;d} + t^c p^d t^a{}_{;d;c} \\
&= p^c t^d{}_{;c} t^a{}_{;d} + t^c p^d t^a{}_{;d;c} && \text{[using Eq.(3.141)]} \\
&= p^c \left(t^d t^a{}_{;d}\right)_{;c} - p^c t^d t^a{}_{;d;c} + t^c p^d t^a{}_{;d;c} \\
&= 2p^c t^d t^a{}_{;[d;c]}
\end{aligned}$$

using Eq.(3.141) and $c \leftrightarrow d$ in the third term. Then, by the Ricci identities

$$\mathcal{A}^a = R^a{}_{bcd} t^b p^c t^d \ . \tag{3.152}$$

Eqs.(3.141) and (3.142) are the basis of the derivation of Einstein's field equations given here. Unfortunately, Einstein missed this derivation and relied on a field theoretic derivation. One might think that it should not matter which derivation is used, but it actually made a crucial difference. A constant of integration arises naturally in this derivation, but not in the field theoretic derivation. The significance of this constant, and the story of the turns of its fortunes will be told later.

We again consider our earlier simple examples for illustrative purposes.

Example 1 : The Christoffel symbols in plane polar coordinates are given by Eqs.(3.87). Thus the geodesic equations are

$$\begin{aligned}
\ddot{x}^1 + \left\{_1{}^1{}_1\right\}\left(\dot{x}^1\right)^2 + \left\{_1{}^1{}_2\right\}\dot{x}^1\dot{x}^2 + \left\{_2{}^1{}_1\right\}\dot{x}^2\dot{x}^1 + \left\{_2{}^1{}_2\right\}\left(\dot{x}^2\right)^2 &= 0 \ , \\
\ddot{x}^2 + \left\{_1{}^2{}_1\right\}\left(\dot{x}^1\right)^2 + \left\{_1{}^2{}_2\right\}\dot{x}^1\dot{x}^2 + \left\{_2{}^2{}_1\right\}\dot{x}^2\dot{x}^1 + \left\{_2{}^2{}_2\right\}\left(\dot{x}^2\right)^2 &= 0 \ .
\end{aligned}$$

Therefore

$$\begin{aligned}
\ddot{r} - r\dot{\theta}^2 &= 0 \ , \\
\ddot{\theta} + 2\dot{r}\dot{\theta}/r &= 0 \ .
\end{aligned}$$

Multiply the latter of these equations by r^2 to obtain

$$(r^2\dot{\theta})^{\cdot} = 0 \ , \text{ or } r^2\dot{\theta} = h \ .$$

Therefore

$$d/ds = \left(h/r^2\right) d/d\theta \ , \tag{1}$$

provides a change of parameter from s to θ. The form of the geodesic equations now becomes

$$\frac{d}{ds}\left(\frac{d}{ds}r\right) - \frac{h^2}{r^3} = 0 \ . \tag{2}$$

Put $r = 1/u$ and use Eq.(1) to get

$$\frac{dr}{ds} = \frac{h}{r^2}\frac{d}{d\theta}\left(\frac{1}{u}\right) = \frac{h}{u^2}u^2\frac{du}{d\theta} = -h\frac{du}{d\theta}$$

Therefore

$$\frac{d}{ds}\left(\frac{dr}{ds}\right) = -h^2u^2\frac{d^2u}{d\theta^2} \ .$$

Thus Eq.(2) becomes

$$\frac{d^2u}{d\theta^2} + u = 0 \ ,$$

which implies that

$$u = \frac{1}{r} = \frac{1}{R}\cos\left(\theta - \theta_0\right) \ .$$

Therefore

$$R = r\cos\left(\theta - \theta_0\right) \ , \tag{3.153}$$

which is the equation of a straight line with least distance from the origin R, making an angle θ_0 with the $\theta = 0$ axis. The two constants (R and θ_0) are determined by the two points through which the line passes: (r_1, θ_1) and (r_2, θ_2).

Since $R^a{}_{bcd} = 0$ here, there is no geodesic deviation. This says that straight lines do not diverge from each other: whatever angle they start with relative to each other, they retain it. In mechanical terms it says that whatever relative velocity particles start with, they retain it. In other words, there is no relative acceleration between particles moving with constant velocities. Here we see the strong connection between Geometry and Mechanics. The Special Theory reduced *kinematics* to the geometry of straight lines in spacetime. Now Einstein and Grossmann reduced *dynamics* to the geometry of curves in spacetime − provided that we limit our attention to gravity.

Example 2 : Of particular interest is the simplest non-trivial example, namely a sphere in spherical coordinates, as we are literally doing *geo*metry, i.e. we can visualise it as the surface of the sphere. In this case, the Christoffel symbols are given by Eqs.(3.79). The geodesic equations are

$$\begin{aligned} \ddot{\theta} - \sin\theta\cos\theta\dot{\phi}^2 &= 0 \ , \\ \ddot{\phi} + 2\cot\theta\dot{\theta}\dot{\phi} &= 0 \ . \end{aligned}$$

Multiplying the latter equation by $\sin^2\theta$ we get

$$\left(\sin^2\theta\dot{\phi}\right)^{\cdot} = 0 \ ,$$

which implies that

$$\sin^2\theta\dot{\phi} = h \ .$$

Therefore

$$d/ds = h\csc^2\theta d/d\phi \ .$$

Hence

$$h\csc^2\theta\frac{d}{d\phi}\left(h\csc^2\theta\frac{d\theta}{d\phi}\right) - h^2\csc^2\theta\cot\theta = 0 \ .$$

Thus

$$\frac{d^2\left(\cot\theta\right)}{d\phi^2} + \cot\theta = 0 \ .$$

Finally

$$\theta = \cot^{-1}\left[A\cos\left(\phi - \phi_0\right)\right] \ . \tag{3.154}$$

The constants A and ϕ_0 are given by the requirement that the geodesic pass through the two points (θ_1, ϕ_1) and (θ_2, ϕ_2). Now $A \to \infty$ corresponds to the (coordinate) singularities $\theta = 0\ , \pi$, and $A = 0$ corresponds to the equator of the sphere, $\theta = \pi/2$. Given any two points we can choose the equator and poles of the sphere such that the equator passes through the two points. This is simply a choice of the coordinate axes. Then the geodesic is just the equatorial circle. Generally, therefore, the geodesic on a sphere lies on the circle of largest radius passing through the two points.

There seems to be an extra constant of integration, h, appearing. In Mechanics it was the angular momentum parameter. It is worthwhile to see what it turns out to be here. Eq.(2.82) becomes

$$1 = a^2\left[\dot{\theta}^2 + \sin^2\theta\dot{\phi}^2\right] \ .$$

Differentiating Eq.(3.144) and using the equation for $\dot{\phi}$ we have

$$\begin{aligned} -\csc^2\theta\,\dot{\theta} &= -A\sin\left(\phi - \phi_0\right)\dot{\phi} \\ &= -Ah\csc^2\theta\sin\left(\phi - \phi_0\right) \ . \end{aligned}$$

Therefore

$$\begin{aligned} 1 &= a^2\left[A^2h^2\sin^2\left(\phi - \phi_0\right) + h^2\cos^2\theta\right] \\ &= a^2h^2\left[A^2\sin^2\left(\phi - \phi_0\right) + 1 + \cot^2\theta\right] \\ &= a^2h^2\left[A^2\sin^2\left(\phi - \phi_0\right) + 1 + A^2\cos^2\left(\phi - \phi_0\right)\right] = a^2h^2\left[1 + A^2\right] \ . \end{aligned}$$

Hence

$$h = \frac{1}{a\sqrt{1 + A^2}} \ ,$$

so h is given in terms of the other constant A. Taking $\phi_0 = 0$, i.e. the initial longitude is the zero meridian (passing through Greenwich on the Earth), the equation of the geodesic, written as a surface curve is

$$\mathbf{x}(\phi) = \frac{1}{\sqrt{1 + A^2\cos^2\phi}}\left(\cos\phi\ ,\ \ \sin\phi\ ,\ \ A\cos\phi\right) \ .$$

It is easy to verify that this is a circle rotated through some angle, since it has a constant curvature $\kappa = 1/a$ and the normal vector is proportional to $\mathbf{x}(\phi)$. Thus, once again, h gets related to a function of an angle.

The Riemann tensor for a sphere was given in Example 2 in section 3.6. Thus we can work out non-trivial geodesic deviation. Let the curves to be taken as the geodesics which are deviating, be $\phi =$const. (i.e. lines of longitude on the Earth's surface). Then in these coordinates $\mathbf{t}$ is given by $\partial/\partial\theta = \partial/\partial x^1$, or $t^a = (1, 0)$. Then $\mathbf{p}$ satisfies Eq.(3.141), which reduces to

$$p^a_{\ ,1} = p^b t^a_{\ ,b} = 0 \ ,$$

which implies that

$$p^a = \left(p^1\left(\phi\right), p^2\left(\phi\right)\right) \ .$$

Choose the separation vector so that $p^1(\phi) = 0$, i.e. there is no θ component to the separation vector, and $p^2(\phi) = a\phi \sin\theta_0$, i.e. the distance of two geodesics from each other at latitude θ_0. Therefore

$$\left.\begin{array}{l} \mathcal{A}^1 = -R^1_{212} t^2 p^1 t^2 = 0 \ . \\ \mathcal{A}^2 = -R^2_{121} t^1 p^2 t^1 = -a\phi \ \sin\theta_0 \ . \end{array}\right\} \tag{3.155}$$

Thus the greater the distance between the two lines the greater is the acceleration! Notice that $\mathcal{A} \perp \mathbf{t}$.

For completeness we evaluate the constant of integration, h, appearing here. Eq.(2.81) becomes

$$1 = a^2[\dot\theta^2 + \sin^2\theta\dot\phi^2] \ .$$

Differentiating Eq.(3.154) and using the equation for $\dot\phi$ we have

$$\begin{aligned} -\csc^2\theta\,\dot\theta &= -A\sin(\phi-\phi_0)\,\dot\phi \\ &= -Ah\csc^2\theta\sin(\phi-\phi_0) \ . \end{aligned}$$

Therefore

$$\begin{aligned} 1 &= a^2\left[A^2h^2\sin^2(\phi-\phi_0) + h^2\cos^2\theta\right] \\ &= a^2h^2\left[A^2\sin^2(\phi-\phi_0) + 1 + \cot^2\theta\right] \\ &= a^2h^2\left[A^2\sin^2(\phi-\phi_0) + 1 + A^2\cos^2(\phi-\phi_0)\right] = a^2h^2\left[1+A^2\right] \ . \end{aligned}$$

Hence

$$h = \frac{1}{a\sqrt{1+A^2}} \ .$$

Thus the equation of the geodesic, written as a surface curve, is (taking $\phi_0 = 0$)

$$\mathbf{x}(\phi) = \frac{1}{\sqrt{1+A^2\cos^2\phi}}\left(\cos\phi,\ \sin\phi,\ A\cos\phi\right) \ .$$

It is easy to verify that this is a circle rotated through some angle, since it has a constant curvature $\kappa = 1/a$ and the normal vector is proportional to $x(\phi)$.

Example 3 : For 3-dimensional Euclidean space in spherical polar coordinates use Eqs.(3.91).

$$\ddot x^1 + \left\{_2\ ^1\ _2\right\}\left(\dot x^2\right)^2 + \left\{_3\ ^1\ _3\right\}\left(\dot x^3\right)^2 = 0 \ , \tag{1}$$

$$\ddot x^2 + 2\left\{_1\ ^2\ _2\right\}\dot x^1\dot x^2 + \left\{_3\ ^2\ _3\right\}\left(\dot x^3\right)^2 = 0 \ , \tag{2}$$

$$\ddot x^3 + 2\left\{_1\ ^3\ _3\right\}\dot x^1\dot x^3 + 2\left\{_2\ ^3\ _3\right\}\dot x^2\dot x^3 = 0 \ . \tag{3}$$

Therefore

$$\begin{aligned} \ddot r - r\dot\theta^2 - r\sin^2\theta\dot\phi^2 &= 0 \ , \\ \ddot\theta + 2\dot r\dot\theta/r - \sin\theta\cos\theta\dot\phi^2 &= 0 \ , \\ \ddot\phi + 2\dot r\dot\theta/r + 2\cot\theta\dot\theta\dot\phi &= 0 \ . \end{aligned}$$

Multiply Eq.(3) by $r^2 \sin^2\theta$ to obtain

$$\left(r^2 \sin^2\theta\dot{\phi}\right)^{\cdot} = 0$$

which implies that

$$\dot{\phi} = h/r^2 \sin^2 \ . \tag{4}$$

Thus

$$\frac{d}{ds} = \frac{h}{r^2 \sin^2\theta}\frac{d}{d\phi} \ . \tag{5}$$

Multiply Eq.(2) by r^2 to obtain

$$\left(r^2\dot{\theta}\right)^{\cdot} = r^2 \sin\theta\cos\theta\dot{\phi}^2 \ .$$

Use Eqs.(4) and (5) to rewrite this equation as

$$\frac{h}{r^2 \sin^2\theta}\frac{d}{d\phi}\left(\frac{h}{\sin^2\theta}\frac{d\theta}{d\phi}\right) = \frac{h^2\cot\theta}{\sin^2\theta} \ .$$

Hence

$$\theta = \cot^{-1}\left[A\cos\left(\phi - \phi_0\right)\right] \tag{6}$$

once again. Now the metric is

$$ds^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2 \ ,$$

which can be rewritten to give

$$r\dot{\theta}^2 + r\sin^2\theta\dot{\phi}^2 = \left(1 - \dot{r}^2\right)/r \ .$$

Inserting this into Eq.(1) and multiplying by r,

$$r\ddot{r} + \dot{r}^2 = 1 \text{ or } \dot{r}^2 = 2 \ .$$

Therefore

$$r^2 = s^2 + Bs + C = \left(s - s_0\right)^2 + D^2 \ .$$

By appropriate choice of s_0 and D, we see that r can be chosen to be the arc length parameter, $r = s$. Thus Eq.(4) can be solved using Eq.(6). However, the equation is not easy to solve in general. We choose coordinates, as before, such that $A = 0$ for the geodesic being considered. Then Eq.(1) becomes

$$hu^2\frac{d}{d\phi}\left[-hu^2\frac{d}{d\phi}\left(\frac{1}{u}\right)\right] - h^2u^3 = 0 \ .$$

whence

$$R = r\cos\left(\phi - \phi_0\right)$$

as before in Eq.(3.144). This example carries algebraic features of both previous examples, but is geometrically and physically like the first example.

3.8 Isometries and Killing's Equations

If the metric remains invariant when Lie transported along a curve, the tangent vector to the curve is called an *isometry* or a *Killing* vector (named after Wilhelm Killing). In this case the Lie derivative of the metric tensor along the tangent vector, $\mathbf{k}$, must be zero,

$$\mathcal{L}_{\mathbf{k}}\mathbf{g} = 0 \ , \tag{3.156}$$

or, in the abstract index notation,

$$k^{\underline{c}}\nabla_{\underline{c}}g_{\underline{ab}} + g_{\underline{ac}}\nabla_{\underline{b}}k^{\underline{c}} + g_{\underline{cb}}\nabla_{\underline{a}}k^{\underline{c}} = 0 \ . \tag{3.157}$$

As the first term is zero and $\mathbf{g}$ commutes with ∇ the covariant Killing vector satisfies

$$\nabla_{(\underline{a}}k_{\underline{b})} = 0 \ , \tag{3.158}$$

which can be written in component form as

$$k_{(\underline{a};\underline{b})} = 0 \ . \tag{3.159}$$

Putting Eq.(3.126) for the metric tensor in Eq.(3.147) gives

$$(K_{ab}): \qquad k^c g_{ab,c} + g_{ac}k^c_{\ ,b} + g_{bc}k^c_{\ ,a} = 0 \ . \tag{3.160}$$

Eqs.(3.147) to (3.150) are alternate forms of *Killing's equations*.

Since the Killing equations are symmetric in two indices in n dimensions there are $n(n+1)/2$ coupled, linear, first order, partial differential equations for n functions of n variables. There can be, at most, $n(n+1)/2$ arbitrary constants in their solutions. This may seem odd as there can be at most n linearly independent Killing vectors (or any vectors) in n-dimensional spaces. The point is that linear combinations of Killing vectors with different functional factors will not be Killing vectors, i.e. if $\mathbf{k}_1$ and $\mathbf{k}_2$ are Killing vectors $f_1\mathbf{k}_1 + f_2\mathbf{k}_2$ will generally not be a Killing vector for f_1, $f_2 \in \mathcal{T}$. There are at most $n(n+1)/2$ *functionally independent* Killing vectors. There can be spaces with no isometries. Such spaces would be more difficult to perform calculations in. Also, as there would be no symmetry group under which there is invariance, there may be no geometrical conserved quantities.

As the simplest possible example consider n-dimensional Euclidean space in Cartesian coordinates. Eqs.(3.148) or (3.149) can be written in the form

$$k_{a,b} + k_{b,a} = 0 \ . \tag{3.161}$$

We drop the summation convention for the present as it is not convenient for this calculation. Now, taking $a = b$ in Eq.(3.151),

$$k_{a,a} = 0 \ . \tag{3.162}$$

Differentiating Eq.(3.151) relative to x^a, using the symmetry of partial derivatives in two or more variables,

$$k_{a,ab} + k_{b,aa} = 0 \ . \tag{3.163}$$

Using Eq.(3.152) we see that the first term is identically zero. Hence k_a is linear in x^b. The most general linear functions in x^b can be written, resumeing the summation convention, as

$$k_a = C_{ab}x^b + A_a \ . \tag{3.164}$$

For Eq.(3.151) to hold we see that

$$C_{(ab)} = 0 \text{ or } C_{ab} = C_{[ab]} = -C_{[ba]} = -C_{ba} \ . \tag{3.165}$$

Thus there are $n(n-1)/2$ independent coefficients of x^b and n constants, A_a. The total is, then, $n(n+1)/2$. The skew bracket provides a generalization of the cross product. Here the cross product comes from the position vector $\mathbf{x}$ (in covariant form) and ∇, i.e. $\mathbf{x} \times \nabla$. Thus, it corresponds to $n(n-1)/2$ rotations. The second term is $\mathbf{A}.\nabla$, which are the displacements (or translations) given by the constant vector $\mathbf{A}$.

Writing $\mathbf{k}$ as a derivation in Cartesian coordinates for 3-d Euclidean space

$$\begin{aligned} \mathbf{k} = &\left[C_{12}\left(y\frac{\partial}{\partial x} - x\frac{\partial}{\partial y}\right) + C_{23}\left(z\frac{\partial}{\partial y} - y\frac{\partial}{\partial z}\right) + C_{31}\left(x\frac{\partial}{\partial z} - z\frac{\partial}{\partial x}\right)\right] \\ &+ \left[D_1\frac{\partial}{\partial x} + D_2\frac{\partial}{\partial y} + D_3\frac{\partial}{\partial z}\right] \ . \end{aligned} \tag{3.166}$$

By Noether's theorem there are three conserved quantities related to rotation, i.e. the three components of the angular momentum vector. Thus the first set of terms is the angular momentum operator. The second part, also consisting of three terms, is obviously the linear momentum operator. In curved spaces these operators need not represent isometries. Hence the corresponding quantities would not, generally, be conserved. It is very important, therefore, to look into the isometries and the symmetries in any physical description. The former is what is used in Relativity while the latter is used in Classical and Quantum Mechanics. It is worth while to see the connection between the two descriptions.

Let $(G, \cdot)$ be a group. (It is assumed that the reader is familiar with the definition of a group and a ring.) We define the *Lie product* of $a, b \in G$ by

$$(a \ o \ b) = a \cdot b \cdot a^{-1} \cdot b^{-1} = c \in G \ . \tag{3.167}$$

The pair (G, o) is called a *Lie group*. If the group is Abelian the Lie group is trivial. If it is a continuous group, i.e. each element is continuously connected to the identity, 1, and there is a ring structure, we can write

$$a = 1 + \epsilon_1 A \ , \ b = 1 + \epsilon_2 B \ , \ c = 1 + \epsilon_3 C \ , \tag{3.168}$$

where $\epsilon_1, \epsilon_2, \epsilon_3$ are infinitesimals (or nilpotents), so that we can take $\epsilon_1^2 = \epsilon_2^2 = \epsilon_3^2 = 0$. Thus

$$a^{-1} = 1 - \epsilon_1 A \ , \ b^{-1} = 1 - \epsilon_2 B \ . \tag{3.169}$$

Therefore

$$\begin{aligned} (a \ o \ b) &= (1 + \epsilon_1 A)(1 + \epsilon_2 B)(1 - \epsilon_1 A)(1 - \epsilon_2 B) \\ &= 1 + \epsilon_1 A + \epsilon_2 B - \epsilon_1 A - \epsilon_2 B + \epsilon_1\epsilon_2 AB - \epsilon_1^2 A^2 - \epsilon_1\epsilon_2 AB \\ &\quad -\epsilon_1\epsilon_2 BA - \epsilon_2^2 B^2 + \epsilon_1\epsilon_2 AB - \epsilon_1^2\epsilon_2 ABA + \epsilon_1^2\epsilon_2 A^2 B \\ &\quad -\epsilon_1\epsilon_2^2 AB^2 + \epsilon_1\epsilon_2^2 BAB + \epsilon_1^2\epsilon_2^2 ABAB \\ &= 1 + \epsilon_1\epsilon_2 (AB - BA) := 1 + \epsilon_1\epsilon_2 [A, B] = 1 + \epsilon_3 C \ , \end{aligned}$$

$$C = [A, B] \ . \tag{3.170}$$

The commutators of the infinitesimal generators of the group are called the *Lie algebra* corresponding to the group. Two globally distinct groups can have the same Lie algebra. In Classical and Quantum Mechanics one generally refers to the algebra and not to the group itself. A (generally non-unique) minimal set of generators, $\mathbf{X}_i$, the commutators of whose elements can always be expressed in terms of linear combinations of those generators, is called a *basis* for the Lie algebra. The Lie algebra is then given by the structure constants in Eq.(1.22). An example of a Lie group is the group of rotations in n-dimensional space. This can be represented by the set of $n \times n$ unimodular (whose determinant is unity) orthogonal matrices ($O^{\mathbf{T}} = O^{-1}$). It is denoted by $SO(n)$ if the space is Riemannian, or $SO(p,q)$ if the space is Lobachevskian, where $p+q=n$.

In the example of Eq.(3.156), the Killing vectors have the generators of 3-d rotations, $SO(3)$, and 3-d translations, $\mathbb{R}^3$. Since rotations and translations do not commute (see SR) we can not take a direct product of the two groups. We therefore take the *semi-direct product*, "$\otimes_s$" (to indicate that the groups do not commute), to obtain the 3-dimensional symmetry group of Euclidean space

$$E_3 = SO(3) \otimes_s \mathbb{R}^3 \ . \tag{3.171}$$

For n-dimensional Euclidean space the Killing vectors given by Eq.(3.154) have the group, E_n with 3 replaced by n in Eq.(3.161). If the space is Lobachevskian we have the Poincaré group

$$E_n = SO(p,q) \otimes_s \mathbb{R}^n \ . \tag{3.172}$$

For Minkowski space it is

$$\mathbb{P}_4 = SO(1,3) \otimes_s \mathbb{R}^4 \ , \tag{3.173}$$

which has 3 generators of ordinary rotation (corresponding to the usual conserved angular momentum), 3 generators of Lorentz transformations (corresponding to the intrinsic spin) and 4 generators of translation (corresponding to energy-momentum conservation). There can be no further geometrically conserved quantities in 4-d.

Let us again consider the same examples to illustrate the solution of the Killing equations.

Example 1 : The metric tensor in plane polar coordinates is given by Eq.(3.76), and Eqs.(3.151), written as (K_{ab}), become

$$\begin{aligned}
(K_{11}): &\quad g_{11,1}k^1 + g_{11,2}k^2 + g_{11}k^1_{,1} + g_{12}k^2_{,1} + g_{11}k^1_{,1} + g_{21}k^2_{,1} = 0 \ , \\
(K_{12}): &\quad g_{12,1}k^1 + g_{12,2}k^2 + g_{11}k^1_{,2} + g_{12}k^2_{,2} + g_{21}k^1_{,1} + g_{22}k^2_{,1} = 0 \ , \\
(K_{22}): &\quad g_{22,1}k^1 + g_{22,2}k^2 + g_{21}k^1_{,1} + g_{22}k^2_{,2} + g_{12}k^1_{,2} + g_{22}k^2_{,2} = 0 \ .
\end{aligned}$$

These equations reduce to

$$\begin{aligned}
(K_{11}): &\quad 2g_{11}k^1_{,1} = 2k^1_{,1} = 0 \quad \text{or } k^1_{,1} = 0 \ , \\
(K_{12}): &\quad g_{11}k^1_{,2} + g_{22}k^2_{,1} = 0 \quad \text{or } k^1_{,2} + r^2 k^2_{,1} = 0 \ , \\
(K_{22}): &\quad g_{22,1} + 2g_{22}k^2_{,2} = 0 \quad \text{or } 2rk^1 + 2r^2 k^2_{,2} = 0 \ .
\end{aligned}$$

Integrating (K_{11}) gives $k^1 = f(\theta)$. Insert this into (K_{12}) to get

$$k^2_{,1} = -f_\theta(\theta)/r^2 \ ,$$

which is integrated to yield

$$k^2 = f_\theta(\theta)/r + g(\theta) \ .$$

Insert these values into (K_{22}) to obtain

$$f_\theta + f_{\theta\theta}(\theta) + r g_\theta(\theta) = 0 \ .$$

Since the last term is r-dependent while the first two are not, both parts are separately zero. Thus

$$f(\theta) = c_1 \cos\theta + c_2 \sin\theta \ , \ g(\theta) = c_3 \ ,$$

where c_1, c_2, c_3 are arbitrary constants of integration. Therefore

$$\left.\begin{array}{l} k^1 = c_1 \cos\theta + c_2 \sin\theta \ , \\ k^2 = (-c_1 \sin\theta + c_2 \cos\theta)/r + c_3 \ . \end{array}\right\} \tag{3.174}$$

Thus the Killing vector can be written as the derivation

$$\mathbf{k} = c_1 \left(\cos\theta \frac{\partial}{\partial r} - \frac{\sin\theta}{r} \frac{\partial}{\partial \theta} \right) + c_2 \left(\sin\theta \frac{\partial}{\partial r} + \frac{\cos\theta}{r} \frac{\partial}{\partial \theta} \right) + c_3 \frac{\partial}{\partial \theta} \ . \tag{3.175}$$

Hence there are 3 independent Killing vectors given by Eqs.(3.174) or (3.175) corresponding to the choice of only one of the constants non-zero, while the other two are taken to be zero. Putting $c_1 = c_2 = 0$, $c_3 = 1$ gives the generator of rotations in the plane, while the other two give the translations. Writing them as $\mathbf{k}_1, \mathbf{k}_2$ and $\mathbf{k}_3$ we have

$$\begin{aligned} [\mathbf{k}_1, \mathbf{k}_2] &= 0 = -[\mathbf{k}_2, \mathbf{k}_1] \ , \\ [\mathbf{k}_2, \mathbf{k}_3] &= -\cos\theta \frac{\partial}{\partial r} + \sin\theta \frac{\partial}{\partial r} = -\mathbf{k}_1 = -[\mathbf{k}_3, \mathbf{k}_2] \ , \\ [\mathbf{k}_3, \mathbf{k}_1] &= -\sin\theta \frac{\partial}{\partial r} - \cos\theta \frac{\partial}{\partial r} = -\mathbf{k}_2 = -[\mathbf{k}_1, \mathbf{k}_3] \ . \end{aligned}$$

Thus the algebra is closed and given by the structure constants $C_{12}{}^3 = -C_{21}{}^3 = 0$, $C_{32}{}^1 = -C_{23}{}^1 = 1$, $C_{13}{}^1 = 0 = -C_{31}{}^1$, $C_{13}{}^2 = -C_{31}{}^2 = 1$. This is $SO(2) \otimes_s \mathbb{R}^2$.

Example 2 : Using the metric for a sphere given by Eq.(3.78) or (2.81)

$$\begin{array}{lll} (K_{11}): & g_{11,c} k^c + 2 g_{11} k^1_{,1} = 0 & \text{or } k^1_{,1} = 0 \ , \\ (K_{12}): & g_{11} k^1_{,2} + g_{22} k^2_{,1} = 0 & \text{or } k^1_{,2} + \sin^2\theta k^2_{,1} = 0 \ , \\ (K_{22}): & g_{22,1} k^1 + 2 g_{22} k^2_{,2} = 0 & \text{or } k^1 + \tan\theta k^2_{,2} = 0 \ . \end{array}$$

$(K_{11}) \Rightarrow k^1 = f(\phi)$. Insert into (K_{12}) to give

$$k^2_{,1} = -\csc^2\theta f_\phi(\phi) \Rightarrow k^2 = f_\phi(\phi) \cot\theta + g(\phi) \ .$$

$(K_{22}) \Rightarrow f(\phi) + f_{\phi\phi}(\phi) + g_\phi(\phi) \tan\theta = 0$.
Therefore

$$f(\phi) = c_1 \cos\phi + c_2 \sin\phi \ , \ g(\phi) = c_3 \ .$$

Thus

$$\left.\begin{array}{l} k^1 = c_1 \cos\phi + c_2 \sin\phi \ , \\ k^2 = (-c_1 \sin\phi + c_2 \cos\phi)\cot\theta + c_3 \ . \end{array}\right\} \tag{3.176}$$

Hence

$$\mathbf{k} = c_1\left(\cos\phi\frac{\partial}{\partial\theta} - \frac{\sin\phi}{\tan\theta}\frac{\partial}{\partial\phi}\right) + c_2\left(\sin\phi\frac{\partial}{\partial\theta} + \frac{\cos\phi}{\tan\theta}\frac{\partial}{\partial\phi}\right) + c_3\frac{\partial}{\partial\phi} \ . \tag{3.177}$$

These are the three generators of rotations of the sphere. Notice that even though this is a curved space it has maximum symmetry, i.e. 3 independent Killing vectors. The algebra can be checked to be $SO(3)$ i.e. $C_{12}{}^3 = -C_{21}{}^3 = 1, C_{31}{}^2 = -C_{13}{}^2 = 1$, $C_{23}{}^1 = -C_{32}{}^1$, $= 1$. Notice that there are no translations available here.

Example 3 : For the Euclidean 3-dimensional metric tensor given by Eqs.(3.90)

$$\begin{array}{lll} (K_{11}): & 2g_{11}k^1_{,1} = 0 & \text{or } k^1_{,1} = 0 \ , \\ (K_{12}): & g_{11}k^1_{,2} + g_{22}k^2_{,1} = 0 & \text{or } k^1_{,2} + r^2 k^2_{,1} = 0 \ , \\ (K_{13}): & g_{11}k^1_{,3} + g_{33}k^3_{,1} = 0 & \text{or } k^1_{,3} + r^2\sin^2\theta k^3_{,1} = 0 \ , \\ (K_{22}): & g_{22,1}k^1 + 2g_{22}k^2_{,2} = 0 & \text{or } k^1 + rk^2_{,2} = 0 \ , \\ (K_{23}): & g_{22}k^2_{,3} + g_{33}k^3_{,2} = 0 & \text{or } k^2_3 + \sin^2\theta k^3_{,2} = 0 \ , \\ (K_{33}): & g_{33,1}k^1 + g_{33,2}k^2 + 2g_{33}k^3_{,3} = 0 & \text{or } k^1 + r\cot\theta k^2_{,} + rk^3_{,3} = 0 \ . \end{array}$$

$(K_{11}) \Rightarrow k^1 = f(\theta,\phi)$. Insert into (K_{12}) and (K_{13}) to obtain

$$k^2_{,1} = -f_\theta(\theta,\phi)/r^2 \Rightarrow k^2 = f_\theta(\theta,\phi)/r + g(\theta,\phi) \ , \tag{1}$$

$$k^3_{,1} = -f_\phi(\theta,\phi)/r^2\sin^2\theta \Rightarrow k^3 = \csc^2\theta f_\phi(\theta,\phi)/r + h(\theta,\phi) \ . \tag{2}$$

Put Eq.(1) in (K_{22}) to get

$$f(\theta,\phi) + f_{\theta\theta}(\theta,\phi) + rg_\theta(\theta,\phi) = 0 \ .$$

Therefore, as the third term has r in it but the first two do not, it is zero and

$$f(\theta,\phi) = a(\phi)\cos\theta + b(\phi)\sin\theta \ , \ g(\theta,\phi) \equiv g(\phi) \ . \tag{3}$$

Thus the Killing vectors reduce to

$$k^1 = a(\phi)\cos\theta + b(\phi)\sin\theta \ , \tag{4}$$

$$k^2 = [-a(\phi)\sin\theta + b(\phi)\cos\theta]/r + g(\phi) \ . \tag{5}$$

Using Eqs.(3) to (5) in (K_{33}) we obtain

$$\left.\begin{array}{l} a(\phi)\cos\theta + b(\phi)\sin\theta + \cot\theta\,[-a(\phi)\sin\theta + b(\phi)\cos\theta] \\ +r\cot\theta + \csc^2\theta\,[a_{\phi\phi}(\phi) + b_{\phi\phi}(\phi)\sin\theta] + rh_\phi(\theta,\phi) = 0 \end{array}\right\} \tag{6}$$

Collecting the r-dependent terms we see that

$$h_\phi(\theta,\phi) + \cot\theta g(\phi) = 0$$

or

$$h(\theta,\phi) = -\cot\theta \int g(\phi)\, d\phi + l(\theta) \ . \tag{7}$$

Hence

$$\csc\theta\left[\cot\theta a_{\phi\phi}(\phi) + \{b_{\phi\phi}(\phi) + b(\phi)\}\right] = 0 \ .$$

Since $\cot\theta \neq 0$ in general, both terms are separately zero. Therefore

$$a(\phi) = a_1 + a_2\phi \ , \quad b(\phi) = c_4\cos\phi + c_5\sin\phi \ .$$

Using Eq.(7) and the above equations, Eqs.(1) to (3) become

$$k^1 = (a_1 + a_2\phi)\cos\theta + (c_1\cos\phi + c_2\sin\phi)\sin\theta \ , \tag{8}$$

$$k^2 = \left[-(a_1 + a_2\phi)\sin\theta + (c_1\cos\phi + c_2\sin\phi)\cos\theta\right]/r + g(\phi) \ , \tag{9}$$

$$k^3 = \left[a_2 + (-c_1\sin\phi + c_2\cos\phi)\sin\theta\right]\csc^2\theta/r - \cot\theta\int g(\phi)\, d\phi + l(\theta) \ . \tag{10}$$

Using Eqs.(9) and (10) in (K_{23}) gives

$$\begin{aligned} &-a_2\sin\theta + (-c_1\sin\phi + c_2\cos\phi) + r g_\phi(\phi) - a_2\csc\theta\left(1 + 2\cot^2\theta\right) \\ &-(-c_1\sin\phi + c_2\cos\phi)\cos\theta + r\int g(\phi)\, d\phi + l(\theta) = 0 \ . \end{aligned}$$

Therefore

$$\int g(\phi)\, d\phi + g_\phi(\phi) = 0 \quad \text{or} \quad g(\phi) = c_4\cos\phi + c_5\sin\phi \ .$$

Also, since $\sin\theta$ is generally non-zero, $a_2 = 0$. Replace a_1 by c_3. Further $l(\theta) = c_6$.

Finally then

$$\left.\begin{aligned} k^1 &= (c_1\cos\phi + c_2\sin\phi)\sin\theta + c_3\cos\theta \ , \\ k^2 &= \left[-(c_1\cos\phi + c_2\sin\phi)\cos\theta c_3\sin\theta\right]/r + (c_4\cos\phi + c_5\sin\phi) \ , \\ k^3 &= (-c_1\sin\phi + c_2\cos\phi)/r\sin\theta + (c_4\cos\phi + c_5\sin\phi)\cot\theta + c_6 \ , \end{aligned}\right\} \tag{3.178}$$

or as derivations

$$\left.\begin{aligned} \mathbf{k} =& c_1\left(\cos\phi\sin\theta\frac{\partial}{\partial r} + \frac{\cos\phi\cos\theta}{r}\frac{\partial}{\partial\theta} - \frac{\sin\phi}{r\sin\theta}\frac{\partial}{\partial\phi}\right) \\ &+ c_2\left(\sin\phi\sin\theta\frac{\partial}{\partial r} + \frac{\sin\phi\cos\theta}{r}\frac{\partial}{\partial\theta} + \frac{\cos\phi}{r\sin\theta}\frac{\partial}{\partial\phi}\right) \\ &+ c_3\left(\cos\theta\frac{\partial}{\partial r} - \frac{\sin\theta}{r}\frac{\partial}{\partial\theta}\right) + c_4\left(\cos\phi\frac{\partial}{\partial\theta} + \sin\phi\cot\theta\frac{\partial}{\partial\phi}\right) \\ &+ c_5\left(\sin\phi\frac{\partial}{\partial\theta} - \cos\phi\cot\theta\frac{\partial}{\partial\phi}\right) + c_6\frac{\partial}{\partial\phi} \ . \end{aligned}\right\} \tag{3.179}$$

Therefore there are 6 independent Killing vectors. These correspond to the 6 vectors given by Eq.(3.156) using spherical polar coordinates. They contain the three dimensional rotations and the three dimensional translations. The algebra satisfied by these Killing vectors is $SO(3)\otimes_s \mathbb{R}^3$.

3.9 Miscellaneous Topics in Geometry

Some further geometrical topics of relevance for Relativity will be presented in this section.

Since the contravariant basis vector, $\delta^{\underline{a}}_{a}$ or $\mathbf{e}_a$, corresponds to $\partial/\partial x^a$, its dual, $\delta^{a}_{\underline{a}}$ or $\mathbf{e}^a$, must correspond to dx^a. Thus a covariant vector, $\mathbf{X}$, can be written as $X_a dx^a$. It is called a *differential form*, or more precisely a *differential 1-form,* often abbreviated to "*1-form*". A totally skew covariant tensor of rank p is called a (*differential*) *p-form*. Thus

$$\left.\begin{array}{l} \mathbf{A} = A_{a\ldots c} dx^a \wedge \ldots \wedge dx^c \text{ or} \\ A_{\underline{a}\ldots\underline{c}} = A_{[\underline{a}\ldots\underline{c}]} = A_{a\ldots c}\delta^{a}_{[\underline{a}}\ldots\delta^{c}_{\underline{c}]} \end{array}\right\} , \tag{3.180}$$

where '$\wedge$' denotes a skew product, which absorbs the skew brackets into the "wedge product", and limits the range of indices so that $a < \ldots < c$, thus obviating the need to divide by $p!$.

The sum of two p-forms is a p-form. The exterior product of a p-form and a q-form is a $(p+q)$-form.

$$\mathbf{A} \wedge \mathbf{B} = A_{a\ldots c} B_{d\ldots f} dx^a \wedge \ldots \wedge dx^c \wedge dx^d \wedge \ldots \wedge dx^f \ . \tag{3.181}$$

The *exterior derivative* of a p-form is a $(p+1)$-form

$$\begin{aligned} \mathbf{d} \wedge \mathbf{A} &= A_{a\ldots c;d} dx^d \wedge dx^a \wedge \ldots \wedge dx^c \\ &= A_{a\ldots c,d} dx^d \wedge dx^a \wedge \ldots \wedge dx^c \ , \end{aligned} \tag{3.182}$$

where the covariant derivative reduces to the partial derivative as the skew of the Christoffel symbols in the lower indices gives zero. Clearly, for all p-forms

$$\mathbf{d}^2 \wedge \mathbf{A} := \mathbf{d} \wedge (\mathbf{d} \wedge \mathbf{A}) = 0 \ . \tag{3.183}$$

This is the generalization of the result that the curl of the gradient of a scalar is zero. Of course, it is a very big generalization, going into higher dimensions and arbitrary rank tensors.

Of particular interest are the *volume* and *hypersurface forms*. These are defined using the n-dimensional *Levi-Civita symbols*

$$\left.\begin{array}{ll} \epsilon_{a\ldots c} = +1 & \text{if } a\ldots c \text{ are an even permutation of } 1\ldots n \\ = -1 & \ldots\ldots \text{"} \ldots\ldots \text{ odd } \ldots\ldots \text{"} \ldots\ldots \\ = 0 & \ldots\ldots \text{"} \ldots\ldots \text{ not a } \ldots\ldots \text{"} \ldots\ldots \end{array}\right\} , \tag{3.184}$$

which is totally skew in all its n indices, i.e. interchanging any pair of indices reverses the sign of the symbol. Then

$$dV = \epsilon_{a\ldots d} dx^a \wedge \ldots \wedge dx^d \ , \tag{3.185}$$

$$dS_a = \epsilon_{ab\ldots d} dx^b \wedge \ldots \wedge dx^d \ , \tag{3.186}$$

are the volume n-forms and the hypersurface $(n-1)$-form respectively. Forms are particularly useful for discussing integration of tensors. However, forms are *not* tensors as they are not invariant (or appropriately covariant or contravariant) under coordinate transformations. This fact will be demonstrated shortly.

Given a 1-vector (i.e. a contravariant vector) $\mathbf{A}^*$, the *generalised Gauss* (divergence) *theorem* states that

$$\int \mathbf{d}\cdot\mathbf{A}^* dV = \oint_S \mathbf{A}^*\cdot\mathbf{d}S = \oint_S A^{*a} dS_a \ , \tag{3.187}$$

where S is the hypersurface that bounds the n-volume V. The dual of this theorem is the generalised Stokes theorem for an $(n-1)$-form $\mathbf{A}$, which is the dual of $\mathbf{A}^*$, having components $A_{b\ldots d} = \epsilon_{ab\ldots d}A^{a*}$, (up to a scaling factor which will become clear later). Stokes' theorem is

$$\oint_S \mathbf{A} = \int \mathbf{d}\wedge\mathbf{A} \ . \tag{3.188}$$

Here we do not need to write the left side in the form of an infinitesimal as that is already contained in the $(n-1)$-form. The $\mathbf{d}\wedge\mathbf{A}$ behaves like a generalised curl. We will not go into the proofs, or details, of these theorems here.

The scaling factor involved in the dual and the fact that the volume form is not a tensor are both related to the Levi-Civita symbol not being a tensor. Before explaining this point it is useful to point out that the usual mass density is not a scalar. Consider the mass written as the integral of density in two coordinate systems, in 3-d space

$$M = \int \rho(\mathbf{x})dV(x) = \int \rho(\hat{\mathbf{x}})dV(\hat{\mathbf{x}}) = \int \rho(\hat{\mathbf{x}})\Delta^{-1}dV(\mathbf{x}) \ , \tag{3.189}$$

where Δ is the Jacobian determinant, $\Delta = \det(\delta^a_{\hat{a}})$, referred to in section 2.4 (after Eq.(2.58)). Hence the density in the transformed coordinates is given by

$$\rho(\hat{\mathbf{x}}) = \Delta\rho(\mathbf{x}) \ . \tag{3.190}$$

A quantity satisfying such a transformation law is called a *scalar density*. Generally, if a quantity satisfies the transformation law

$$T^{\hat{a}\ldots\hat{c}}{}_{\hat{d}\ldots\hat{f}}\,(\hat{\mathbf{x}}) = \Delta^w \delta^{\hat{a}}_a\ldots\delta^{\hat{c}}_c\delta^d_{\hat{d}}\ldots\delta^f_{\hat{f}}T^{a\ldots c}{}_{d\ldots f}\,(\mathbf{x}) \tag{3.191}$$

it is called a *tensor density of weight* w.

Now, the determinant of an $n\times n$ matrix, A^a_i, is given by

$$\det(A^a_i) = \epsilon_{a\ldots c}\epsilon^{i\ldots k}A^a_i\ldots A^c_k/n! \ , \tag{3.192}$$

where $\epsilon^{i\ldots k}$ is defined exactly as in Eq.(3.174). Thus

$$g(\mathbf{x}) := \det(g_{ai}) = \epsilon^{a\ldots c}\epsilon^{i\ldots k}g_{ai}(\mathbf{x})\ldots g_{ck}(\mathbf{x})/n! \ . \tag{3.193}$$

Changing the coordinate system we see that

$$\begin{aligned} g(\hat{\mathbf{x}}) &= \det(g_{\hat{a}\hat{i}}) = \epsilon^{\hat{a}\ldots\hat{c}}\epsilon^{\hat{i}\ldots\hat{k}}\delta^a_{\hat{a}}\ldots\delta^c_{\hat{c}}\delta^i_{\hat{i}}\ldots\delta^k_{\hat{k}}g_{ai}(\mathbf{x})\ldots g_{ck}(\mathbf{x})/n! \\ &= \Delta^2 g(\mathbf{x}) \ . \end{aligned} \tag{3.194}$$

Thus the determinant of the metric tensor is a scalar density of weight 2. Now $\epsilon^{i\ldots k}$ is *not* defined in terms of $\epsilon_{a\ldots c}$ by raising indices with the inverse metric

tensor. Thus it is not a tensor. Let $e_{a\ldots c}$ be the components of a tensor proportional to $\epsilon_{a\ldots c}$ such that $e^{i\ldots k}$ is defined by using g^{ai}. Then it is clear that the factor of proportionality is the Jacobian $\Delta = \sqrt{g}$. Thus

$$e_{a\ldots c} = \sqrt{g}\left(\epsilon_{a\ldots c}/\sqrt{n!}\right) \ , \quad e^{i\ldots k} = \left(\epsilon^{i\ldots k}/\sqrt{n!}\right)/\sqrt{g} \ . \tag{3.195}$$

Clearly, we have, for a general volume element,

$$dV(\hat{\mathbf{x}}) = \det(\delta^a_{\hat{a}})dy(\mathbf{x}) = \Delta^{-1}dV(\mathbf{x}) \ . \tag{3.196}$$

Thus the volume and surface elements are not scalar and vector quantities but the corresponding densities. Similarly, the dual of a 1-vector is defined using the "e" and not the "ϵ".

It is often useful in Relativity to consider length re-scaling by a position dependant scalar function, as gravity will change length and time intervals

$$ds^2 \to d\tilde{s}^2 = \Omega^2(\mathbf{x})ds^2 \ , \tag{3.197}$$

where $\Omega^2(\mathbf{x})$ is called the *conformal factor*. It is written as a whole square to ensure that the sign of the metric is not changed by the re-scaling. Such transformations are called *conformal transformations*. This is equivalent to maintaining the "size of objects", dx^a, while changing the "size of the measuring rod", g_{ab}. Thus Eq.(3.187) can be obtained from

$$g_{\underline{ab}} \to \tilde{g}_{\underline{ab}} = \Omega^2(\mathbf{x})g_{\underline{ab}} \ . \tag{3.198}$$

Conformal transformations do not alter the angle between vectors. This may be seen easily. Consider two contravariant vectors $\mathbf{A}$ and $\mathbf{B}$ with an angle θ between them. Then by Eq.(2.50)

$$\cos\theta = \frac{\mathbf{A}\cdot\mathbf{B}}{|\mathbf{A}|\,|\mathbf{B}|} = \frac{g_{\underline{ab}}A^{\underline{a}}B^{\underline{b}}}{\sqrt{g_{\underline{cd}}A^{\underline{c}}A^{\underline{d}}}\sqrt{g_{\underline{ef}}B^{\underline{e}}B^{\underline{f}}}} \ . \tag{3.199}$$

If $\mathbf{A}$ and $\mathbf{B}$ are unaltered but $g_{\underline{ab}}$ is transformed according to Eq.(3.198), it is clear that $\cos\theta$, and hence θ, is unaltered.

A useful quantity constructed from the curvature tensor is its trace-free part, possessing the same symmetries as the Riemann tensor. Thus

$$C^{\underline{ab}}{}_{\underline{cd}} = R^{\underline{ab}}{}_{\underline{cd}} + \alpha\delta^{[\underline{a}}_{[\underline{c}}R^{\underline{b}]}_{\underline{d}]} + \beta\delta^{[\underline{a}}_{[\underline{c}}\delta^{\underline{b}]}_{\underline{d}]}R \tag{3.200}$$

subject to the requirement that

$$C^{\underline{ab}}{}_{\underline{ad}} = 0 \ . \tag{3.201}$$

Therefore

$$\begin{aligned} 0 &= R^{\underline{ab}}{}_{\underline{ad}} + \frac{1}{4}\alpha\left(\delta^{\underline{a}}_{\underline{a}}R^{\underline{b}}_{\underline{d}} - \delta^{\underline{a}}_{\underline{d}}R^{\underline{b}}_{\underline{a}} - \delta^{\underline{b}}_{\underline{a}}R^{\underline{a}}_{\underline{d}} + \delta^{\underline{b}}_{\underline{d}}R^{\underline{a}}_{\underline{a}}\right) \\ &\quad + \frac{1}{2}\beta\left(\delta^{\underline{a}}_{\underline{a}}\delta^{\underline{b}}_{\underline{d}} - \delta^{\underline{a}}_{\underline{d}}\delta^{\underline{b}}_{\underline{a}}\right)R \\ &= \left[1 + \frac{1}{4}\alpha\,(4-1-1)\right]R^{\underline{b}}_{\underline{d}} + \left[\frac{1}{4}\alpha + \frac{1}{2}\beta\,(4-1)\right]\delta^{\underline{b}}_{\underline{d}}R \\ &= (1+\alpha/2)\,R^{\underline{b}}_{\underline{d}} + \left(\frac{3}{2}\beta - \frac{1}{4}\alpha\right)R^{\underline{b}}_{\underline{d}}R \ . \end{aligned}$$

Thus

$$\alpha = -2 \ , \quad \beta = 1/3 \ .$$

Hence

$$C^{\underline{ab}}{}_{\underline{cd}} = R^{\underline{ab}}{}_{\underline{cd}} - 2\delta^{[\underline{a}}_{[\underline{c}} R^{\underline{b}]}_{\underline{d}]} + 1/3\delta^{[\underline{a}}_{[\underline{c}}\delta^{\underline{b}]}_{\underline{d}]} R \ . \tag{3.202}$$

The corresponding trace-free quantity for the Ricci tensor, called the *Segre tensor*, is

$$Q^{\underline{a}}_{\underline{c}} = R^{\underline{a}}_{\underline{c}} - \frac{1}{4}\delta^{\underline{a}}_{\underline{c}} R \ . \tag{3.203}$$

An important property of $C^{\underline{ab}}{}_{\underline{cd}}$ is that

$$\left.\begin{array}{ll} \tilde{C}_{\underline{abcd}} = \Omega^2 C_{\underline{abcd}} \ , & \tilde{C}^{\underline{a}}{}_{\underline{bcd}} = C^{\underline{a}}{}_{\underline{bcd}} \ , \\ \tilde{C}^{\underline{ab}}{}_{\underline{cd}} = \Omega^{-2} C^{\underline{ab}}{}_{\underline{cd}} \ , & \tilde{C}^{\underline{abc}}{}_{\underline{d}} = \Omega^{-4} C^{\underline{abc}}{}_{\underline{d}} \ , \\ \tilde{C}^{\underline{abcd}} = \Omega^{-6} C^{\underline{abcd}} & \end{array}\right\} \tag{3.204}$$

It is called the *Weyl*, or the *conformal*, *curvature tensor* on account of the above property. It gives the matter-free part, or the pure gravitational part, of the curvature. The structure of this tensor can be better understood in terms of spinors. In particular, the property given by Eq.(3.194) is very difficult and tedious to prove using tensors, but comes out relatively easily using spinors. We will not dwell on conformal geometry, or spinors, here but will mention them in the last chapter.

3.10 Exercises

1. Construct a spherical spiral for a curve embedded in $\mathbb{R}^3$. Now generalize to a curve on a hypersphere embedded in $\mathbb{R}^4$. Construct the moving quadrad (4 basis vectors). Work out the first second and third curvatures for it.

2. A particle in Minkowski space is moving on a spiral path with constant rate of change of speed. Construct the moving quadrad for it.

3. An $\mathbb{S}^3$ is defined by $w^2+x^2+y^2+z^2=a^2$. Construct a parameterization for it in terms of three parameters, say $\mathbf{x}(p,q,r)$ and hence obtain the metric tensor components in terms of these coordinates. Use the metric tensor to obtain the Riemann and Ricci tensors and the Ricci scalar for it. Also compute the Weyl tensor.

4. Extend the above analysis to the hyper-hyperboloids: (a) $w^2+x^2+y^2-z^2=a^2$; (b) $w^2+x^2-y^2-z^2=a^2$. Are there any more distinct types?

5.(a) Construct the geodesic equations on a spheroid and try to find a class of simple solutions for the system.
(b) Repeat the process for a right hyperboloid.
(c) Try to find the general solutions for both the above. If you cannot manage to do so, explain why.

6.(a) On an $\mathbb{S}^3$ take the family of geodesics passing through $\phi=0,\pi$. First take a geodesic lying along the equator and the initial point, P_o, at $\phi=\pi/4$, as the observer's path. Construct, $\mathbf{t}$ for it. Now, consider the position vector, $\mathbf{p}$, joining the initial point to Q_o at $(\pi/4,\pi/4)$. There will be a $\mathbf{t}$ along the geodesic passing through the second point. Lie transport $\mathbf{p}$ along $\mathbf{t}$ moving along the equator. This gives the position vector for the moving point Q_ϕ as seen from P_ϕ, for any $\phi>\pi/4$.
(b) Now regard the geodesic through Q as that of the observer and P as that of the particle being observed. This is done by reversing the sign of $\mathbf{p}$. Now, use the $\mathbf{t}$ for the second geodesic to Lie transport $-\mathbf{p}$ and verify that you recover the original $\mathbf{t}$ and P.

7.(a) Compute the isometries on the $\mathbb{S}^3$ and demonstrate that the Lie algebra is $SO(4)$.
(b) Extend the above analysis to a right hyper-hyperboloid of 1-sheet. What is its symmetry group? What about the second type of hyper-hyperboloid?
(c) Now obtain the isometries of the hyper-paraboloid and the hyper-spheroid, and obtain their symmetry groups.

8. There is no Weyl scalar as the tensor is traceless. Construct a scalar quantity from it and work it out for the other hypersurfaces given above.

Chapter 4

Einstein's Unrestricted Theory of Relativity

In the Geometry developed so far either very general n-d manifolds or some specific 2 and 3-d examples have been discussed (apart from the formalism for curves in a flat 4-d space). For standard Relativity we use 4-d Lobachevskian manifolds with *signature* $(+,-,-,-)$, i.e. the diagonalised metric tensor has a positive time component[1]. Manifolds with this signature are called spacetimes. Points in such manifolds are called *events* to indicate something happening somewhere at some time. Curves representing the paths of particles are called *world-lines*. If the particles are massive the tangent vectors to the world-lines at all points will be time-like (i.e. $\mathbf{t}\cdot\mathbf{t} > 0$ and in fact $\mathbf{t}\cdot\mathbf{t} = 1$) while for massless particles they will be null (i.e. $\mathbf{t}\cdot\mathbf{t} = 0$ in that $ds^2 = 0 = g_{\mu\nu}dx^\mu dx^\nu$). In no case can the tangent vector to the world-line of a particle be spacelike (i.e. $\mathbf{t}\cdot\mathbf{t} < 0$ and in fact $\mathbf{t}\cdot\mathbf{t} = -1$). A spacelike vector represents a body having spatial extension. The set of world-lines traced out by a body having spatial extension is called a *world-tube*. Fig. 1.2 depicts a world-tube, for example.

There are higher dimensional extensions of standard Relativity and there are competing theories of gravity. However there is no experimental or observational evidence supporting any of them against Relativity, i.e. they only agree with experiment (at most) to the extent that Relativity does. Generally all of them appeal to some additional guiding principle or attempts at unification of some kind. Here we will deal only with standard 4-d Relativity, only occasionally referring to generalizations, extensions or modifications in passing so as to retain contact with modern trends in Theoretical Physics.

4.1 The Matter-Stress-Energy-Momentum Tensor (Density)

General Relativity deals with the gravitational field. This field depends upon the spatial distribution of matter and on its temporal evolution. Thus we need

[1] The reverse signature $(-,+,+,+)$ could have been used equally well. In fact this convention was used first, with time as the fourth instead of the zeroth coordinate. Relativists are divided about equally between the two conventions. We will use the former convention, i.e. $(+,-,-,-)$.

a mathematical description of the distribution of matter in spacetime. Since Special Relativity inextricably links mass and energy, the relativistic (spacetime) description must incorporate the distribution of energy as well. The energy could either be carried by matter or stored in a field. In particular, it could be contained in stresses set up in a medium. (We will approximate the medium by a continuum.) Before going on to the full relativistic description it is worthwhile to very briefly review the classical, 3-d description of stresses.

"Stress" is a generalization of the concept of "pressure", which is force per unit area. Since "force" and "area" are both vector quantities in 3-d, it would not generally be meaningful to divide one by the other. The concept of "pressure" is applicable if the directions of the vectors do not matter and only the magnitudes are relevant. A medium for which this requirement holds is called *isotropic*. For example the pressure of a gas, as deduced from the kinetic theory of gases, is the same in all directions. Similarly, water is an isotropic medium. However, consider a helical spring (such as is used for weighing or in air guns) stretched or squashed. The energy stored in it can be released by motion in only one direction and not orthogonal to it. Such a medium is *anisotropic*. For example geological faults store energy anisotropically. Clearly the usual concept of "pressure" will not apply here. The generalization to anisotropic media is called *stress.*

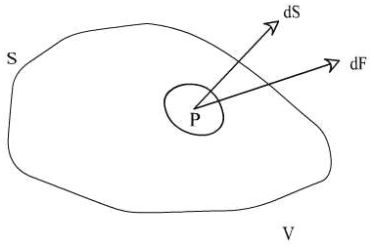


Figure 4.1: A region of volume V is bounded by a surface S. The force $d\mathbf{F}$ acts on an area element $d\mathbf{S}$ with centre at P. Had both vectors been in the same direction the ratio of their magnitudes would have given the pressure at P.

The stress is given by the *stress-tensor*

$$\sigma^{ij} = \frac{dF^i}{dS_j} \ \ (i,\ j = 1,\ 2\ 3) \ , \tag{4.1}$$

Where dF^i is the force acting on the area element dS_j (in some coordinates), see Fig 4.1. The skew part of this tensor, $\sigma^{[ij]}$, gives the *rotation*. From this quantity we can define the *vorticity vector* by using the totally skew tensor in 3-d, e_{ijk},

$$\Omega_k = \frac{1}{2} e_{ijk}\sigma^{ij} \ . \tag{4.2}$$

Assuming an irrotational medium, so that $\Omega_k = 0$ the stress tensor will be symmetric, $\sigma^{ij} = \sigma^{(ij)}$. It is this symmetric part of the stress tensor that is the generalization of pressure.

For the relativistic generalization of the stress tensor we need to adjoin something to σ^{ij} to make it into a 4×4 matrix that will give the components of a 4-d second rank tensor. Pressure (or stress) has units of energy density. The

only part of the 4-d tensor that would accommodate the energy density, ρc^2, in a manner consistent with the requirement of Lorentz covariance (i.e. invariance of the tensor under Lorentz transformations) is the time-time part. In the rest-frame there can only be the rest-energy and the stresses with no kinetic component. Thus, in the rest-frame in Minkowski spacetime the 4-d tensor is

$$T^{\mu\nu} = \rho c^2 \delta^\mu_0 \delta^\nu_0 + \sigma^{ij} \delta^\mu_i \delta^\nu_j \ . \tag{4.3}$$

Of course, as we shall see later, the spacetime will not remain flat if matter is present. Eq.(4.3) can be generalized for an arbitrary frame and an arbitrary manifold by

$$T^{\mu\nu} = \rho u^\mu u^\nu + \sigma^{ij} \delta^\mu_i \delta^\nu_j \ , \tag{4.4}$$

where u^μ is the velocity 4-vector, $\dot{x}^\mu$ (the dot referring to a derivative with respect to proper time). This vector has magnitude c. Thus in the rest-frame

$$u^\mu := \dot{x}^\mu = c\delta^\mu_0 \left(g_{00}\right)^{-1/2} = \left(c, 0, 0, 0\right) / \sqrt{g_{00}} \ . \tag{4.5}$$

(Remember that $x^0 = ct$.) In the case that there are no stresses in an arbitrary frame, $T^{\mu\nu}$ gives the total energy and the momentum of any portion of the fluid. It is, therefore, called the *energy-momentum tensor*. As it generally also gives stresses it is also called the *stress-energy tensor*. Since it also gives the distribution of matter (and energy) it is called the *matter tensor*. (Since all these names are used in the literature I put them all together in the section heading.) Further, since the components of the tensor are energy *densities* (or equivalents), it can not really be a tensor at all. It is in fact a tensor density of weight 1. It is only the volume integral over a spacelike hypersurface (i.e. a hypersurface with a time-like normal) that can give the *energy* as components, and hence be a genuine tensor.

We will need to refer to two conservation equations of continuum mechanics (which will not be proved rigorously here). The first is the *equation of continuity* arising from the principle of conservation of mass. For mass to be conserved it can neither be created nor destroyed. Hence, the total time derivative of the matter density must vanish

$$D\rho/Dt = 0 \ . \tag{4.6}$$

Writing this in terms of partial derivatives with respect to time and position, Eq.(4.6) becomes

$$\partial\rho/\partial t + \nabla \cdot (\rho \mathbf{u}) = 0 \ . \tag{4.7}$$

(Notice the analogy with the equation of continuity of electromagnetism in which mass density would be replaced by charge density.) The other is the momentum conservation equation which says that the total force at a point in equilibrium is zero. The body force on a matter element must be balanced by the release of tension. The former is given by $\rho\mathbf{a}$, where

$$\mathbf{a} = D\mathbf{u}/Dt \ , \tag{4.8}$$

while the latter is given by the divergence of the stress-tensor $\sigma^{ij}{}_{,j}$. Hence the momentum conservation equation is

$$\rho Du^i/Dt + \sigma^{ij}{}_{,j} = 0 \ . \tag{4.9}$$

There is also an energy conservation equation in classical continuum mechanics. That is contained in the above equations if we use the relativistic mass, or energy, density instead of only rest-mass density. Taking ρ to be the relativistic mass density Eqs.(4.7) and (4.8) can be written in relativistic, 4-vector form as

$$\begin{aligned} \left(\rho u^{\nu}\right)_{,\nu} &= 0 \ , & (4.10) \\ \rho u^{\nu} u^{k}_{,\nu} + \sigma^{kj}_{\ \ ,j} &= 0 \ . & (4.11) \end{aligned}$$

Now consider the divergence of the stress-energy tensor in a space that is approximately Minkowski, using Cartesian coordinates. Since the Christoffel symbols are (approximately) zero here, the covariant derivative can be replaced by the partial derivative. The divergence of $T^{0\nu}$ in Eq.(4.4) is

$$\begin{aligned} T^{0\nu}{}_{,\nu} &= \left(\rho c u^{\nu}\right)_{,\nu} + \left(\sigma^{ij}\delta^{0}_{i}\delta^{\nu}_{j}\right)_{,\nu} \\ &= c\left(\rho u^{\nu}\right)_{,\nu} \ . & (4.12) \end{aligned}$$

From Eq.(4.10) we see that this is zero. Similarly, consider the divergence of $T^{k\nu}$ in Eq.(4.4),

$$\begin{aligned} T^{k\nu}_{\ \ ,\nu} &= \left(\rho u^{k} u^{\nu}\right)_{,\nu} + \left(\sigma^{ij}\delta^{k}_{i}\delta^{\nu}_{j}\right)_{,\nu} \\ &= u^{k}\left(\rho u^{\nu}\right)_{,\nu} + \left[\rho u^{\nu} u^{k}_{,\nu} + \sigma^{kj}_{\ \ ,j}\right] \ . & (4.13) \end{aligned}$$

The first term is zero by Eq.(4.10) and the second by Eq.(4.11). Thus the laws of conservation of mass-energy and momentum could be stated as the requirement that the stress-energy tensor be divergence-free, i.e. its divergence be zero. For an arbitrary spacetime in arbitrary coordinates the natural generalization of this requirement is

$$T^{\mu\nu}_{\ \ ;\nu} = 0 \ . \qquad (4.14)$$

Notice that $u^{\nu}u^{k}_{,\nu} = \dot{u}^{k} = \ddot{x}^{k}$. Thus the first term of Eq.(4.11) is the "kinetic" force density while the second is the "potential" force density. In an arbitrary spacetime $\ddot{x}^{k}$ would be replaced by the left side of the geodesic equations, Eq.(3.136). For completeness let me add that considerations of angular momentum conservation lead to the identification of $\sigma^{[ij]}$ with rotational forces.

4.2 The Stress-Energy Tensor for Fields

To extend the concept of the stress-energy tensor to fields we look for a divergence-free quantity with units of energy density and which reduces to the usual stress-energy tensor for a field describing a fluid. For this purpose first consider the Lagrangian density for a scalar field, $\mathcal{L}[\phi, \phi_{,\rho}]$, whose gradient is given by

$$\frac{\partial \mathcal{L}}{\partial x^{\mu}} = \frac{\partial \phi}{\partial x^{\mu}}\frac{\delta \mathcal{L}}{\delta \phi} + \frac{\partial \phi_{,\nu}}{\partial x^{\mu}}\frac{\delta \mathcal{L}}{\delta \phi_{,\nu}} \ . \qquad (4.15)$$

We can use Eqs.(1.18), the EL equations, to replace the functional derivative in the first term on the right side of Eq.(4.15), which becomes

$$\begin{aligned} \mathcal{L}_{,\mu} &= \phi_{,\mu}\left(\frac{\delta \mathcal{L}}{\delta \phi_{,\nu}}\right)_{;\nu} + \left(\phi_{,\mu}\right)_{;\nu}\frac{\delta \mathcal{L}}{\delta \phi_{,\nu}} \\ &= \left(\phi_{,\mu}\frac{\delta \mathcal{L}}{\delta \phi_{,\nu}}\right)_{;\nu} \ . & (4.16) \end{aligned}$$

Also it is trivially obvious that

$$\mathcal{L}_{,\mu} = \left(\delta^{\nu}_{\mu}\mathcal{L}\right)_{;\nu} \ . \tag{4.17}$$

Thus, for the scalar field

$$T^{\nu}_{\mu} = \phi_{,\mu}\frac{\delta\mathcal{L}}{\delta\phi_{,\nu}} - \delta^{\nu}_{\mu}\mathcal{L}, \tag{4.18}$$

is divergence-free, i.e. it satisfies Eq.(4.14). Clearly it has units of energy density. If the Lagrangian is written as the difference between the kinetic and potential energy, T^0_0 would be the *sum* of the kinetic and potential energy as $\delta\mathcal{L}/\delta\phi_{,\nu}$ corresponds to p_ν. Thus T^0_0 corresponds to the total energy density, as required. Hence $T^{\mu\nu}$, as given by Eq.(4.18) (with the index raised), is the stress-energy tensor for the scalar field.

Generally, the field need not be a scalar but can be a vector valued function, ϕ_r. The index r could be a spacetime vector or tensor index, or it could refer to some other space, as for example the space spanned by the generators of some "internal" symmetry group. Or, as in Yang-Mills fields, it could be both. For such fields the Lagrangian density is $\mathcal{L}[\phi_r, \phi_{r,\mu}]$. Notice that here it is *not* the covariant derivative of ϕ_r that enters into the Lagrangian but the partial derivative. This arises from it being a scalar density and not a scalar. The covariant derivative, when r is a spacetime vector or tensor index, introduces a Christoffel symbol $\{_\mu\ ^\nu\ _\nu\}$ which is $(\ln\sqrt{|g|})_{,\mu}$, by Eq.(3.95). This involves $|\ g\ |_{,\mu}$ which converts the vector into a vector density. Following the previous procedure here we obtain

$$T^{\mu\nu} = g^{\mu\rho}\phi_{r,\rho}\frac{\delta\mathcal{L}}{\delta\phi_{r,\nu}} - g^{\mu\nu}\mathcal{L} \ . \tag{4.19}$$

Correctly speaking, if angular momentum is conserved, the stress-energy tensor is given by symmetrizing over the indices μ and ν. Whereas the second term on the right side of Eq.(4.19) is already symmetric the first term is generally not symmetric.

As an example consider the case of electromagnetism. Here $\mathcal{L}$ is given by Eq.(1.21). Then, for the source-free case $j^\mu = 0$, Eq.(4.19) gives

$$T^{\mu\nu}_{(em)} = -\frac{1}{4\pi}\left(F^{\mu\rho}F^{\nu}{}_{\rho} - \frac{1}{4}g^{\mu\nu}F^{\rho\sigma}F_{\rho\sigma}\right) \ . \tag{4.20}$$

In the flat spacetime approximation it is easily verified that $T^{00}$and T^{0i} are the usual Hamiltonian (total energy) density and the Poynting vector (momentum density)

$$\mathcal{H} = T^{00}_{(em)} = \left(E^2 + H^2\right)/8\pi \ , \tag{4.21}$$

$$\mathcal{P}^i = T^{0i}_{(em)} = \left(\mathbf{E}\times\mathbf{H}\right)^i/8\pi = \epsilon^{ijk}E_jH_k/8\pi \ , \tag{4.22}$$

where $\mathbf{E}$ and $\mathbf{H}$ are the electric and magnetic field intensities. The purely spatial part of the stress-energy tensor gives the *Maxwell stress-tensor*

$$T^{ij} = -\frac{1}{4\pi}\left[E^iE^j + H^iH^j + \frac{1}{2}g^{ij}\left(\mathbf{E}\cdot\mathbf{E} + \mathbf{H}\cdot\mathbf{H}\right)\right] \ . \tag{4.23}$$

Notice that the trace of $T_{(em)}$, $T^{\mu}_{\mu} = T$ is zero from Eq.(4.20). Thus the "pressure" in the electromagnetic field exactly balances the energy density of the field. If sources were included we would have to modify Eq.(4.20) by subtracting $g^{\mu\nu}A_{\rho}j^{\rho}$ from the right side of the equation.

4.3 The Einstein Field Equations

In the rest-frame, the world-line of an observer, O, is a geodesic if there is no "external force" acting on the observer . Let the unit tangent vector to this geodesic, $\dot{x}^{\mu}/c$ be denoted by $\mathbf{t}$. Then $\mathbf{t}$ is a time-like vector. Suppose this observer sees another observer, O', on whom there is also no "external force", as being at a position $\mathbf{p}(s)$. The geodesics of the two observers belong to the same family and hence $\mathbf{t}(s)$ is the same vector field for both observers. Since $\mathbf{p}(s)$ continues to join the geodesics of the same family it is Lie transported along the geodesic. The position vector is obviously space-like. Then, by Eq.(3.142) the geodesic deviation, $\mathcal{A}^{\mu}$, is related to the curvature. Now

$$\mathcal{A}^{\mu} = \ddot{p}^{\mu} = t^{\rho}\left[t^{\nu}\left(p^{\mu}\right)_{;\nu}\right]_{;\rho} = R^{\mu}{}_{\nu\rho\pi}\, t^{\nu}p^{\rho}t^{\pi} \tag{4.24}$$

is the usual acceleration, being the second derivative of the position with respect to proper-time. Thus the acceleration is related to the curvature of spacetime. Classically we would say that "gravitation causes acceleration and is caused by the presence of matter". Therefore, we would expect to be able to find a relationship between the distribution of matter (and energy) and curvature. Let the relationship be expressed by

$$\mathcal{E}^{\mu\nu}\left[g_{\alpha\beta}, R^{\alpha}{}_{\beta\gamma\delta}\right] = \kappa T^{\mu\nu} \tag{4.25}$$

where $\mathcal{E}^{\mu\nu}$ is a tensor function of the metric tensor and the curvature tensor and κ is a constant of proportionality.

We will determine $\mathcal{E}^{\mu\nu}$ now, leaving κ for the next section. The tensor $\mathcal{E}^{\mu\nu}$ must be divergence-free so that Eq.(4.25) can be consistent with Eq.(4.14) to ensure that mass-energy-momentum are conserved. Also $\mathcal{E}^{\mu\nu}$ must be a symmetric tensor. The procedure adopted will be to take the simplest possible function that can be chosen consistently with physical requirements. Simplicity will be, essentially, measured by the order of dependence on the curvature tensor. Thus the simplest function which satisfies the mathematical requirements is $\mathcal{E}^{\mu\nu} \propto g^{\mu\nu}$. This is not only divergence-free it is even "gradient-free"! It gives a constant stress-energy tensor. Since $T^{\mu\nu}$ need not, generally, be constant this is a trivial case. The simplest non-trivial choice would be a linear function of curvature.

For the function to be divergence-free we consider the contraction of Eq.(3.121) over a and b

$$\begin{aligned} 0 &= R^{ab}{}_{ac;d} + R^{ab}{}_{cd;a} + R^{ab}{}_{da;c} \\ &= R^{b}{}_{c;d} + R^{ab}{}_{cd;a} - R^{b}{}_{d;c} \ . \end{aligned} \tag{4.26}$$

Now contracting b and d gives

$$\begin{aligned} 0 &= R^{d}{}_{c;d} + R^{ad}{}_{cd;a} - R^{d}{}_{d;c} \\ &= R^{d}{}_{c;d} + R^{a}{}_{c;a} - R_{;c} \\ &= 2\left(R^{d}_{c} - \frac{1}{2}\delta^{d}_{c}R\right)_{;d} \end{aligned} \tag{4.27}$$

Hence, for the 4-dimensional spacetime the linear, symmetric, divergence-free function of the curvature is

$$\mathcal{E}^{\mu\nu} = R^{\mu\nu} - \frac{1}{2}Rg^{\mu\nu} \ . \tag{4.28}$$

This is called the *Einstein tensor*. Inserting this equation into Eq.(4.25) we get the *Einstein field equations*

$$R^{\mu\nu} - \frac{1}{2}Rg^{\mu\nu} = \kappa T^{\mu\nu} \ . \tag{4.29}$$

These were the equations originally stated by Einstein. *This* was the crucial step that Grossmann and Einstein had missed. You will notice that this derivation requires a constant of integration, since it is the divergence of the Einstein tensor that is to be proportional to the stress-energy tensor. Einstein missed this constant because he derived the equations from the variational principle using the Einstein-Hilbert Lagrangian. Later, when he applied General Relativity to Cosmology, he was unable to find what he considered a satisfactory solution to these equations. He consequently inserted a constant, Λ, to the Lagrangian, by hand, nowadays called the *cosmological* constant, to $\mathcal{E}^{\mu\nu}$, so that Eq.(4.29) becomes

$$R^{\mu\nu} - \frac{1}{2}Rg^{\mu\nu} + \Lambda g^{\mu\nu} = \kappa T^{\mu\nu} \ . \tag{4.30}$$

This constant would have arisen naturally if he had followed through with his earlier approach to derive the equations as given here. The difference made is that when added to the Lagrangian, it appears as an energy that is present everywhere — what is called a "vacuum energy". When it comes as a constant of integration it is part of the geometry as a general constant curvature in the background. In the latter way of looking at it, it is naturally there and we can only determine what the curvature is by observing the Universe. In observations at the scale of the solar system, and even of the galaxy (and nearby galactic clusters), it would appear to be zero. At a larger scale it appears to be nonzero. Einstein later regretted putting in this term, calling it "the greatest blunder in my life". This issue will be discussed in chapter 7, on Cosmology.

4.4 The Newtonian Limit of the Einstein Field Equations

Apart from developing various theories of Physics and the use of Geometry, Einstein developed a Philosophy of Science. In it he formulated the procedure required to develop scientific theories — a sort of "do-it yourself construction kit" for them, based on his methods for developing theories that were fundamentally new. The main problem, he noted, was that of the infinitely many possible generalizations of any previously established theory, we need to select one; as if

we are pioneers exploring unchartered territory and need to guess what direction to take when there are no paths to follow. Previously new theories were forced on us by the weight of observational evidence; as if we were pulled in one direction by a "divining rod". now we needed to use other signs to guide us. It was in this spirit that he stated his "principles of SR", and later his "principles of GR". (Everybody got stuck in trying to similarly formulate "principles of Quantum Theory", and we still have none.)

In general, some small parameter involved in the new theory was effectively disregarded in the earlier one. His *correspondence principle* says that the new theory must reduce to Newtonian mechanics in the limit of that small parameter becoming zero. Thus for Quantum theory, $h \to 0$, must yield Newtonian theory. In the case of SR, you might think that the parameter is not small and the limit taken is infinity and not zero. However, actually Newtonian mechanics takes it that no time may be taken between any cause and the consequent effect. As such, for SR, $1/c \to 0$. Then what would be the requirement for GR? Since Einstein's "happiest thought" involved gravity, one might expect it to be $G \to 0$, but that cannot be the case since we would then "throw the baby out with the bath-water", as we would then lose Newton's gravity as well. As such, we should put it in a rather odd way, as $G^2 \to 0$.

The classical limit for General Relativity comes from the requirements that there are no non-classical effects due to motion, either due to high speeds (Special Relativity) or high accelerations (due to gravity). Thus $c \longrightarrow \infty$, or equivalently the speeds $u^i = \dot{x}^i$ are negligible, and the spacetime approximates Minkowski, i.e. $g_{00} \approx -g_{11} \approx -g_{22} \approx -g_{33} \approx 1$ and $g_{\mu\nu} = 0$ otherwise, in Cartesian coordinates. We further assume that the first derivatives of metric coefficients are small, so that their squares can be neglected, and that the geometry is not time varying, $g_{\mu\nu,0} = 0$. To check the validity of the Einstein field equations we need to verify that they reduce to the corresponding equations for the Newtonian gravitational potential and the geodesic equations reduce to the equations of motion in a gravitational field. Before discussing the Einstein equations we will first review Newtonian gravitation theory for our purpose.

The classical equation for the gravitational potential is derived analogously to that for the electrostatic potential. We first consider the gravitational field intensity, $\mathbf{g}$, at a point P at position r relative to a point gravitational source of mass m. The field intensity is defined by

$$\mathbf{g} = \mu \underset{\lim}{\to} 0 \frac{\mathbf{F}}{\mu} \ , \tag{4.31}$$

where $\mathbf{F}$ is the gravitational force on a test particle of mass μ. According to Newton's law of gravity

$$\mathbf{g} = -\frac{Gm\mathbf{r}}{r^3} \ , \tag{4.32}$$

where G is Newton's gravitational constant (6.672×10^{-11} in mks units). Consider a volume V containing the gravitational source and bounded by a surface S. The integral of the field intensity over the surface is then

$$\oint_S \mathbf{g} \cdot d\mathbf{S} = -\oint_S Gm \frac{\mathbf{r} \cdot d\mathbf{S}}{r^3} \ . \tag{4.33}$$

It is clear from Fig. 4.2 that

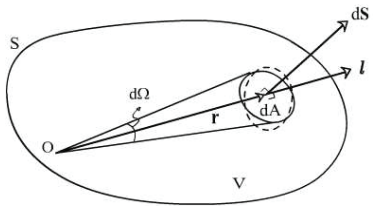


Figure 4.2: A region of volume V is bounded by a surface S. The force $d\mathbf{F}$ acts on an area element $d\mathbf{S}$ with centre at P. Had both vectors been in the same direction the ratio of their magnitudes would have given the pressure at P.

$$\frac{\mathbf{r}\cdot d\mathbf{S}}{r^3} = \frac{\mathbf{1}\cdot d\mathbf{S}}{r^2} = d\Omega \ . \tag{4.34}$$

where $\mathbf{l}$ is the unit vector in the radial direction and $d\Omega$ is the solid angle element subtended by the area element $d\mathbf{S}$ at the gravitational source. Since the solid angle subtended by a closed surface is 4π

$$\oint_S \mathbf{g}\cdot d\mathbf{S} = -4\pi G m \ . \tag{4.35}$$

Now consider n point masses $m_{1,}....,m_n$ enclosed in V. Since the surface integral of the gravitational field intensity only depends on the mass, according to Eq.(4.35), the total field intensity for the n particles must be such that its surface integral is simply the sum of the surface integrals for each of the particles, i.e.

$$\oint_S \mathbf{g}\cdot d\mathbf{S} = -4\pi G \sum_{i\,=\,1}^{n} m_i \ . \tag{4.36}$$

In the continuum limit Eq.(4.36), becomes

$$\oint_S \mathbf{g}\cdot d\mathbf{S} = -4\pi G \int_V \rho dV \ . \tag{4.37}$$

Thus, using Gauss' divergence theorem

$$\int_V \nabla\cdot\mathbf{g} dV = -\int_V 4\pi G\rho dV \ . \tag{4.38}$$

Therefore

$$\nabla\cdot\mathbf{g} = -4\pi G\rho \ . \tag{4.39}$$

The *gravitational potential*, Φ, is defined by

$$\mathbf{g} = -\nabla\Phi \ , \tag{4.40}$$

subject to the requirement that $\Phi \to 0$ sufficiently far from the gravitational source. Hence we obtain the *Poisson equation*

$$\nabla^2\Phi = 4\pi G\rho \ . \tag{4.41}$$

We now return to the relativistic treatment of gravity. First consider the geodesic equations, Eq.(3.146) broken into the time and space equations with

the various terms expanded accordingly:

$$\ddot{x}^0 + \left\{ {}_0 \; {}^0 \; {}_0 \right\} \left(\dot{x}^0\right)^2 + 2 \left\{ {}_0 \; {}^0 \; {}_i \right\} \dot{x}^0 \dot{x}^i + \left\{ {}_i \; {}^0 \; {}_j \right\} \dot{x}^i \dot{x}^j = 0 \; , \tag{4.42}$$

$$\ddot{x}^i + \left\{ {}_0 \; {}^i \; {}_0 \right\} \left(\dot{x}^0\right)^2 + 2 \left\{ {}_0 \; {}^i \; {}_j \right\} \dot{x}^0 \dot{x}^j + \left\{ {}_j \; {}^i \; {}_k \right\} \dot{x}^j \dot{x}^k = 0 \; . \tag{4.43}$$

Now, in the classical limit $\dot{x}^0 \to c$ and $\dot{x}^i \to u^i$, and $u^i/c \approx 0$. In other words we can set $\dot{x}^i = 0$ in the above equations. Also, as there is no time variation, Eqs.(3.85) give $\left\{ {}_0 \; {}^0 \; {}_0 \right\} = \left\{ {}_i \; {}^0 \; {}_j \right\} = \left\{ {}_0 \; {}^i \; {}_j \right\} = 0$. The only non-zero Christoffel symbols, again given by Eqs.(3.85), are

$$\left\{ {}_0 \; {}^0 \; {}_i \right\} = \frac{1}{2} g^{00} g_{00,i} \; , \; \left\{ {}_0 \; {}^i \; {}_0 \right\} = -\frac{1}{2} g^{ij} g_{00,j} \; . \tag{4.44}$$

Clearly, Eq.(4.12) reduces to $\ddot{x}^0 = 0$ or $\dot{x}^0 = c$. Eq.(4.43) becomes

$$\ddot{x}^i = \frac{1}{2} c^2 g^{ij} g_{00,j} \; . \tag{4.45}$$

Since the force on a test particle of mass μ would be $\mu \ddot{x}^i$, $\ddot{\mathbf{x}} = \mathbf{g}$. Lowering the index in Eq.(4.45) and comparing with Eq.(4.40) we see that

$$\Phi = \frac{1}{2} c^2 \left(g_{00} - 1 \right) \; , \tag{4.46}$$

so that $\Phi = 0$ far away from the gravitational source when $g_{00} = 1$. Hence g_{00} plays the role of the Newtonian gravitational potential.

We now come to Eq.(4.29), the Einstein field equations. Contracting the indices gives

$$R^\mu_\mu - \frac{1}{2} \delta^\mu_\mu R = R - 2R = -R = \kappa T \; . \tag{4.47}$$

Thus Eq.(4.29) can be re-written as

$$R_{\mu\nu} = \kappa \left(T_{\mu\nu} - \frac{1}{2} T g_{\mu\nu} \right) \; . \tag{4.48}$$

Therefore

$$R_{00} = \kappa \left(T_{00} - \frac{1}{2} T g_{00} \right) \; . \tag{4.49}$$

Consider a gravitational source with no stresses or other fields but consisting only of matter with density ρ. In its rest-frame $T_{00} = \rho c^2$ and $T_{\mu\nu} = 0$ otherwise. Thus

$$T = g^{\mu\nu} T_{\mu\nu} = g^{00} T_{00} = \rho c^2 \; . \tag{4.50}$$

Hence

$$R_{00} = \kappa \left(\rho c^2 - \frac{1}{2} \rho c^2 \right) = \frac{1}{2} \kappa \rho c^2 \; . \tag{4.51}$$

From Eq.(3.91) using Eq.(4.44) we have

$$\left.\begin{aligned} R_{00} = {} & \left\{ {}_0 \; {}^0 \; {}_0 \right\}_{,0} + \left\{ {}_0 \; {}^i \; {}_0 \right\}_{,i} - \left(\ln \sqrt{|g|} \right)_{,00} + \left(\ln \sqrt{|g|} \right)_{,0} \left\{ {}_0 \; {}^0 \; {}_0 \right\} \\ & + \left(\ln \sqrt{|g|} \right)_{,i} \left\{ {}_0 \; {}^i \; {}_0 \right\} - \left\{ {}_0 \; {}^0 \; {}_0 \right\}^2 - 2 \left\{ {}_0 \; {}^0 \; {}_i \right\} \left\{ {}_0 \; {}^i \; {}_0 \right\} \\ & - \left\{ {}_0 \; {}^i \; {}_j \right\} \left\{ {}_0 \; {}^j \; {}_i \right\} \; . \end{aligned}\right\} \tag{4.52}$$

Since products of Christoffel symbols involve quadratic expressions in the derivatives of the metric coefficients, they can be neglected. Thus

$$\begin{aligned} R_{00} &= \left(-\frac{1}{2}g^{ij}g_{00,j}\right)_{,i} + \left(\ln\sqrt{|g|}\right)_{,i}\left(-\frac{1}{2}g^{ij}g_{00,j}\right) \\ &\quad - 2\left(\frac{1}{2}g^{00}g_{00,i}\right)\left(-\frac{1}{2}g^{ij}g_{00,j}\right) \\ &\approx -\frac{1}{2}g^{ij}g_{00,ij} \ . \end{aligned} \tag{4.53}$$

From Eqs.(4.46), (4.51) and (4.53) we have

$$\nabla^2\Phi \approx \frac{1}{2}\kappa\rho c^4 \ . \tag{4.54}$$

Thus Einstein's equations *do* reduce to the classical gravitational field equations in the appropriate limit, as Eq.(2.24) has the same form as the Poisson equation, Eq.(4.41). In fact comparing the two equations we obtain

$$\kappa = 8\pi G/c^4 \ . \tag{4.55}$$

4.5 The Schwarzschild Solution

The Einstein field equations are a set of 10 partial differential equations for 10 functions ($g_{\mu\upsilon}$) of 4 independent variables (x^μ). They are second order and highly non-linear. The non-linearity is in both the first and second derivatives and can not be assigned any degree as the derivatives are multiplied by the inverse metric tensor. The problem of solving the Einstein equations is obviously too complicated to solve in any degree of generality. Instead of attempting a general solution some physically relevant matter tensor is chosen and assumptions are made about the spacetime symmetries of the solution. In this way the number of independent variables and the number of functions to be determined is effectively reduced. A further reduction is achieved by an appropriate choice of coordinate system (e.g. spherical-like, cylindrical-like, Cartesian-like) and coordinate frame (e.g. at rest relative to the gravitational source, an observer at infinity, a co-moving observer). Nevertheless the problem is very complicated.

The simplest case is a vacuum, $T^{\mu\nu} = 0$ and maximum symmetry. From section 3.8 we know that in 4-d maximum symmetry is achieved with 10 [$= 4\times(4+1)/2$] Killing vectors. It turns out that the unique solution of Eq.(4.29) in this case is the trivial one — Minkowski space. There *do* exist two non-trivial solutions of Eq.(4.30), rather than Eq.(4.29), which will be discussed later. A classification of solutions according to their symmetries has been achieved for *Einstein spaces*, which assume $T^{\mu\nu} \propto g_{\mu\upsilon}$ (see [19]). There are non-vacuum solutions having spherical symmetry with 7, 6, 4 and 3 Killing vectors.On account of its physical relevance, here we will be interested in solving the vacuum field equations with spherical symmetry.

Even through we are taking $T_{\mu\nu} = 0$, that does not mean that there is no matter *anywhere*. We only assume that there is none in the region under discussion. Similarly there are no stresses or energy in that region. Since there will still be gravitational effects, it should be borne in mind that gravity will not be regarded as providing any energy. As such the energy of test particles

will not generally be conserved. It will be conserved if there exists a time-like isometry in the spacetime. In that case the energy will be $E = p_\mu k^\mu$ where p_μ is the momentum 4-vector of the test particle and k^μ is the normalized time-like killing vector.

Consider a point gravitating source of mass m, in its own rest-frame and situated at the origin. This clearly implies that the solution is spherically symmetric and, excluding the origin, there is a vacuum. We will later apply our analysis to the Solar System, with the Sun as the gravitating source and the planets as the test particles. It can be rigorously shown (see [20]) that the most general spherically symmetric metric, written in spherical polar-like coordinates, is

$$ds^2 = e^{\nu(t,r)}c^2dt^2 - e^{\lambda(t,r)}dr^2 - R^2\left(t,r\right)d\Omega^2 \ , \tag{4.56}$$

where ν, λ and R are arbitrary functions of the time coordinate, t, and the radial coordinate, r. The metric coefficients are written in the above form so that the signature of the metric tensor is manifestly maintained and for convenience in subsequent calculations. For the point mass the gravitational field cannot vary with time. For example the Sun's field must remain the same. Hence all the functions depend on r only, and Eq.(4.56) reduces to

$$ds^2 = e^{\nu(r)}c^2dt^2 - e^{\lambda(r)}dr^2 - R^2\left(r\right)d\Omega^2 \ . \tag{4.57}$$

There are two possibilities as regards $R^2(r)$: it can be a constant function or a varying function. In the former case the area subtending a given solid angle, at the origin, will be independent of r. This is clearly unrealistic for describing a gravitating point mass. (Such spacetimes have six Killing vectors and are generically known as Bertotti-Robinson metrics, see [21].) In the latter case we could redefine the radial coordinate to be R instead of r. This will automatically satisfy the requirement that the area subtending a solid angle $d\Omega$ at the origin be $R^2d\Omega^2$. In terms of this new radial oordinate, we would get new functions ν and λ with the new radial distance element dR . We could now rewrite the new radial coordinate as r (just changing the symbol). Thus Eq.(4.57) becomes

$$ds^2 = e^{\nu(r)}c^2dt^2 - e^{\lambda(r)}dr^2 - r^2d\Omega^2 \ . \tag{4.58}$$

In other words, in Eq.(4.57), we choose to define the measurement of r as the square root of the ratio of the area element, dS, to the square of the solid element, $d\Omega^2$.

Thus we finally have to determine two functions of one variable instead of the three functions of two variables in Eq.(4.56). That was already a great reduction from the original 10 functions of 4 variables. Since $T_{\mu\nu} = 0$ the Einstein field equations, Eq.(4.29), become

$$\left.\begin{aligned} R_{\mu\nu} &= \{_\mu \ ^\rho \ _\nu\}_{,\rho} - \left(\ln\sqrt{|g|}\right)_{,\mu\nu} + \left(\ln\sqrt{|g|}\right)_{,\rho}\{_\mu \ ^\rho \ _\nu\} - \{_\pi \ ^\rho \ _\mu\}\{_\rho \ ^\pi \ _\nu\} \\ &= 0 \ . \end{aligned}\right\} \tag{4.59}$$

To solve these equations we need to evaluate the non-zero Christoffel symbols.

Now the metric tensor and its inverse have the components

$$g_{\mu\nu} = \begin{pmatrix} e^{\nu(r)} & 0 & 0 & 0 \\ 0 & -e^{\lambda(r)} & 0 & 0 \\ 0 & 0 & -r^2 & 0 \\ 0 & 0 & 0 & -r^2\sin^2\theta \end{pmatrix} , \tag{4.60}$$

$$g^{\mu\nu} = \begin{pmatrix} e^{-\nu(r)} & 0 & 0 & 0 \\ 0 & -e^{-\lambda(r)} & 0 & 0 \\ 0 & 0 & -1/r^2 & 0 \\ 0 & 0 & 0 & -1/r^2\sin^2\theta \end{pmatrix} , \tag{4.61}$$

and the determinant of the metric tensor components is

$$g = -e^{\nu+\lambda}r^4\sin^2\theta, \quad \text{so} \quad \ln\sqrt{|g|} = \frac{1}{2}(\nu+\lambda) + 2\ln r + \ln\sin\theta \ . \tag{4.62}$$

Thus the derivatives of the metric coefficients are

$$\left.\begin{array}{l} g_{00,1} = \nu' e^{\nu},\ g_{11,1} = -\lambda' e^{\lambda},\ g_{22,1} = -2r,\ g_{33,1} = -2r\sin^2\theta\ , \\ g_{33,2} = -2r^2\sin\theta\cos\theta,\ g_{\mu\nu,\rho} = 0 \text{ otherwise,} \end{array}\right\} \tag{4.63}$$

and the derivatives of the logarithm of the determinant are

$$\left.\begin{array}{l} \left(\ln\sqrt{|g|}\right)_{,1} = \frac{1}{2}(\nu'+\lambda') + 2/r\ , \\ \left(\ln\sqrt{|g|}\right)_{,11} = \frac{1}{2}\left(\nu'' + \lambda''\right) - 2/r^2\ , \\ \left(\ln\sqrt{|g|}\right)_{,2} = \cot\theta\ ,\ \left(\ln\sqrt{|g|}\right)_{,22} = -\csc^2\theta\ , \\ \left(\ln\sqrt{|g|}\right)_{,\mu} = \left(\ln\sqrt{|g|}\right)_{,\mu\nu} = 0 \text{ otherwise.} \end{array}\right\} \tag{4.64}$$

Since the metric tensor is diagonal we can use Eqs.(3.85). Further, since the $g_{\mu\nu}$ are independent of t and ϕ, we have

$$\left.\begin{array}{r} \left\{{}_0\ {}^0\ {}_0\right\} = \left\{{}_i\ {}^0\ {}_j\right\} = \left\{{}_0\ {}^i\ {}_j\right\} = \left\{{}_r\ {}^3\ {}_s\right\} = \left\{{}_s\ {}^r\ {}_2\right\} = \left\{{}_s\ {}^r\ {}_3\right\} = \left\{{}_3\ {}^3\ {}_3\right\} = 0 \\ (i,j = 1,2,3\ ;\ r,s = 0,1,2)\ . \end{array}\right\} \tag{4.65}$$

The non-zero Christoffel symbols, then, are

$$\left.\begin{array}{l} \left\{{}_0\ {}^0\ {}_1\right\} = \frac{1}{2}\nu'\ ,\ \left\{{}_0\ {}^1\ {}_0\right\} = \frac{1}{2}\nu' e^{\nu-\lambda}\ ,\ \left\{{}_1\ {}^1\ {}_1\right\} = \frac{1}{2}\lambda'\ , \\ \left\{{}_2\ {}^1\ {}_2\right\} = -re^{-\lambda}, \left\{{}_3\ {}^1\ {}_3\right\} = -r\sin^2\theta\ e^{-\lambda}\ , \\ \left\{{}_1\ {}^2\ {}_2\right\} = 1/r = \left\{{}_1\ {}^3\ {}_3\right\}, \\ \left\{{}_3\ {}^2\ {}_3\right\} = -\sin\theta\cos\theta\ ,\ \left\{{}_2\ {}^3\ {}_3\right\} = \cot\theta\ . \end{array}\right\} \tag{4.66}$$

It should be pointed out here that some procedure needs to be adopted to ensure that no Christoffel symbols are missed out. A convenient method that I follow is to start with the lowest indices and work one's way up. The lowest will be when all are zero. Then keep zero above zero to the left and run up the other indices to the right. Due to symmetry the reverse order need not be calculated. Now, keeping zero above take one to the left and run up from *one* (not from zero due to symmetry). Complete the full range in general, i.e. $\left\{{}_1\ {}^0\ {}_2\right\}$ and $\left\{{}_1\ {}^0\ {}_3\right\}$. However, for a diagonal metric tensor these are zero anyhow, as all

three indices are unequal. Thus, for a diagonal metric tensor one could evaluate only $\left\{_0\ ^0\ _0\right\}$, $\left\{_0\ ^0\ _1\right\}$, $\left\{_0\ ^0\ _2\right\}$, $\left\{_0\ ^0\ _3\right\}$, $\left\{_1\ ^0\ _1\right\}$, $\left\{_2\ ^0\ _2\right\}$, $\left\{_3\ ^0\ _3\right\}$. Now follow the same procedure with one on top, then two and then three.

We can now write down Eqs.(4.59) explicitly, for the metric given by Eq.(4.58) by using the Christoffel symbols given by Eqs.(4.64) - (4.66).

$$\left.\begin{array}{rcl} R_{00} & = & \left\{_0\ ^\rho\ _0\right\}_{,\rho} - \left(\ln\sqrt{|g|}\right)_{,00} + \left(\ln\sqrt{|g|}\right)_{,\rho}\left\{_0\ ^\rho\ _0\right\} - \left\{_\pi\ ^\rho\ _0\right\}\left\{_\rho\ ^\pi\ _0\right\} \\ & = & \left\{_0\ ^1\ _0\right\}_{,1} + \left(\ln\sqrt{|g|}\right)_{,1}\left\{_0\ ^1\ _0\right\} - \left\{_1\ ^0\ _0\right\}\left\{_0\ ^1\ _0\right\} - \left\{_0\ ^1\ _0\right\}\left\{_1\ ^0\ _0\right\} \\ & = & \left(\frac{1}{2}\nu' e^{\nu-\lambda}\right)' + \left[\frac{1}{2}\left(\nu'+\lambda'\right)+\frac{2}{r}\right]\left(\frac{1}{2}\nu' e^{\nu-\lambda}\right) - 2\left(\frac{1}{2}\nu'\right)\left(\frac{1}{2}\nu' e^{\nu-\lambda}\right) \\ & = & \frac{1}{2}\left[\nu'' + \nu'\left(\nu'-\lambda'\right) + \frac{1}{2}\nu'\left(\nu'+\lambda'\right) + \frac{2}{r}\nu' - \nu'^2\right] e^{\nu-\lambda} \\ & = & \frac{1}{2}\left[\nu'' + \frac{1}{2}\nu'\left(\nu'-\lambda'\right) + \frac{2}{r}\nu'\right] e^{\nu-\lambda} = 0\ . \end{array}\right\} \quad (4.67)$$

$$\left.\begin{array}{rcl} R_{11} & = & \left\{_1\ ^\rho\ _1\right\}_{,\rho} - \left(\ln\sqrt{|g|}\right)_{,11} + \left(\ln\sqrt{|g|}\right)_{,\rho}\left\{_1\ ^\rho\ _1\right\} - \left\{_\pi\ ^\rho\ _1\right\}\left\{_\rho{}^\pi\ _1\right\} \\ & = & \left\{_1\ ^1\ _1\right\}_{,1} - \left(\ln\sqrt{|g|}\right)_{,11} + \left(\ln\sqrt{|g|}\right)_{,1}\left\{_1\ ^1\ _1\right\} - \left\{_0\ ^0\ _1\right\}^2 \\ & & - \left\{_1\ ^1\ _1\right\}^2 - \left\{_2\ ^2\ _1\right\}^2 - \left\{_3\ ^3\ _1\right\}^2 \\ & = & \left(\frac{1}{2}\lambda'\right)' - \left[\frac{1}{2}\left(\nu''+\lambda''\right) - 2/r^2\right] + \left[\frac{1}{2}\left(\nu'+\lambda'\right)+2/r\right]\left(\frac{1}{2}\lambda'\right) \\ & & - \left(\frac{1}{2}\nu'\right)^2 - \left(\frac{1}{2}\lambda'\right)^2 - 1/r^2 - 1/r^2 \\ & = & -\frac{1}{2}\left[\nu'' + \frac{1}{2}\nu'\left(\nu'-\lambda'\right) - \frac{2}{r}\lambda'\right] = 0\ . \end{array}\right\} \quad (4.68)$$

Comparing Eqs.(4.67) and (4.68) we see that

$$\nu(r) + \lambda(r) = constant\ . \quad (4.69)$$

If this constant of integration were non-zero we could define a new function $\bar{\nu}(r) = \nu(r) - constant$. Thus the coefficient of the time interval square would be $e^{\bar{\nu}(r)} \cdot e^{constant} \cdot c^2$. This would amount to re-scaling c. Since that coefficient is, physically, the square of the speed of light, we can take the constant to be zero. We say that we have "absorbed the constant into the units of measurement of time". Thus we have

$$\nu(r) = -\lambda(r)\ . \quad (4.70)$$

We can use this result in the remaining two equations.

$$\left.\begin{array}{rcl} R_{22} & = & \left\{_2\ ^\rho\ _2\right\}_{,\rho} - \left(\ln\sqrt{|g|}\right)_{,22} + \left(\ln\sqrt{|g|}\right)_{,\rho}\left\{_2\ ^\rho\ _2\right\} - \left\{_\pi\ ^\rho\ _2\right\}\left\{_\rho\ ^\pi\ _2\right\} \\ & = & \left\{_2\ ^1\ _2\right\}_{,1} + \left(\cos ec^2\theta\right) + \left(\ln\sqrt{|g|}\right)_{,1}\left\{_2\ ^1\ _2\right\} - 2\left\{_2\ ^1\ _2\right\} \\ & & + \left\{_1\ ^2\ _2\right\} - \left\{_3\ ^3\ _2\right\}^2 \\ & = & \left(-re^{-\lambda}\right)' + \csc^2\theta + \left[\frac{1}{2}\left(\nu'+\lambda'\right)+\frac{2}{r}\right]\left(-re^{-\lambda}\right) \\ & & - 2\left(-re^{-\lambda}\right)(1/r) - \cot^2\theta \\ & = & \left(-re^{-\lambda}\right)' + 1 = 0\ . \end{array}\right\} \quad (4.71)$$

Integrating Eq.(4.71) and dividing through by r

$$e^{-\lambda(r)} = e^{\nu(r)} = 1 + \alpha/r\ , \quad (4.72)$$

where α is the constant of integration.

$$\left.\begin{array}{rcl} R_{33} & = & \left\{_3\ ^\rho\ _3\right\}_{,\rho} - (\ln\sqrt{|g|})_{,33} + (\ln\sqrt{|g|})_{,\rho}\left\{_3\ ^\rho\ _3\right\} - \left\{_\pi\ ^\rho\ _3\right\}\left\{_\rho\ ^\pi\ _3\right\} \\ & = & \left\{_3\ ^1\ _3\right\}_{,1} + \left\{_3\ ^2\ _3\right\}_{,2} + (\ln\sqrt{|g|})_{,1}\left\{_3\ ^1\ _3\right\} + (\ln\sqrt{|g|})_{,2}\left\{_3\ ^2\ _3\right\} \\ & & -2\left\{_3\ ^1\ _3\right\}\left\{_1\ ^3\ _3\right\} - 2\left\{_3\ ^2\ _3\right\}\left\{_2\ ^3\ _3\right\} \\ & = & (-r\sin^2\theta e^{-\lambda})' + (-\sin\theta\cos\theta)_{,2} + [\tfrac{1}{2}(\nu'+\lambda') + \tfrac{2}{r}](-r\sin^2\theta e^{-\lambda}) \\ & & + (\cot\theta)(-\sin\theta\cos\theta) - 2(-r\sin^2\theta e^{-\lambda})(1/r) \\ & & -2(-\sin\theta\cos\theta)\cot\theta \\ & = & \left(-re^{-\lambda}\right)'\sin^2\theta + \sin^2\theta - \cos^2\theta - \cos^2\theta + 2\cos^2\theta \\ & = & [\left(-re^{-\lambda}\right)' + 1]\sin^2\theta = R_{22}\sin^2\theta = 0\ . \end{array}\right\} \quad (4.73)$$

The fact that $R_{33} = R_{22}\sin^2\theta$ comes from the spherical symmetry of the metric. It is easy to verify that Eq.(4.72) is a solution of Eqs.(4.67) and (4.68). It is worth verifying (as an exercise) that the other components of the Ricci tensor are identically zero.

We have, therefore, a complete solution of the Einstein vacuum field equations with spherical symmetry

$$ds^2 = c^2(1+\alpha/r)\,dt^2 - \frac{dr^2}{(1+\alpha/r)} - r^2 d\Omega^2\ . \quad (4.74)$$

It was derived by Karl Schwarzschild within three months of Einstein's paper on GR being published. (It is interesting that Schwarzschild was fighting at the Russian front of the German push in World War I and he was in the infirmary on account of an illness contracted at the front. He sent this work to Einstein, who presented it at the Warsaw Academy on behalf of Schwarzschild. Schwarzschild went on to derive another exact solution of the Einstein equations (which will be discussed in the next chapter) within months of this one. He died in May 1916.) The above solution is known as the *Schwarzschild solution*. The only part of it still to be determined is α. This metric can be deduced by considering small accelerations in Special Relativity, which is the weak-field limit of General Relativity. By comparing the metric so obtained with the Schwarzschild solution, using the correspondence principle α may be determined. This comparison gives (see SR)

$$\alpha = -2Gm/c^2\ . \quad (4.75)$$

In the next section we will see how this value can be obtained from the relativistic equations of motion.

4.6 The Relativistic Equation of Motion

We will now proceed to derive the relativistic equations of motion and require that in the classical limit they reduce to the classical equations of motion in a gravitational field. The relativistic equations required are simply obtained by writing the geodesic equations for the Schwarzschild metric. Before proceeding to them, however, we will first obtain the classical equations of motion for the sake of comparison.

For a point gravitating source of mass m (at rest at the origin) we have a central force which has magnitude $-Gm\mu/r^2$, acting on a test particle of mass μ at position r. Due to conservation of angular momentum, in the absence of any

torque the direction (and of course magnitude) of the angular momentum vector will remain constant. Since this vector is orthogonal to the osculating plane for the orbit, the curve followed is coplanar. Thus the equation of motion can be obtained from Eq.(3.76) by setting the radial part equal to the gravitational field intensity and the polar part equal to zero.

$$\begin{aligned} \ddot{r} - r\dot{\theta}^2 &= -Gm/r^2 \ , & (4.76)\\ r\ddot{\theta} + 2\dot{r}\dot{\theta} &= 0 \ . & (4.77) \end{aligned}$$

Multiplying Eq.(4.77) by r and integrating gives

$$r^2\dot{\theta} = h \quad \text{or} \quad \frac{d}{dt} = \frac{h}{r^2}\frac{d}{d\theta} \ , \tag{4.78}$$

where h is a constant of integration. Using Eq.(4.78) in Eq.(4.76) gives

$$\frac{h}{r^2}\frac{d}{d\theta}\left(\frac{h}{r^2}\frac{d}{d\theta}r\right) - \frac{h^2}{r^3} = -\frac{Gm}{r^2} \ . \tag{4.79}$$

Putting $r = 1/u$ as before, we obtain the classical equation of motion

$$\frac{d^2u}{d\theta^2} + u = \frac{Gm}{h^2} \ , \tag{4.80}$$

which has the solution

$$\begin{aligned} u = \frac{1}{r} &= \frac{Gm}{h^2} + A\cos\left(\theta - \theta_0\right) \\ &= \frac{1}{R}\left[1 + \varepsilon\cos\left(\theta - \theta_0\right)\right] \ , & (4.81) \end{aligned}$$

where θ_0 and ε (or A) are arbitrary constants of integration. Eq.(4.81) is the polar form of the equation of an ellipse with semi-major axis

$$a = h^2/Gm\left(1-\varepsilon^2\right) = R/\left(1-\varepsilon^2\right) \ , \tag{4.82}$$

and eccentricity ε, with the vertices of the ellipse along the direction $\theta = \theta_0$ and $\theta = \theta_0 + \pi$.

For the relativistic equations of motion we use Eqs.(4.65) in Eq.(3.136) to obtain

$$\left.\begin{array}{l} \ddot{x}^0 + 2\left\{{}_0\ {}^0\ {}_1\right\}\dot{x}^0\dot{x}^1 = 0 \ , \\ \ddot{x}^1 + \left\{{}_0\ {}^1\ {}_0\right\}\left(\dot{x}^0\right)^2 + \left\{{}_1\ {}^1\ {}_1\right\}\left(\dot{x}^1\right)^2 + \left\{{}_2\ {}^1\ {}_2\right\}\left(\dot{x}^2\right)^2 + \left\{{}_3\ {}^1\ {}_3\right\}\left(\dot{x}^3\right)^2 = 0 \ , \\ \ddot{x}^2 + 2\left\{{}_1\ {}^2\ {}_2\right\}\dot{x}^1\dot{x}^2 + \left\{{}_3\ {}^2\ {}_3\right\}\left(\dot{x}^3\right)^2 = 0 \ , \\ \ddot{x}^3 + 2\left\{{}_1\ {}^3\ {}_3\right\}\dot{x}^1\dot{x}^2 + 2\left\{{}_2\ {}^3\ {}_3\right\}\dot{x}^2\dot{x}^3 = 0 \ . \end{array}\right\} \tag{4.83}$$

Using Eq.(4.66) we can rewrite Eq.(4.83) in terms of t, r, θ, ϕ as

$$\begin{aligned} c\ddot{t} + \nu' c\dot{t}\dot{r} &= 0 \ , & (4.84)\\ \ddot{r} + \frac{1}{2}\nu' e^{\nu-\lambda}c^2\dot{t}^2 + \frac{1}{2}\lambda'\dot{r}^2 - re^{-\lambda}\dot{\theta}^2 - r\sin^2\theta e^{-\lambda}\dot{\phi}^2 &= 0 \ , & (4.85)\\ \ddot{\theta} + \frac{2}{r}\dot{r}\dot{\theta} - \sin\theta\cos\theta\dot{\phi}^2 &= 0 \ , & (4.86)\\ \ddot{\phi} + \frac{2}{r}\dot{r}\dot{\phi} + 2\cot\theta\dot{\theta}\dot{\phi} &= 0 \ . & (4.87) \end{aligned}$$

Eqs.(4.86) and (4.87) are identical with the polar geodesic equations in 3-dimensional Euclidean space (see Example 3 in section. 3.7, whose solution is given by Eq.(6) there). By taking an appropriate choice of the constant ($A = 0$) the motion will be in the equatorial plane

$$\theta = \pi/2 \ , \dot{\theta} = 0 \ , \tag{4.88}$$

$$\dot{\phi} = \frac{h}{r^2} \quad \text{or} \quad \frac{d}{d\tau} = \frac{h}{r^2}\frac{d}{d\phi} \ , \tag{4.89}$$

where τ is the proper-time. Further, since

$$(e^{\nu})^{\cdot} = \dot{\nu}e^{\nu} = \nu'\dot{r}e^{\nu} \ , \tag{4.90}$$

we can multiply Eq.(4.84) by e^{ν} and integrate to obtain

$$\left(e^{\nu}\dot{t}\right)^{\cdot} = 0 \quad \text{or} \quad \dot{t} = ke^{-\nu} \ , \tag{4.91}$$

where k is a constant of integration. Since the spacetime must be Minkowski sufficiently far from the gravitational source, i.e. as $r \to \infty$, and the proper-time is the coordinate time (it being the rest-frame), $\dot{t} \to 1$ there. From Eq.(4.72) $e^{\nu} \to 1$ in this limit. Since s has units of length, $k = c$ in Eq.(4.91).

We now have to simplify Eq.(4.85) and reduce it to a form which can be compared with Eq.(4.80). To this end we use Eq.(4.58) to eliminate the $\dot{r}^2$ term in Eq.(4.85) and (4.91) we get

$$1 = e^{-\nu} - \left(\dot{r}^2 e^{\lambda} + \frac{h^2}{r^2}\right) \ . \tag{4.92}$$

Using Eq.(4.70) we get

$$c^2 - \dot{r}^2 = \left(c^2 - h^2/r^2\right) e^{-\lambda} \ . \tag{4.93}$$

Also, differentiating Eq.(4.72) we see that

$$\nu' e^{\nu} = -\lambda' e^{-\lambda} = -\alpha/r^2 \ . \tag{4.94}$$

Using Eqs.(4.66), (4.88), (4.89), (4.91), (4.92) and (4.94) we obtain

$$\ddot{r} - \frac{\alpha}{2r^2}\left(c^2 + \frac{h^2}{r^2}\right) - \frac{h^2}{r^3}\left(1 + \frac{\alpha}{r}\right) = 0 \ . \tag{4.95}$$

Again using Eq.(4.89) we have

$$\frac{h}{r^2}\frac{d}{d\phi}\left(\frac{h}{r^2}\frac{d}{d\phi}r\right) - \frac{h^2}{r^3} = \frac{\alpha c^2}{2r^2} + \frac{3\alpha h^2}{2r^4} \ . \tag{4.96}$$

Putting $r = 1/u$ we finally obtain

$$\frac{d^2u}{d\phi^2} + u = -\frac{\alpha c^2}{2h^2} - \frac{3\alpha}{2}u^2 = \alpha c^2\left(-\frac{1}{2h^2} - 3u^2\right) \ . \tag{4.97}$$

Comparing Eqs.(4.97) and (4.80) we see that we require that α be given by Eq.(4.75), in which case the second term in Eq.(4.97) will disappear in the limit $c \to \infty$. Thus the relativistic equation of motion is

$$\frac{d^2u}{d\phi^2} + u = \frac{Gm}{h^2} + \frac{3Gm}{c^2}u^2 \ . \tag{4.98}$$

Notice that we have ϕ instead of the θ which appears in Eq.(4.80) because of the difference of definition of θ in plane and spherical polar coordinates. Also notice that Eq.(4.98) provides a good definition of what we mean by "weak gravitational field" in the context of the classical limit. We require that in this limit the last term be negligible. Negligible compared with what? Negligible compared with u. In other words we require that

$$3Gmu/c^2 \ll 1 \text{ or } 3Gm/c^2r \ll 1 \ . \tag{4.99}$$

This is the regime in which we can calculate "relativistic corrections" to classical theory instead of requiring new concepts.

4.7 The First "Three Tests" of General Relativity

When Einstein finally presented his un-restricted theory of Relativity he was able to demonstrate that it not only reproduced the successes of Newton's theory of gravitation but it explained an outstanding problem of the time. Mercury does not follow the orbit prescribed for it by Newtonian Physics! He also deduced two additional effects which could provide tests for his theory. One of these was indistinguishable from the Special Relativistic effects, not only with the limits of accuracy of the time but even with the technology of today. This is the frequency shift of light due to the presence of a gravitational field. The other gives a value significantly different from the Special Relativistic effects. This is the bending of the path of light due to a gravitational pull. These three were regarded as the tests of the theory, originally. Before going on to the relativistic analysis let us take a brief look at the history of these "effects".

As we saw, in the last section, the classical equations of motion predict that a test particle should move in an ellipse about a point gravitational source. To the extent that a planet can be regarded as a test particle and the Sun as a point gravitating source with spherical symmetry, we should expect the planet to move in an ellipse with the Sun at one focus. To this extent the prediction does hold. However, a planet is not massless. The mass of the planet acts on the Sun. For one thing there are tidal effects. The Sun pulls different parts of the planet by different amounts and, contrariwise, the planet pulls different parts of the Sun by different amounts — the closer parts being pulled more than the farther parts. We will neglect these effects in our present discussion and regard the Sun and planets as point particles of different, but finite, masses.

First consider the Sun with only one planet and let their masses be M and m respectively. Newton's equations can be solved exactly for this system of two particles by introducing the "reduced mass" $mM/\left(m+M\right)$, which is slightly less than the mass of the planet, being $m(1-m/M+...)$. Both bodies behave as if they had this mass and were orbiting about a body at the centre of mass with a mass $(m+M)$. They have distance from the centre of mass in the inverse ratio of their masses. If we consider more than one planet along with the Sun there is generally no exact solution of the equations of motion for the system of particles.

For an n-body system it is most convenient to use Lagrange's generalized coordinates and apply perturbation theory to obtain an approximate solution. For the Solar System this procedure gave an exact fit with observation (within the limits of observational error) for eight of the nine planets. The orbits predicted are rosette shapes corresponding to precessing ellipses, see Fig. 4.3. The

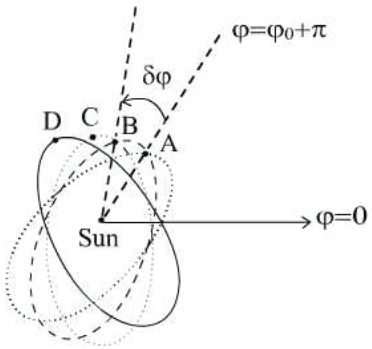


Figure 4.3: An elliptical orbit with the Sun at a focus has perihelion (point of closest approach) at A, which is at $\varphi = \varphi_0 + \pi$. The precessing ellipse traces out a rosette shape. From one orbit to the next the perihelion shifts from A to B through an angle $\delta\varphi$. It then shifts by $\delta\varphi$ from B to C and then from C to D and so on.

precession is most clearly apparent at the perihelion, the point closest to the Sun. From the diagram one might have expected to be best able to see the precession of the *apphelion* (the point of the orbit furthest from the Sun). However, thinking of planetary orbits in terms of these ellipses drawn about the Sun is grossly misleading. In the actual observations there are no orbits seen in space. The full period for one orbit of a planet is measured in months or years and not in minutes or seconds. What we see easily is *not* the distance of the planet from the Sun, but the speed with which the planet is moving, which could be seen by making measurements during one night of observation, and can currently be easily seen by the Doppler shift in the light from the planet. Also, the shift is easiest to see for highly eccentric orbits. The perihelion shift of Mercury was classically predicted to be $521.12''$ of arc per century. This value is $42.12''$ of arc per century more than the observed value. (The classical effect is due to the gravitational pull of the other planets.)

To explain the discrepancy between theory and observation a new planet was postulated. It was to lie in an orbit between Mercury and the Sun. From the requirement that it produce the required perihelion shift of Mercury without affecting the other planets significantly its mass and orbit could be worked out. So sure were the Astronomers that, like Uranus, Neptune and Pluto, the planet would be found, they named the planet Vulcan, after the Greek god connected with Hell. (*This*, presumably, is where Spock's Vulcan of *Star-Trek* fame came from.) It would not have been seen earlier as it was so close to the Sun that it would be visible only briefly, at dusk or dawn, since it would be swamped by the Sun's light in the sky for the rest of the day. From the mass, making some assumption about the density and *albedo* (the fraction of incident light that is reflected) its brightness could be estimated. Making allowances for all uncertainties it should have been detectable, once one knew where to look, either to the side of the Sun, or in transit across its face. However, it was not detected even with the best telescopes of the time.

Giving up Vulcan, there was an idea that the planet could be elsewhere and have the required effect on Mercury. The masses of the other planets could be

then fitted with observation. The problem with this suggestion was that the planet would then be seen. One way to avoid the planet being seen was by having it always stay on the other side of the Sun relative to the Earth. For this purpose it would have to lie in the same orbit as the Earth. To provide the required perihelion shift of Mercury it would have a mass more or less the same as that of the Earth. On account of its position this hypothetical planet was called anti-Earth. This suggestion was not viable on two counts. The first is that it would have perturbed Venus and Mars substantially. The second is that it would have been in an unstable equilibrium position. Let me elaborate further. While there is no general solution to the n-body problem for three bodies there are two equilibrium configuration which are exact solutions. One is when the three bodies lie along a straight line and the other when they form an equilateral triangle. The former is unstable while the latter is stable. The Earth-Sun-anti Earth system would be unstable. Hence there was no satisfactory solution to the problem of the perihelion shift of Mercury. Similar ideas have been suggested to explain how comets get pushed out of their orbits and fall in towards the Sun. None of these suggestions have worked out either.

To his great delight Einstein was able to explain the perihelion shift of Mercury in terms of his field equations. We will use the Schwarzschild solution to derive this result, but remember that Einstein did not have that solution to provide the answer and had to use other methods to obtain the effect. This led to a delay in his publishing the paper. In the event, that meant that he published it *after* Hilbert had published his paper on the Lagrangian for GR. However, since Hilbert acknowledged Einstein's priority, there was no serious loss to Einstein in the delay. It is worth remarking that Hilbert, unlike Newton, had no desire to take credit that should rightfully go to someone else. To me, *that* establishes his greatness.

Newton had contributed to two major branches of Physics — Mechanics, with special reference to Gravitation, and Light. He believed that light should be subject to gravity. This belief may have had its roots in his corpuscular theory of light. Certainly, if light consisted of massive particles a ray of light should be deflected by the gravitational pull of massive bodies. The actual mass of the particles would be irrelevant as the mass appears linearly in the law of inertia and the law of gravitation. As such, so long as it is non-zero, it cancels and the effect depends only on the gravitational field encountered. It is worth mentioning that Newton took the analogy between Optics and Mechanics further. He talked of the paths of the planets being "refracted" about the Sun. This very geometrical view of light and mechanics under gravity, presaging General Relativity, was unfortunately lost. (It is ironic that Newton generally gets much credit that should, by rights, go to others. However, in this instance he is denied credit for a profound insight.) The reasons why Newton's prediction of the gravitational deflection of light did not gain currency are probably two-fold. For one his prediction was not quantitative. For another the general belief was that "light was a wave and not a particle", whatever that might mean. As such usual mechanical considerations did not apply. The idea was revived later but never gained much credence.

The idea was effectively revived by Einstein when he used kinematical consideration to deduce a formula for the gravitational deflection of light. The argument appealed to Special Relativity with corrections due to small accelerations (see SR). He further managed to suggest an astronomical test for his

prediction. Luckily for him World War I broke out before his theory could be tested. In the mean time he had worked out his unrestricted theory whereby he could perform a complete dynamical calculation. The result came out to be twice his previous prediction. When the experiment was finally performed in 1919 it verified the prediction here using the equations of motion, Eq.(4.98), for the Schwarzschild solution.

Einstein had postulated the "photon" as the fundamental constituent of light. This theory was not simply a re-hash of Newton's corpuscular theory but had the weight of experimental evidence behind it. Using that theory and Special Relativity with small accelerations (see SR) he had deduced a frequency shift of light in moving from higher to lower gravitational potentials, or *vice versa*. The result depended not only on the mass-energy relation but also on the energy-frequency relation for light. Using the Schwarzschild solution he obtained substantially the same result with no reference to the Quantum nature of light. We will derive this result in the next section.

4.8 The Gravitational Red-Shift

Consider a light signal consisting of n waves sent from a point A, at a distance r_A from the gravitational source, to the point B at distance r_B. Let the corresponding frequency of light emission and reception be ν_A and ν_B respectively. The proper time interval for emission and reception are obtained by putting $dr = d\theta = d\phi = 0$ in Eq.(4.74), using Eq.(4.75) to give α.

$$\left.\begin{aligned} d\tau_A &= \sqrt{1 - 2Gm/c^2 r_A}\, dt \ , \\ d\tau_B &= \sqrt{1 - 2Gm/c^2 r_B}\, dt \ . \end{aligned}\right\} \tag{4.100}$$

Since $\nu_A d\tau_A = n = \nu_B d\tau_B$ the change in frequency is given by

$$\frac{\nu_B}{\nu_A} = \sqrt{\frac{1 - 2Gm/c^2 r_A}{1 - 2Gm/c^2 r_B}} \ . \tag{4.101}$$

Let us consider particular applications of this result.

Let the light emitted from the surface of a star of radius R with a frequency ν be received very far away with a frequency $\nu + \delta\nu$. We can put $r_A = R$ and $r_B = \infty$. Thus

$$\frac{\delta\nu}{\nu} = \sqrt{1 - \frac{2Gm}{c^2 R}} - 1 \approx -\frac{Gm}{c^2 R} \ , \tag{4.102}$$

which gives a red-shift of light. As shown in SR this effect may be understood in terms of Special Relativity as due to light climbing out of a gravitational potential well. Correspondingly, light entering into a gravitational potential well would suffer a blue-shift. In particular light entering the gravitational potential well for the Earth would be blue-shifted by an amount

$$\frac{\delta\nu}{\nu} \approx -\frac{Gm_E}{c^2 R_E} \ , \tag{4.103}$$

where m_E is the mass of the Earth and R_E its radius. Similarly for light climbing out of the gravitational potential well of another galaxy and entering our galaxy. For a normal star, such as our Sun, $\delta\nu/\nu \approx -10^{-6}$. For the Earth

$\delta\nu/\nu \approx 7 \times 10^{-10}$. The blue-shift at the Earth due to the Sun $\approx 4 \times 10^{-10}$. For light coming from another galaxy the red-shift $\sim -9 \times 10^{-7}$ and the blue-shift for it entering our galaxy's gravitational well to the place where our Sun is $\sim -6 \times 10^{-7}$. Thus we need accuracies better than 10^{-6} to see the red-shift of stars and better than 10^{-7} to see the red-shift of galaxies. Due to the line-broadening of stellar spectra (because of the thermal motion of stellar matter in the photo-sphere causing a Doppler shift either side) there are too many uncertainties for this to provide a reliable test of GR. Of course, once the effect is established it could provide a valuable tool for studying stars and galaxies.

To find a more reliable test we will consider a purely terrestrial experiment. Light is sent up from the surface of the Earth, at distance R_E from the centre, to a distance d above it. From Eq.(4.101) here

$$\frac{\delta\nu}{\nu} = \left(1 - \frac{2Gm_E}{c^2 R_E}\right)^{1/2} \left[1 - \frac{2Gm_E}{c^2\,(R_E + d)}\right]^{-1/2} - 1\ . \tag{4.104}$$

To lowest order in Gm_E/c^2R_E and d/R_E this gives

$$\frac{\delta\nu}{\nu} \approx -\frac{2Gm_E}{c^2 R_E}\frac{d}{R_E}\ . \tag{4.105}$$

Taking $d \approx 70m$, $d/R_E \sim 10^{-5}$. Thus $\delta\nu \approx -8 \times 10^{-15}$. Frequency shifts a hundred times less can be measured by the Mössbauer effect. This effect was tested by Pound and Rebka [22] in 1959 (well after Einstein had died). The prediction holds within the limits of experimental error. Due to the high accuracies required and the corresponding technological problems this was the last of the three tests to be performed.

The above result agrees exactly with the prediction of Special Relativity. The question arises whether we could experimentally distinguish between the predictions of the Special and General theories. The difference would arise at a higher order in either of the parameters, G and d, taken to lowest order so far. Clearly

$$\frac{d}{R_E} \gg \frac{Gm_E}{c^2 R_E}\ . \tag{4.106}$$

Thus we can continue to neglect higher order terms in the latter and go to next order in the former parameter. Expanding Eq.(4.104) by the binomial expansion we get

$$\begin{aligned}\left(\frac{\delta\nu}{\nu}\right)_{GR} &= \left[1 - \frac{Gm_E}{c^2R_E} - \frac{G^2m_E^2}{2c^4R_E^2} - \mathcal{O}\left(\frac{Gm_E}{c^2R_E}\right)^3\right]\left[1 + \frac{Gm_E}{c^2\,(R_E+d)}\right. \\ &\quad \left. + \frac{3G^2m_E^2}{2c^4\,(R_E+d)^2} + \mathcal{O}\left(\frac{Gm_E}{c^2R_E}\right)^3\right] - 1 \\ &= \left.\begin{aligned}&\frac{Gm_E}{c^2\,(R_E+d)} - \frac{Gm_E}{c^2R_E} + \frac{G^2m_E^2}{2c^4R_E^2}\left[\frac{3R_E^2}{(R_E+d)^2} - 1\right] \\ &+ \left(\frac{Gm_E}{c^2R_E}\right)^3\ .\end{aligned}\right\}\end{aligned} \tag{4.107}$$

Now, from SR it can be seen that for Special Relativity

$$\left(\frac{\delta\nu}{\nu}\right)_{SR} = \frac{Gm_E}{c^2\,(R_E+d)} - \frac{Gm_E}{c^2R_E}\ . \tag{4.108}$$

Thus we can distinguish experimentally between the two theories if we have an accuracy of

$$\left.\begin{aligned}\frac{(\delta\nu/\nu)_{GR}-(\delta\nu/\nu)_{SR}}{(\delta\nu/\nu)_{SR}} &= \frac{Gm_E}{2c^2R_E}\frac{\left(2R_E^2-2R_Ed-d^2\right)}{\left(R_E+d\right)^2}\Big/\left(\frac{-d}{R_E+d}\right)\\ &\quad+\mathcal{O}\left(\frac{Gm_E}{cm^2R_E}\right)^2\\ &\approx -\frac{Gm_E}{c^2R_E}\cdot\frac{R_E}{d}\sim -7\times10^{-5}\ .\end{aligned}\right\}\tag{4.109}$$

Since the effect itself $\sim 8\times10^{-15}$ the accuracy required $\sim 6\times10^{-19}$ in the measurement of $\delta\nu/\nu$. This is beyond the present limits of experimental accuracy.

4.9 The Gravitational Deflection of Light

For light $d\tau = 0$. Thus from Eq.(4.89) $h\to\infty$ for light and hence the first term on the right side in Eq.(4.98) vanishes. Thus, for light the equation of motion becomes

$$\frac{d^2u}{d\phi^2}+u=\frac{3Gm}{c^2}u^2\ .\tag{4.110}$$

This is a nonlinear differential equation and cannot be solved exactly by any of the standard techniques. Therefore it is solved iteratively. The procedure assumes that the term on the right is negligible compared with u, i.e. $3Gm/c^2r \ll 1$. As we saw, this result would hold for terrestrial experiments and for Solar System based experiments.

The simplest solution to this equation is the special case of a circular orbit, i.e. with r, and hence $u = constant$. In this case the first term on the left in Eq.(4.110) is zero. Hence *light will go into a circular orbit at* $u=\frac{3Gm}{c^2}u^2$, *or* $r=\frac{3Gm}{c^2}$! It is obvious that this is a thoroughly "GR result", and could not come from SR, or Newtonian gravity, as it relies on the nonlinearity of the equation.

First let us put the right side of Eq.(4.110) equal to zero. We denote the solution of this equation by u_0. Then

$$u_0=1/r_0=\cos\left(\phi-\phi_0\right)/R\ ,\tag{4.111}$$

where ϕ_0 and R are the two constants of integration of the second order linear differential equation solved. The sub-script zero for u and r refer to the "zero order approximation". Clearly this function approximates the solution of Eq.(4.110).

We can obtain a better approximation, the "first order approximation", by solving the equation with u_0 replacing u in the right side of Eq.(4.110), i.e.

$$\frac{d^2u_1}{d\phi^2}+u_1=\frac{3Gm}{c^2R^2}\cos^2\left(\phi-\phi_0\right)\ .\tag{4.112}$$

We can now solve Eq.(4.112) as an inhomogeneous, second order, linear, ordinary differential equation. The complementary function is just $u_0(\phi)$ given by Eq.(4.111). Since

$$\cos^2\left(\phi-\phi_0\right)=\frac{1}{2}\left[\cos\left(2\phi-2\phi_0\right)+1\right]\ ,\tag{4.113}$$

and

$$\begin{aligned} \left(D^2+1\right)^{-1}\cos\left(2\phi-2\phi_0\right) &= \left(-4+1\right)^{-1}\cos\left(2\phi-2\phi_0\right)\ , \\ \left(D^2+1\right)^{-1}1 &= 1\ , \end{aligned} \tag{4.114}$$

the particular integral of Eq.(4.112) is

$$\begin{aligned} u_{1PI}\left(\phi\right) &= \frac{3Gm}{c^2R^2}\frac{1}{\left(D^2+1\right)}\cos^2\left(\phi-\phi_0\right) \\ &= -\frac{Gm}{c^2R^2}\frac{1}{2}\left[\cos\left(2\phi-2\phi_0\right)-3\right] \\ &= \frac{Gm}{c^2R^2}\left[\cos^2\left(\phi-\phi_0\right)+2\sin^2\left(\phi-\phi_0\right)\right]\ . \end{aligned} \tag{4.115}$$

Hence the first order approximation to the solution of Eq.(4.110) is

$$u_1\left(\phi\right)=\frac{1}{r_1}=\frac{1}{R}\cos\left(\phi-\phi_0\right)+\frac{Gm}{c^2R^2}\left[\cos^2\left(\phi-\phi_0\right)+2\sin^2\left(\phi-\phi_0\right)\right]\ . \tag{4.116}$$

This procedure can be continued to the second order approximation, which would satisfy the equation

$$\frac{d^2u_2}{d\phi^2}+u_2=\frac{3Gm}{c^2}u_1^2\ , \tag{4.117}$$

and the third order and so on to the n^{th} order approximation which has u_n on the left side of the equation and u_{n-1}^2 on the right side. A final solution would be the limit of the sequence of functions u_0, u_1, ..., u_{n-1}, u_n, ..., provided it converges. The limiting function, u, will be the solution of Eq.(4.110). Remember that here the convergence is in the sense of functions and not of numbers.

For practical purposes we do not need to find the limiting function, so long as we know that the sequence converges and that the higher order corrections are negligible compared with the accuracy of the measurement. This is the case for our present purpose. We can evaluate u_0 to get the classical path of the light, R, being the "impact parameter", and ϕ_0 the choice of phase, see Fig. 4.4. The experiment takes light from a distant star grazing by the Sun and observed at the Earth. Now, evaluating u_1 gives the "corrected" path of the light and hence gives the gravitational deflection of the light. Put $x=r_1\cos\left(\phi-\phi_0\right)$ and $y=r_1\sin\left(\phi-\phi_0\right)$ in Eq.(4.116). Thus $r_1=\sqrt{x^2+y^2}$. Multiplying Eq.(4.116) by Rr_1 and transposing

$$x=R-\frac{2Gm}{c^2R}\frac{x^2+2y^2}{\sqrt{x^2+y^2}}\ . \tag{4.118}$$

It is clear from Fig. 4.4 that $x\sim R$, which is the Solar radius, while $y\sim r_1$, which is the approximate distance of the Earth from the Sun. Hence $x\ll y$. Thus the deflection of the light is given by

$$\theta\approx\tan\theta=\frac{R-x}{r_1}\approx-\frac{2Gm}{c^2R}\ . \tag{4.119}$$

The SR result is exactly half that given by Eq.(4.119). The reason for the difference is that here distances are measured along space-like geodesics while

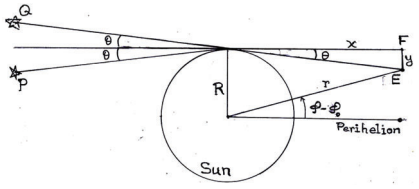


Figure 4.4: Light from a distant star at P grazes by the Sun at a distance R equal to the radius of the Sun, from the Solar centre. Instead of going on to point F, at a distance x from the point where it grazed by the Sun, it is received at the Earth at a distance r. The light is clearly deflected a distance $x = FE$. The angle of deflection θ is given by $\tan\theta = x/y$. Clearly, $y \sim r$ and $x \ll R \ll r$ and so $\theta \approx \tan\theta$. The star appears to be at point Q. Thus the apparent deflection is through 2θ.

the special relativistic analysis uses null lines for the purpose. The experiment is discussed in SR. Observations are made during a Solar eclipse of a star that should have been hidden behind the disc of the Sun but is visible due to the bending of light. As is clear from Fig. 4.4, the observed angular displacement is 2θ and not θ. As explained in section 7 the test could not be performed till after the end of World War I. In 1919 Sir Arthur Eddington led an expedition to South America where there was to be a total Solar eclipse. The observations confirmed the predictions of GR to a sufficient accuracy to be able to exclude the SR prediction.

The confirmation of GR led to great fame for Einstein and Eddington and that fame appears to have gone to Eddington's head. The story goes that a reporter asked Eddington whether it was true that only three people understood Relativity. Eddington did not reply, so the reporter repeated the question. Eddington said: "I was thinking. Who is the third?" The mathematical nature of the theory and its heavy dependence on Geometry supported the impression put across by people of Eddington's stamp, that General Relativity was incomprehensible. As should be clear from this book that is by no means the case. Also, nowadays many branches of Theoretical Physics use equally (or more) profound Mathematics.

4.10 The Perihelion Shift of Mercury

Historically this was the first test of GR. It was not a test in the sense that the prediction preceded the observation, but that GR was not designed to explain the observation. This was the first non-Newtonian prediction of the theory and it happened to fit observation precisely. In fact, when Einstein was told that his prediction of the deflection of light had been verified and the reporter asked him if he was not excited, he said: "No. I knew that it would work." This certainty was based on the fact that the theory had given the right answer for an outstanding problem. When the reporter asked, "But what if it had not been verified?", he answered "Then I would have been sorry for the Good Lord."

Consider Eq.(4.98) for the orbit of a material test particle (which therefore moves on a time-like geodesic) in a nearly classical regime. In this case the second term on the right of the equation will be very small. Following the

procedure adopted in the previous section, u_0 will be given by Eq.(4.81) with θ and θ_0 replaced by ϕ and ϕ_0. Then

$$\frac{du_1}{d\phi^2} + u_1 = \frac{Gm}{h^2} + \frac{3Gm}{c^2}\frac{G^2m^2}{h^4}\left[1 + \varepsilon\cos\left(\phi - \phi_0\right)\right]^2 \ . \tag{4.120}$$

Since ε is small for all natural planets (as opposed to space-craft put into orbit) we can neglect terms in ε^2, as they will merely give a correction to a correction. Defining

$$U = u_1 - \frac{Gm}{h^2} - \frac{3G^3m^3}{c^2h^4} \ , \tag{4.121}$$

Eq.(4.120) can be re-written as

$$\frac{dU}{d\phi^2} + U = \frac{6G^3m^3}{c^2h^2}\varepsilon\cos\left(\phi - \phi_0\right) \ . \tag{4.122}$$

Using the inverse operator method of solving ordinary differential equations, since $(D^2+1)^{-1}\cos(\phi-\phi_0)$ can not be solved by replacing D^2 by -1 (as that would give infinity) we apply the shift theorem to get

$$(D^2+1)^{-1}\cos(\phi-\phi_0) = \frac{1}{2}(\phi-\phi_0)\sin(\phi-\phi_0) \ . \tag{4.123}$$

One could, instead, have used the comparison of coefficients to get the same answer. Essentially, one of the derivative operators acts on the *sin*, while the other acts on the linear term, or both can act on the *sin*. One of the expressions cancels, leaving only the other. We get the factor of half, since either operator could act on either part. This result is easily verified by inserting the expression on the right side of Eq.(4.123) in place of U on the left side of Eq.(4.122). Using this value for the particular integral of U and using the previous solution (u_0) for the complementary function, Eq.(4.121) gives

$$u_1\left(\phi\right) = \frac{Gm}{h^2}\left[1 + \varepsilon\cos\left(\phi - \phi_0\right)\right] + \frac{3G^3m^3}{c^2h^4} + \frac{3G^3m^3}{c^2h^4}(\phi-\phi_0)\sin(\phi-\phi_0) \ . \tag{4.124}$$

The first term on the right side of Eq.(4.124) is the usual solution of the equations for planetary motion. The second term gives a very small distance correction. Since we use the planetary orbit to estimate the mass of the Sun, this "correction" would essentially modify our value for the Solar mass. It does not, therefore, provide a testable prediction. Now

$$\cos\left(\phi - \phi_0 - \delta\phi\right) = \cos\left(\phi - \phi_0\right)\cos\delta\phi + \sin(\phi - \phi_0)\sin\delta\phi \ . \tag{4.125}$$

Thus to first order in $\delta\phi$, for $\delta\phi \ll 1$,

$$\cos\left(\phi - \phi_0 - \delta\phi\right) \approx \cos\left(\phi - \phi_0\right) + \delta\phi\sin(\phi - \phi_0) \ . \tag{4.126}$$

Hence Eq.(4.124) can be rewritten as

$$u_1\left(\phi\right) \approx \frac{Gm}{h^2}\left[1 + \varepsilon\cos\left(\phi - \phi_0 - \delta\phi\right)\right] + \frac{3G^3m^3}{c^2h^2} \ , \tag{4.127}$$

where

$$\delta\phi = \frac{3G^2m^2}{c^2h^2}\left(\phi - \phi_0\right) \ . \tag{4.128}$$

Using Eq.(4.82) to replace h^2 in term of the semi-major axis, a, Eq.(4.128) reduces to

$$\delta\phi = \frac{6\pi Gm}{c^2 a\,(1-\varepsilon^2)} \text{ rad/revolution} , \tag{4.129}$$

by putting $\phi - \phi_0 = 2\pi$. From Eq.(4.78) it is clear that classically the rate at which an area is swept out by the orbiting particle is the constant $h/2$ (Kepler's second law of planetary motion). Since the area of an ellipse is πab and it is swept out in one time period, T,

$$\frac{1}{2}hT = \pi ab , \tag{4.130}$$

where b is the semi-minor axis. Thus

$$b = a\sqrt{1-\varepsilon^2} . \tag{4.131}$$

Using Eqs.(4.82), (4.130) and (4.131) we get Kepler's third law of planetary motion,

$$T^2 = \frac{4\pi^2 a^3}{Gm} . \tag{4.132}$$

Eq.(4.132) can be used to replace Gm in Eq.(4.129) and obtain

$$\delta\phi = \frac{24\pi^2}{1-\varepsilon^2}\frac{a^2/T^2}{c^2} \text{ rad/revolution} . \tag{4.133}$$

Since $2\pi a/T$ is the average speed of the planet, we can write the perihelion shift in terms of $\bar{v}/c$ as

$$\delta\phi = \frac{6\pi}{1-\varepsilon^2}\left(\frac{\bar{v}}{c}\right)^2 \text{ rad/revolution} . \tag{4.134}$$

The advantage of re-casting Eq.(4.128) in this form is that other uncertainties and inaccuracies are avoided. To the order of approximation used, $\delta\phi$ is limited in accuracy only to twice the possible error in $\bar{v}$, which is extremely small. Hence we know $\delta\phi$ quite adequately for the purpose of testing the theory.

For Mercury $\varepsilon = .206$ and its average orbital speed is $\bar{v} \approx 4.8 \times 10^4$ m/sec. Eq.(4.134) gives $\delta\phi \approx 5.0 \times 10^{-7}$ rad/revolution. Since Mercury takes ~ 88 days this gives $\delta\phi \approx 43''$ arc/century, which is exactly the amount required to explain the observed discrepancy! This was the most striking possible confirmation of the theory. The table below gives the status of observations for some of the planets and a planetoid, Icarus.

Planet	$\Delta\delta$ **(theory)**	$\Delta\phi$**(observed)**	ε
Mercury	43.03	43.11 ± 0.45	0.2065
Venus	8.63	8.4 ± 4.8	0.0068
Earth	3.84	5.0 ± 1.2	0.017
Mars	1.35	—	0.193
Jupiter	0.06	—	0.050
Icarus	10.0	9.8 ± 0.8	0.827

Since the early days there have been much better tests that became available. The light from distant quasi-stellar objects, called *QSOs* or *quasars* gets bent by

intervening galaxies. leading to double, triple or more images. This effect has been seen many times over. Einstein had predicted that the image may come in the form of a cross (called the Einstein cross) or even a ring (called, oddly enough, an Einstein ring). One can visualise the effect by the analogy of seeing an object through a bowl or vase containing water. These have also been seen. These effects are collectively called *gravitational lensing*. What Einstein had not predicted was the possibility of the focussing of light from a star due to an intervening condensed object, like planets. This process is called *microlensing*. This has been used to find planets outside the solar system; freely floating "planets" not associated with a star; and even planets in our neighbouring galaxy, M31. Strong bending is caused by stars that have undergone total gravitational collapse, which are called black holes. These objects will be discussed in Chapter 6.

The equivalent of the perihelion shift has been seen by observing binary pulsars (consisting of a pulsar in a binary pair with some other gravitationally collapsed object, like a white dwarf or another pulsar). The Hulse-Taylor binary pulsar, using the parameterised post-Newtonian (PPN) formalism [15], provided a laboratory to test GR against other, competing, theories of gravity. (Since Einstein provided the do-it-yourself their own theories.) The tests exclude all competing theories to the extent that the theories have to be made observationally indistinguishable from GR to fit with the observations.

4.11 Exercises

1. Construct the Ricci tensor for the class of plane-symmetric metrics:

$$ds^2 = A^2(t,x)c^2dt^2 - B^2(t,x)dx^2 - C^2(t,x)(dy^2 + dz^2) \ ,$$

and hence obtain the Einstein equations for a vacuum plane-symmetric static spacetime. Try to solve the equations. Now drop the requirement that it be static and try to solve it.

2. The most general cylindrically-symmetric metric is:

$$ds^2 = e^{2(\gamma-\psi)}(c^2dt^2 - d\rho^2) - e^{-2\psi}\rho^2 d\varphi^2 - e^{2\psi}dz^2 \ ,$$

where γ and ψ are arbitrary functions of (t,ρ). Construct the Ricci tensor for it and hence obtain the Einstein equations for a vacuum cylindrically-symmetric static spacetime. Solve the equations. Now drop the requirement that it be static and solve the equations.

3. Construct a stress-energy tensor density for a massive scalar field, with Lagrange density $\mathcal{L} = q\sqrt{-g}\phi + \frac{1}{2}m^2\phi^2 + g^{\mu\nu}\phi_{,\mu}\phi_{,\nu}$. Obtain the conservation law for the scalar field. Now construct the coupled scalar-Einstein field equations. What difference is made to the conservation law by the coupling? Can one obtain a static spherically symmetric solution for the coupled system with no other stress-energy tensor than the scalar field?

4. Construct a stress-energy tensor density for a massive spin one field, the Proca field, i.e. with Lagrange density $\mathcal{L} = qA_\mu j^\mu + \frac{1}{2}m^2 A_\mu A^\mu + 4g^{\mu\rho}g^{\nu\pi}A_{[\mu,\nu]}A_{[\rho,\pi]}$, where j^μ is the current density for the Proca field. Obtain the conservation law for the Proca field. Now construct the coupled Einstein-Proca field equations. What difference is made to the conservation law by the coupling? Can one obtain a static spherically symmetric solution for the coupled system with no other stress-energy tensor than the Proca field?

5. Construct a stress-energy tensor density for a massless spin one field, with an "internal symmetry" with structure constants C^i_{jk}, the Yang-Mills field, with the field tensor $F^i_{\mu\nu} = C^i_{jk}A^j_\mu \wedge A^k_\nu + 2A^i_{[\nu,\mu]}$ and Lagrange density $\mathcal{L} = qA^i_\mu j^\mu_i \frac{1}{16\pi}F^i_{\mu\nu}F^{\mu\nu}_i$, where j^μ_i is the current density for the Yang-Mills field. Obtain the conservation law for the Yang-Mills field. Now construct the coupled Einstein-Yang-Mills field equations. What difference is made to the conservation law by the coupling? Can one obtain a static spherically symmetric solution for the coupled system with no other stress-energy tensor than the Yang-Mills field?

6. Why is GR, as presented so far, *not* a field theory? What is missing?

Chapter 5

General Relativity as a Field Theory of Gravity

We now proceed with the original derivation of the Einstein field equations from the Einstein-Hilbert Lagrangian density. This derivation is better adapted to the incorporation of other fields than gravity, into General Relativity. There has been no successful unification of gravity with other fields so far. However, no attempt at unification, or quantization for that matter, can afford to ignore field theory, even if it attempts to avoid it. In this chapter we will not only re-derive the Einstein field equations but also see how they can be extended to deal with fields coupled with gravity, such as the Einstein-Maxwell fields, for example. We will also discuss some of the more well known solutions to the field equations.

5.1 Re-derivation of the Einstein Field Equations

The action we are to deal with, S, will generally consist of a matter part, S_m, and a part for gravity, S_G, apart from the action for other fields that are present, such as S_{em} for electromagnetism, for example. We start with

$$S = S_m + S_G \ . \tag{5.1}$$

The matter action involves the metric tensor as

$$S_m = \int_\Omega T^{(\text{matter})} d\Omega = \int_\Omega T^{(\text{matter})}_{\mu\nu} g^{\mu\nu} d\Omega \ , \tag{5.2}$$

where $T^{(\text{matter})}_{\mu\nu}$ is the stress-energy tensor referred to in section 4.1. In the absence of any fields other than gravity this tensor will be the total stress-energy tensor. Hence we can drop the "matter" superscript in what follows. Let the action be varied due to a variation in $g^{\mu\nu}$, $\delta g^{\mu\nu}$. Now the total action must remain invariant. Hence

$$0 = \delta S = \delta S_m + \delta S_G = \int_\Omega T_{\mu\nu} \delta g^{\mu\nu} d\Omega + \delta S_G \ . \tag{5.3}$$

The action for gravity, independently proposed by Einstein and Hilbert, is the purely geometric quantity

$$S_G = -\gamma \int_\Omega \sqrt{|g|} R d\Omega \ , \tag{5.4}$$

where R is the Ricci scalar and γ an arbitrary constant that will be involved in the coupling of gravity to matter. Thus the Lagrangian density for gravity is

$$\left.\begin{aligned} \mathcal{L}_G[g^{\mu\nu}, g^{\mu\nu}_{,\rho}] \ &= \sqrt{|g|} R = \sqrt{|g|} g^{\mu\nu} R_{\mu\nu} \\ &= \sqrt{|g|} g^{\mu\nu} \Big[\left\{ \mu \ \ \rho \ \ \nu \right\}_{,\rho} - \left\{ \mu \ \ \rho \ \ \rho \right\}_{,\nu} + \left\{ \pi \ \ \rho \ \ \rho \right\} \left\{ \mu \ \ \pi \ \ \nu \right\} \\ &\quad - \left\{ \pi \ \ \rho \ \ \mu \right\} \left\{ \rho \ \ \pi \ \ \nu \right\} \Big] \ . \end{aligned} \right\} \tag{5.5}$$

The above Lagrangian density depends not only on $g^{\mu\nu}$ and combinations of its derivatives, in the form of Christoffel symbols, but *their* derivatives as well. However, as will be demonstrated shortly, the extra derivative terms can be eliminated by using a standard trick of field theory. These terms will be expressed as a whole divergence. Using the generalised Gauss theorem this divergence can be expressed as an integral over the hypersurface bounding the four volume. This "surface term" can be regarded as negligible compared with the bulk action (to the extent that the surface energy density can be regarded as negligible compared with the bulk energy density for a sufficiently large volume). The argument may be understood in terms of its 3-dimensional analogue for a sphere, where the volume $V \sim A^{3/2}$, A being the surface area of the sphere. Thus as $r \to \infty$ the volume terms dominate over the area terms.

To remove the higher derivative terms we first write

$$\left.\begin{aligned} 2[\sqrt{|g|} g^{\mu\nu} \left\{ \mu \ \ \rho \ \ [\nu \right\}_{,\rho]}] = &\sqrt{|g|} g^{\mu\nu} \left[\left\{ \mu \ \ \rho \ \ \nu \right\}_{,\rho} - \left\{ \mu \ \ \rho \ \ \rho \right\}_{,\nu} \right] \\ &+ 2 \left[\sqrt{|g|} g^{\mu\nu} \right]_{,[\rho} \left\{ \nu] \ \ \rho \ \ \mu \right\} \ . \end{aligned} \right\} \tag{5.6}$$

Hence the higher derivative terms in Eq.(5.5) can be re-expressed as

$$\left.\begin{aligned} \sqrt{|g|} g^{\mu\nu} \left[\left\{ \mu \ \ \rho \ \ \nu \right\}_{,\rho} - \left\{ \mu \ \ \rho \ \ \rho \right\}_{,\nu} \right] = &2 \Big[\sqrt{|g|} g^{\mu\nu} \left\{ \mu \ \ \rho \ \ [\nu \right\}_{,\rho]} \Big] \\ &+ 2 \left[\sqrt{|g|} g^{\mu\nu} \right]_{,[\nu} \left\{ \rho] \ \ \rho \ \ \mu \right\} \ . \end{aligned} \right\} \tag{5.7}$$

The first term on the right can be written as a whole divergence

$$\varsigma^\rho_{;\rho} = \left(\sqrt{|g|} \varsigma^\rho \right)_{,\rho} \ . \tag{5.8}$$

By the generalised Gauss divergence theorem we have

$$\int_\Omega \varsigma^\rho_{;\rho} d\Omega = \oint_{\partial\Omega} \varsigma^\rho d\Sigma_\rho \ , \tag{5.9}$$

where $d\Sigma_\rho$ is the 3-dimensional hypersurface volume element and $\partial\Omega$ is the boundary of Ω depicted in Fig. 1.2. As just explained above this term can be neglected compared with the rest of the action. Hence we can leave the

first two terms on the right hand side of Eq.(5.5) out of the Lagrangian density and replace them by the right hand side of Eq.(5.7), to obtain the effective Lagrangian density

$$\left.\begin{aligned}\mathcal{L}_G[g^{\mu\nu}, g^{\mu\nu}_{,\rho}] = &\left(\sqrt{|g|}g^{\mu\nu}\right)_{,\nu}\left\{\rho\ \ \mu\right\} - \left(\sqrt{|g|}g^{\mu\nu}\right)_{,\rho}\left\{\mu\ \ \nu\right\} \\ &+ \sqrt{|g|}g^{\mu\nu}\left[\left\{\rho\ \ \pi\right\}\left\{\mu\ \ \nu\right\} - \left\{\pi\ \ \mu\right\}\left\{\rho\ \ \nu\right\}\right] ,\end{aligned}\right\} \tag{5.10}$$

having neglected the surface term.

The expressions appearing in Eq.(5.10) can be further reduced by using a couple of identities. First notice that

$$\sqrt{|g|}_{,\rho} = \sqrt{|g|}\left(\ln\sqrt{|g|}\right)_{,\rho} = \sqrt{|g|}\left\{\pi\ \ \rho\right\} . \tag{5.11}$$

Next (using the symmetriser brackets of section 3.4) consider the expression

$$2\left\{\pi\ \ {}^{(\mu}\ \ \rho\right\} g^{\nu)\pi} = g^{\sigma(\mu}g^{\nu)\pi}\left(g_{\sigma\pi,\rho} + g_{\sigma\rho,\pi} - g_{\rho\pi,\sigma}\right) . \tag{5.12}$$

The last two terms inside the brackets on the right hand side of Eq.(5.12) are skew-symmetric in σ and π, while the product of the g's outside the brackets is explicity symmetric in μ and ν and hence is symmetric in σ and π (as the g's are symmetric tensors). Thus those two terms cancel out. The remaining expression is explicitly symmetric in σ and π (and hence in μ and ν). Therefore the symmetriser brackets can be dropped. Further, since

$$0 = \left(\delta^\nu_\sigma\right)_{,\rho} = \left(g^{\nu\pi}g_{\sigma\pi}\right)_{,\rho} = g^{\nu\pi}_{,\rho}g_{\sigma\pi} + g^{\nu\pi}g_{\sigma\pi,\rho} , \tag{5.13}$$

we can shift the derivative from $g_{\sigma\pi}$ to $g^{\nu\pi}$ and introduce a negative sign. Also, since $g^{\sigma\mu}g_{\sigma\pi} = \delta^\mu_\pi$, we finally have

$$g^{\mu\nu}_{,\rho} = -\left\{\pi\ \ \mu\ \ \rho\right\} g^{\nu\pi} - \left\{\pi\ \ \nu\ \ \rho\right\} g^{\mu\pi} . \tag{5.14}$$

Using the identities given by Eqs.(5.11) and (5.14) in (5.10) we get

$$\left.\begin{aligned}\mathcal{L}_G[g^{\mu\nu}, g^{\mu\nu}_{,\rho}] = &\sqrt{|g|}[g^{\mu\nu}\left\{\pi\ \ \nu\right\}\left\{\rho\ \ \mu\right\} - g^{\nu\pi}\left\{\pi\ \ \mu\ \ \nu\right\}\left\{\rho\ \ \mu\right\} \\ &- g^{\mu\pi}\left\{\pi\ \ \nu\ \ \nu\right\}\left\{\rho\ \ \mu\right\} - g^{\mu\nu}\left\{\pi\ \ \pi\ \ \rho\right\}\left\{\mu\ \ \nu\right\} \\ &+ g^{\nu\pi}\left\{\pi\ \ \mu\ \ \rho\right\}\left\{\nu\ \ \mu\right\} + g^{\mu\pi}\left\{\pi\ \ \nu\ \ \rho\right\}\left\{\nu\ \ \mu\right\} \\ &+ g^{\mu\nu}\left\{\pi\ \ \rho\ \ \rho\right\}\left\{\mu\ \ \nu\right\} - g^{\mu\nu}\left\{\pi\ \ \rho\ \ \mu\right\}\left\{\rho\ \ \nu\right\}] .\end{aligned}\right\} \tag{5.15}$$

By interchanging dummy indices it is easy to see that the first and third terms cancel, as do the second and seventh terms and the fifth and eighth terms, respectively. The effective gravitational Lagrangian density then becomes

$$\mathcal{L}_G[g^{\mu\nu}, g^{\mu\nu}_{,\rho}] = -\sqrt{|g|}g^{\mu\nu}[\left\{\mu\ \ \rho\ \ \pi\right\}\left\{\nu\ \ \pi\ \ \rho\right\} - \left\{\pi\ \ \pi\ \ \rho\right\}\left\{\mu\ \ \rho\ \ \nu\right\}] . \tag{5.16}$$

Hence, even though the Lagrangian density is highly non-linear, it depends effectively only on the metric tensor and its first derivatives, as required for the Euler-Lagrange equations to apply.

Now consider the variation of the gravitational Lagrangian, S_G. This will consist of two parts, one due to the variation of $\sqrt{|g|}$ and the other due to the variation of R. Now

$$\delta\sqrt{|g|} = -\frac{1}{2}\sqrt{|g|}g_{\mu\nu}\delta g^{\mu\nu} , \tag{5.17}$$

as may be verified by multiplying both sides by $2\sqrt{|g|}$. Further, since $R = R_{\mu\nu}g^{\mu\nu}$, we have

$$\left.\begin{aligned}\delta S_G &= -\gamma\delta\int_\Omega \sqrt{|g|}R_{\mu\nu}g^{\mu\nu}d\Omega \\ &= -\gamma\int_\Omega \left[\left(\delta\sqrt{|g|}\right)R_{\mu\nu}g^{\mu\nu} + \sqrt{|g|}\delta R_{\mu\nu}g^{\mu\nu} + \sqrt{|g|}R_{\mu\nu}\delta g^{\mu\nu}\right]d\Omega \\ &= -\gamma\int_\Omega \sqrt{|g|}\left(R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu}\right)\delta g^{\mu\nu}d\Omega - \gamma\int_\Omega \sqrt{|g|}g^{\mu\nu}\delta R_{\mu\nu}d\Omega\ .\end{aligned}\right\} \quad (5.18)$$

The last term may be eliminated by using Reimann normal coordinates. This is a frame in which the Christoffel symbol is made negligibly small, but its derivative may be non-negligible. To follow this point consider the function $(x-a)^2$. By changing the coordinate system to $x' = x - a$, we see that near the origin the first derivative, $2x'$, is negligible but the second derivative, 2, is not. Similarly, for a surface we can consider the tangent plane to the point under consideration and use coordinates along two independent vectors in the plane. To first order the slope of the surface relative to the tangent plane is zero, but to second order it is the normal curvature. Riemann normal coordinates give the first derivative of the metric tensor, at the point, nearly zero but the second derivative gives the curvature (which can not be made zero by a choice of coordinates). Thus $\{_\nu{}^\mu{}_\rho\} \approx 0$ but $\{_\nu{}^\mu{}_\rho\}_{,\pi} \neq 0$ in general. Hence

$$g^{\mu\nu}\delta R_{\mu\nu} = g^{\mu\nu}\delta\left[\{_\mu{}^\rho{}_\nu\}_{,\rho} - \{_\mu{}^\rho{}_\rho\}_{,\nu}\right]\ . \quad (5.19)$$

We can express this variation as a whole divergence and some additional terms, as before, by interchanging the ν and μ in the second term

$$\begin{aligned} g^{\mu\nu}\delta R_{\mu\nu} &= [g^{\mu\nu}\delta\{_\mu{}^\rho{}_\nu\} - g^{\mu\rho}\delta\{_\mu{}^\nu{}_\nu\}]_{,\rho} \\ &\quad - \left[g^{\mu\nu}_{,\rho}\delta\{_\mu{}^\rho{}_\nu\} - g^{\mu\rho}_{,\rho}\delta\{_\mu{}^\nu{}_\nu\}\right]\ . \end{aligned} \quad (5.20)$$

The expression in the second bracket on the right hand side is negligible because each of the terms involves the first derivative of the metric tensor. The first bracket is not negligible on this account as it involves the *variation* of the first derivative and is then differentiated. The term can be converted to a covariant divergence by multiplying and dividing by $\sqrt{|g|}$. The numerator can be taken inside the derivative as the additional term is negligible in Riemann normal coordinates. The total divergence can be converted to a surface term and removed (as explained before). Thus

$$\delta S_G = -\gamma\int_\Omega \left(R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu}\right)\delta g^{\mu\nu}d\Omega \equiv -\gamma\int_\Omega \mathcal{E}_{\mu\nu}\delta g^{\mu\nu}d\Omega\ . \quad (5.21)$$

Taking Eqs.(5.3) and (5.21) together by putting $\gamma = 1/\kappa$, we get the Einstein field equation

$$\mathcal{E}_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} = \kappa T_{\mu\nu}\ . \quad (5.22)$$

By re-defining the zero of the potential energy we would not change the classical gravitational field. However, that does not hold here. We can add a constant to the action and thereby modify the gravitational Lagrangian density to

$$\mathcal{L}_G[g^{\mu\nu}, g^{\mu\nu}_{,\rho}] = \sqrt{|g|}\,(R + 2\Lambda)\ , \quad (5.23)$$

where Λ is a constant (called the *cosmological constant*). Following the above procedure we get the Einstein equations with a cosmological term

$$R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} - \Lambda g_{\mu\nu} = \kappa T_{\mu\nu} \ . \tag{5.24}$$

Eqs.(5.22) and (5.24) are the equations derived in the last chapter, Eqs.(4.29) and (4.30). Till recently there was no evidence for a non-negligible Λ, which was therefore normally taken to be zero. It is worth-mentioning that not only was Λ used in deriving the first models of the Universe, it plays an important part in modern cosmology as well. On the other hand it is a serious embarrassment in attempts for theories of Quantum Gravity. We will return to it in Chapter 7 , on Cosmology.

5.2 The Schwarzschild Interior Solution

The simplest case to consider, after the flat Minkowski space, was the case of a simple point gravitational source at the origin, which is clearly spherically symmetric and static. Using the vacuum Einstein field equations we obtained the Schwarzschild metric of section 4.5. The next simplest case is when there is a sphere (of radius a) centred at the origin and composed of an isotropic perfect fluid, i.e. one for which the density, ρ, is just a function of r and the stress-tensor given in Eq (4.1) is

$$\sigma^i_j = \delta^i_j p(r)/c^2 \ . \tag{5.25}$$

Thus the stress-energy tensor for the matter in the sphere becomes

$$T^\nu_\mu = \left[\rho(r) + p(r)/c^2\right] u^\nu u_\mu - \delta^\nu_\mu p(r)/c^2 \ , \tag{5.26}$$

with u^ν given by Eq (4.5) and the metric tensor by Eq (4.4). Hence $u^\nu u_\mu = \delta^\nu_0 \delta^0_\mu$.

Rather than proceeding with the Einstein field equations here, it is convenient to use the energy conservation equation (which is equivalent to using the Bianchi identities) $T^\nu_{\mu;\nu} = 0$. Hence

$$\left.\begin{aligned} 0 = \left[\rho(r) + p(r)/c^2\right]^{\cdot} \delta^0_\mu + \left[\rho(r) + p(r)/c^2\right] \left(\ln\sqrt{|g|}\right)^{\cdot} \delta^0_\mu \\ - \left[p(r)/c^2\right]_{,\mu} - \left[\rho(r) + p(r)/c^2\right] \left\{{}_0{}^0{}_\mu\right\} \ . \end{aligned}\right\} \tag{5.27}$$

The first and second terms on the right hand side are automatically zero, as no time dependent functions are involved. Similarly, even the latter two terms are identically zero for $\mu \neq 1$. Thus three of the four conservation equations are identities and only the fourth, for $\mu = 1$, gives a new equation. Eqs (4.66) give

$$\frac{1}{2}\left[\rho(r)c^2 + p(r)\right] v\nu'(r) + p'(r) = 0 \ . \tag{5.28}$$

Using Eqs. (4.67), (4.68), (4.71) and (4.73) we can evaluate $R_{\mu\nu}$ and hence R and use them to obtain the Einstein tensor. Then Eqs.(5.22) for $\mu = 0 = \nu$ and $\mu = 1 = \nu$, respectively, yield

$$\left[\frac{\lambda'(r)}{r} - \frac{1}{r^2}\right] e^{-\lambda(r)} + \frac{1}{r^2} = \frac{8\pi G}{c^2}\rho(r) \ , \tag{5.29}$$

$$\left[\frac{\nu'(r)}{r} + \frac{1}{r^2}\right] e^{-\lambda(r)} - \frac{1}{r^2} = -\frac{8\pi G}{c^4}p(r) \ . \tag{5.30}$$

For given $\rho(r)$ and $p(r)$ we would have three equations for two functions. Thus there would generally be no solution. On the other hand, if none of them is given the system seems to be under-determined. What is required is one additional equation. This will generally be provided by an "equation of state", which is an equation relating $\rho(r)$ and $p(r)$, such as the ideal gas equation for a given temperature (*isothermal* gas) $p(r) \propto \rho(r)$. Notice that the system of equations is non-linear. Eq.(5.29) can be regarded as linear in the function $e^{-\lambda(r)}$, but the other two equations involve products of the derivative of one function with other functions. As such it may still happen that there are limitations on the permissible functions.

Schwarzschild obtained a solution for a constant density case, $\rho(r) = \rho_0$. Now Eq.(5.28) can be integrated to give

$$ke^{-\frac{1}{2}\nu(r)} = \rho_0 + p(r)/c^2 \ , \tag{5.31}$$

where k is an integration constant. Integrating Eq.(5.29) yields

$$e^{-\lambda(r)} = 1 - 8\pi G\rho_0 r^2/3c^2 + \alpha/r \ , \tag{5.32}$$

where α is another integration constant. Physically it is clearly necessary that there be no singularity at $r = 0$. However, there can be coordinate singularities where there is no *physical* singularity, such as the z-axis in Minkowski space using spherical coordinates. Our solution represents the interior of a sphere of radius a. Outside the sphere ($r > a$) we must have the Schwarzschild (exterior) solution of section 4.5. Requiring continuity at the boundary, $r = a$, we get

$$e^{-\lambda(a)} = 1 - 2Gm/c^2a = 1 - 8\pi G\rho_0 a^2/3c^2 \ . \tag{5.33}$$

Comparing Eq (5.33) with Eq.(5.32) for $r = a$, we see that, $\alpha = 0$.

We now need to eliminate $p(r)$ in Eqs.(5.30) and (5.31) to obtain $\nu(r)$. Adding Eqs.(5.29) and (5.30) and using Eq.(5.31) we get

$$\frac{16\pi G\rho_0}{3c^2} + \frac{\nu'(r)}{r} - \frac{8\pi G\rho_0 r\nu'(r)}{3c^2} = \frac{8\pi G}{c^2}k \ . \tag{5.34}$$

Writing P^2 in place of $3c^2/8\pi G\rho_0$ for convenience and multiplying Eq.(5.34) by $e^{-\frac{1}{2}\nu(r)}$, it can be re-written as

$$\left[e^{\frac{1}{2}\nu(r)}\right]' + \frac{r/P^2}{1 - r^2/P^2}\left[e^{\frac{1}{2}\nu(r)}\right] = \frac{3kr/\rho_0 P^2}{1 - r^2/P^2} \ . \tag{5.35}$$

Solving Eq.(5.35) by introducing an integrating factor we obtain

$$e^{\frac{1}{2}\nu(r)} = 3k/\rho_0 - A\sqrt{1 - r^2/P^2} \ , \tag{5.36}$$

where A is a constant of integration. Again, requiring continuity at $r = a$,

$$\left[3k/\rho_0 - A\sqrt{1 - a^2/P^2}\right]^2 = 1 - 2Gm/c^2a \ . \tag{5.37}$$

Using Eq.(5.37) along with the identity $a^2/P^2 = 2Gm/c^2a$, in Eq.(5.36) gives

$$e^{\nu(r)} = \left[\sqrt{1 - 2Gm/c^2a}\,(1 + A) - \sqrt{1 - 2Gmr^2/c^2a^3}\right]^2 \ . \tag{5.38}$$

Eqs.(5.32), $\alpha = 0$, and (5.38) yield the *Schwarzschild interior* solution

$$\left.\begin{aligned} ds^2 = &\left[\sqrt{1-2Gm/c^2a}\,(1+A) - \sqrt{1-2Gmr^2/c^2a^3}\right]^2 c^2dt^2 \\ &- \frac{dr^2}{1-r^2/P^2} - r^2d\Omega^2 \ , \end{aligned}\right\} \tag{5.39}$$

where A is to be determined from other physical considerations.

From Eqs.(5.37) and (5.39) we find that the pressure is given by

$$p(r) = \frac{\rho_0 c^2}{3}\left[\frac{A\sqrt{1-r^2/P^2}}{(1+A)\sqrt{1-a^2/P^2} - A\sqrt{1-r^2/P^2}} - 2\right] \ . \tag{5.40}$$

If we require that the pressure also be constant, we will have $A = 0$ and $p_0(r) = -2\rho_0 c^2/3$. On the other hand, if we require that $p(a) = 0$, so that the pressure is also continuous across the boundary, we will have $A = 2/\sqrt{1-a^2/P^2}$.

It is worth emphasising that we have assumed that the sphere is homogenous. In general there would be some equation of state, given by thermodynamic considerations, which would relate pressure and density In that case we would have to integrate Eq.(5.28) with ρ as a given function of p, to obtain $\nu(r)$. The interior of a star would be given by the resulting solution if we knew the equation of state well enough and could realistically regard it as spherically symmetric (and not an ellipsoid with only axial symmetry). Again, rotation would have to be taken to be negligible so that the assumption of spherical symmetry is valid.

5.3 The Reissner-Nordström Metric

We now come to gravitational and electromagnetic fields due to a point mass, m, with charge Q at rest at the origin. The metric is again spherically symmetric and static. The solution obtained is not only for the Einstein equations but the Maxwell equations as well. When we require a solution to two different sets of field equations, each of which has an effect on the other, we solve the *coupled* field equations. In this case we have the coupled Einstein-Maxwell field equations to be solved. The gravitational field enters into the Maxwell equations as we require the covariant divergence of the field tensor to be zero (or more generally the current density). In the reverse direction, the electromagnetic stress-energy tensor, given by Eq.(4.20), acts as a source term for the gravitational field. Physically, the energy distribution due to the electromagnetic field has an effective mass, which "causes" a gravitational field. We will not solve the Maxwell equations but will assume that the usual solution works and verify that our assumption is valid at the end.

We take the electromagnetic 4-vector potential to be

$$A_\mu = (Q/cr, \underline{0}) \ . \tag{5.41}$$

Then it is easy to see that

$$F_{\mu\nu} = -F_{\nu\mu} = 2\delta^0_{[\mu}\delta^1_{\nu]}Q/cr^2 \ . \tag{5.42}$$

Inserting this value into Eq.(4.20), writing the stress-energy tensor in mixed form, we get

$$\left.\begin{aligned} T^0_0 = T^1_1 = -T^2_2 = -T^3_3 &= Q^2e^{-(\upsilon+\lambda)}/8\pi c^2r^4 \ , \\ T^\mu_\nu = 0 \qquad &\text{for } \mu \neq \nu \quad . \end{aligned}\right\} \tag{5.43}$$

Since $T_0^0 = T_1^1$, Eqs.(4.67) and (4.68) can be used in the field equations for gravity, to give $\nu'(r)+\lambda'(r) = 0$. Hence, once again we can take $\nu(r)+\lambda(r) = 0$. Remembering that here $T = 0$, the Einstein equation for the "22" component is $R_2^2 = \kappa T_2^2$. Using Eq.(4.71) gives

$$-\frac{1}{r^2}\left[\left(-re^{-\lambda}\right)' + 1\right] = -\frac{8\pi G}{c^2}\cdot\frac{Q^2}{8\pi c^2 r^4}\ . \tag{5.44}$$

Hence, solving the above equation yields

$$e^{\nu(r)} = e^{-\lambda(r)} = 1 + \frac{\alpha}{r} + \frac{GQ^2}{c^4 r^2}\ . \tag{5.45}$$

Again, choosing zero charge should yield the Schwarzschild exterior solution. Hence $\alpha = -2Gm/c^2$. Thus we get

$$\left.\begin{aligned} ds^2 = &\left(1 - \frac{2Gm}{c^2 r} + \frac{GQ^2}{c^4 r^2}\right)c^2 dt^2 - \frac{dr^2}{1 - 2Gm/c^2 r + GQ^2/c^4 r^2} \\ &- r^2 d\Omega^2\ , \end{aligned}\right\} \tag{5.46}$$

which is the *Reissner-Nordström metric* derived and investigated by them in 1916 and 1918 [23, 24].

We must now verify that A_μ given by Eq.(5.41) is a solution of the source-free Maxwell equations when coupled with gravity. We have

$$F^{\mu\nu}_{\ ;\nu} = \frac{1}{\sqrt{|g|}}\left(\sqrt{|g|}F^{\mu\nu}\right)_{,\nu} = 0\ . \tag{5.47}$$

Since $v(r) + \lambda(r) = 0$, $\sqrt{|g|} = r^2\sin\theta$. Also, from Eq.(5.42)

$$F^{\mu\nu} = -2\delta_0^{[\mu}\delta_1^{\nu]}e^{-(v+\lambda)}Q/cr^2 = -2\delta_0^{[\mu}\delta_1^{\nu]}Q/cr^2\ . \tag{5.48}$$

Hence Eq.(5.47) is automatically satisfied. It is truly remarkable that the usual solution of Maxwell's equation holds for the coupled Einstein-Maxwell system.

It should be noted here that we have managed to incorporate the electromagnetic field into GR. However, this does not amount, in itself, to a *unification* of gravity with electromagnetism. The difference is apparent if we consider the equation of motion here

$$\ddot{x}^\mu + \{_\nu\ ^\mu\ _\rho\}\,\dot{x}^\nu\dot{x}^\rho = (q/\mu)\,F^\mu_\nu\dot{x}^\nu\ , \tag{5.49}$$

where q and μ are the charge and mass of a test particle in the field of the charged gravitating source giving rise to the Reissner-Nordström metric. Here the gravitational effect is incorporated into the geometry represented by the left side of the equation, while the electromagnetic effect is entirely contained on the right side. There is another aspect of the problem that will be discussed later.

5.4 The Kerr and Charged Kerr Metrics

There is another important exact solution of Einstein's vacuum field equation obtained by R.P. Kerr [25] using spinors and requiring that the solution be a

rational function. It turns out to be the most general vacuum solution of the field equations with axial symmetry and time independence. It is not static, as there is no global timelike Killing vector, but is called *stationary*. We will not go into the derivation here. Spinors will be touched upon later. For the present we will just write down the metric in Boyer-Lindquist coordinates, which correspond most closely to the usual spherical polar coordinates used for the Schwarzschild and Reissner-Nordström metrics:

$$\left.\begin{aligned} ds^2 = &\left(1-2Gmr/\rho^2c^2\right)c^2dt^2 - \left(\rho^2/\Delta\right)dr^2 - \rho^2 d\theta^2 \\ &- \left[\left(r^2+a^2/c^2\right)\sin^2\theta + 2Gmra^2\sin^4\theta/\rho^2c^2\right]d\varphi^2 \\ &+ \left(2Gmra\sin^2\theta/\rho^2c^2\right)dtd\varphi \ , \end{aligned}\right\} \tag{5.50}$$

where a is a new parameter introduced in this metric (whose physical significance will be discussed later) and

$$\rho^2 = r^2 + a^2\cos^2\theta/c^2 \ , \tag{5.51}$$

$$\Delta = r^2 - 2Gmr/c^2 + a^2/c^2 \ . \tag{5.52}$$

That this is a solution of the Einstein field equations may be verified by working out the Ricci tensor and checking that it is zero. It is not quite so easy to evaluate the Ricci tensor here because of the off-diagonal term in Eq.(5.50). Also, all coefficients are functions of r and θ. Further, the inverse metric tensor, $g^{\mu\nu}$, is not simply given by term by term inversion of the elements. Though g^{11} and g^{22} are simply $1/g^{11}$ and $1/g^{22}$ respectively, g^{00}, $g^{03} = g^{30}$ and g^{33} are obtained by inverting the metrix. The non-zero Christoffel symbols are (dropping the summation convention):

$$\left.\begin{aligned} &\left\{{0\ \ i}^{\ 0}\right\} = \tfrac{1}{2}\left(g^{00}g_{00,i} + g^{03}g_{03,i}\right), && \left\{{i\ \ 3}^{\ 0}\right\} = \tfrac{1}{2}\left(g^{00}g_{03,i} + g^{03}g_{33,i}\right), \\ &\left\{{0\ \ 0}^{\ i}\right\} = -\tfrac{1}{2}g^{ii}g_{00,i} \ , && \left\{{0\ \ 3}^{\ i}\right\} = -\tfrac{1}{2}g^{ii}g_{03,i} \ , \\ &\left\{{0\ \ i}^{\ 3}\right\} = \tfrac{1}{2}\left(g^{03}g_{00,i} + g^{33}g_{03,i}\right), && \left\{{3\ \ i}^{\ 3}\right\} = \tfrac{1}{2}\left(g^{03}g_{03,i} + g^{33}g_{33,i}\right), \\ &\left\{{3\ \ 3}^{\ i}\right\} = -\tfrac{1}{2}g^{ii}g_{33,i} \ , && \left\{{i\ \ i}^{\ i}\right\} = \tfrac{1}{2}g^{ii}g_{ii,i} \ , \\ &\left\{{i\ \ j}^{\ i}\right\} = \tfrac{1}{2}g^{ii}g_{jj,i} \ . \end{aligned}\right\} \tag{5.53}$$

The indices i and j are used one stands for different values of 1 and 2. Where there are two different lower indices the symmetry property can be used to evaluate the Christoffel symbol with the indices reversed. We find $R_{\mu\nu} = 0$.

This metric represents a point mass m with intrinsic angular momentum ma. Due to the "spin" of the gravitational source inertial frames are "dragged along" with the rotation, much as twirling a stick in thick treacle would drag the treacle along. In some ways the spacetime behaves like a sticky fluid being dragged by the gravitational source.

The Kerr metric is a solution of the vacuum equations. Kerr also found a similar solution of the Einstein-Maxwell equations corresponding to a spinning charged mass, the generalization of the Reissner-Nordström metric. This solution was also obtained by a neat trick of complexification by E.T. Newman [26]. It is, therefore, known as the *charged Kerr*, or the *Kerr-Newman*, metric. It differs from the Kerr metric by adding GQ^2/c^4 to Δ in Eq.(5.52) and subtracting it from $2mr/c^2$ in the zero-zero component of the metric tensor, so that

$$\left.\begin{aligned} ds^2 = &\left[1 - \left(2Gmr/c^2 - GQ^2/c^4\right)/\rho^2\right]c^2dt^2 - \left(\rho^2/\Delta\right)dr^2 - \rho^2 d\theta^2 \\ &- \left[\left(r^2+a^2/c^2\right)\sin^2\theta + 2Gmra^2\sin^2\theta/\rho^2c^4\right]d\varphi^2 \\ &+ \left(2Gmra\sin^2\theta/\rho c^2\right)dtd\varphi \ , \end{aligned}\right\} \tag{5.54}$$

where ρ^2 is given by Eq.(5.51) and

$$\Delta = r^2 - 2Gmr/c^2 + GQ^2/c^4 + a^2/c^2 \ , \tag{5.55}$$

Notice that in the limit $a \to 0$ this metric reduces to the Reissner-Nordström metric and in the limit $Q \to 0$ to the Kerr metric. Further, the Reissner-Nordström metric reduces to the usual Schwarzchild metric as $Q \to 0$ and so does the Kerr metric in the limit $a \to 0$. The Kerr and charged Kerr metrics are axially symmetric and stationary, the terms involving 'a' destroying spherical symmetry. When $a \to 0$ spherical symmetry is restored. (A metric is said to be "stationary" if it is time independent i.e. has a time-like Killing vector but there is no space-like hypersurface globally orthogonal to it. If there is such a hypersurface then the metric is said to be "static".)

5.5 Gravitational Waves and Linearised Gravity

One of the most important differences between GR (as a field theory of gravity) and Newtonian Gravity is that GR is non-linear. Newtonian gravity is given by a scalar potential leading to a linear differential equation, Eq.(4.41). Though many of the interesting consequences of GR come from its nonlinearity it is worth while to study its linear approximation. This is useful to be able to compare the results of both theories so as to look at GR as a "correction" of Newtonian theory. It is also useful to be able to see clearly, the consequence of having a *tensor* theory rather than a scalar theory, without getting confused and side-tracked by nonlinearity. Linearization naturally leads one to the prediction of gravitational waves.

Since the field in GR is the metric tensor, its appearance in the field equations nonlinearly gives rise to the nonlinearity of GR. We can not change the way the metric tensor enters into the curvature but we can write the curved spacetime metric tensor as the flat (Minkowski) spacetime metric tensor, which we shall write as $\eta_{\mu\nu}$ and an additional term, $h_{\mu\nu}$. We can then require that $h_{\mu\nu}$ and its derivatives occur only once in the field equations and higher powers be neglected. Thus

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu} \ . \tag{5.56}$$

Let the inverse metric tensor be given as

$$g^{\mu\nu} = \eta^{\mu\nu} + f^{\mu\nu} \ , \tag{5.57}$$

where $\eta^{\mu\nu}$ is the inverse of $\eta_{\mu\nu}$. Then, by definition

$$\begin{aligned} g^{\mu\rho}g_{\rho\pi} &= \delta^\mu_\pi = (\eta^{\mu\rho} + f^{\mu\rho})(\eta_{\rho\pi} + h_{\rho\pi}) \\ &= \delta^\mu_\pi + f^{\mu\rho}\eta_{\rho\pi} + \eta^{\mu\rho}h_{\rho\pi} + f^{\mu\rho}h_{\rho\pi} \ . \end{aligned} \tag{5.58}$$

Cancelling the δ^μ_π on both sides and multiplying through by $\eta^{\nu\pi}$,

$$f^{\mu\nu} + \eta^{\mu\rho}\eta^{\rho\pi}h_{\rho\pi} + \eta^{\nu\pi}f^{\mu\rho}h_{\rho\pi} = 0 \ . \tag{5.59}$$

The last term is clearly quadratic in the difference between the curved and flat spacetime metric tensors. Thus, to the first order

$$f^{\mu\nu} = -\eta^{\mu\rho}\eta^{\nu\pi}h_{\rho\pi} + \mathcal{O}(h^2) \ . \tag{5.60}$$

Using the flat spacetime metric tensor to raise and lower indices we can write $f^{\mu\nu} = -h^{\mu\nu} + \mathcal{O}(h^2)$.Thus Eq.(5.57) becomes

$$g^{\mu\nu} = \eta^{\mu\nu} - h^{\mu\nu} + \mathcal{O}(h^2) \approx \eta^{\mu\nu} - h^{\mu\nu} \ . \tag{5.61}$$

Using this linearization and for the moment taking Cartesian coordinates, so that there are no derivatives of $\eta_{\mu\nu}$, the Christoffel symbols linearize to

$$\{_{\mu\ \nu}^{\ \rho}\} \approx \frac{1}{2}\eta^{\rho\pi}\left(h_{\mu\pi,\nu} + h_{\nu\pi,\mu} - h_{\mu\nu,\pi}\right) \ . \tag{5.62}$$

Clearly, terms quadratic in Christoffel symbols become quadratic in h and can be neglected compared with the linear terms. Thus the linearized Ricci tensor is

$$\begin{aligned} R_{\mu\nu} &\approx \{_{\mu\ \nu}^{\ \rho}\}_{,\rho} - \{_{\mu\ \rho}^{\ \rho}\}_{,\nu} \\ &\approx \frac{1}{2}\left[\eta^{\rho\pi}\left(h_{\mu\pi,\nu} + h_{\nu\pi,\mu} - h_{\mu\nu,\pi}\right)_{,\rho}\right] \\ &\quad -\frac{1}{2}\left[\eta^{\rho\pi}\left(h_{\mu\pi,\rho} + h_{\rho\pi,\mu} - h_{\mu\rho,\pi}\right)_{,\nu}\right] \\ &= \frac{1}{2}\eta^{\rho\pi}\left(h_{\mu\rho,\nu\pi} + h_{\nu\pi,\mu\rho} - h_{\rho\pi,\mu\nu} - h_{\mu\nu,\rho\pi}\right) \ . \end{aligned} \tag{5.63}$$

Notice that putting $h = h^{\mu}_{\mu}$ and $\eta^{\rho\pi}f_{,\rho\pi} = \Box f$, we can rewrite this equation as

$$R_{\mu\nu} \approx \frac{1}{2}\left(h^{\rho}_{\mu,\nu\rho} + h^{\rho}_{\nu,\mu\rho} - h_{,\mu\nu}\right) - \frac{1}{2}\Box h_{\mu\nu} \ , \tag{5.64}$$

If we could get rid of the terms in the bracket, the vacuum Einstein equations would reduce to the wave equation. To do so we break it into two terms, so that

$$R_{\mu\nu} \approx \frac{1}{2}\left(h^{\rho}_{\mu} - \frac{1}{2}h\delta^{\rho}_{\mu}\right)_{,\rho\nu} + \frac{1}{2}\left(h^{\rho}_{\nu} - \frac{1}{2}h\delta^{\rho}_{\nu}\right)_{,\mu\rho} - \frac{1}{2}\Box h_{\mu\nu} \ . \tag{5.65}$$

Now define the infinitesimal transformation

$$x^{\mu} \to x'^{\mu} = x^{\mu} + \xi^{\mu}(x^{\rho}) \ , \tag{5.66}$$

so that terms quadratic in ξ and/or h can be neglected. Since this is only a coordinate transformation it leaves the metric invariant and hence

$$\begin{aligned} ds^2 &= g_{\mu\nu}(x^{\rho})dx^{\mu}dx^{\nu} = g'_{\mu\nu}(x'^{\rho})dx'^{\mu}dx'^{\nu} \\ &= g'_{\mu\nu}(x'^{\rho})\left(dx^{\mu} + \xi^{\mu}_{,\alpha}dx^{\alpha}\right)\left(dx^{\nu} + \xi^{\nu}_{,\beta}dx^{\beta}\right) \ . \end{aligned} \tag{5.67}$$

Using the linearization procedure it is easy to see that

$$h'_{\mu\nu} \approx h_{\mu\nu} - \xi_{\mu,\nu} - \xi_{\nu,\mu} \ . \tag{5.68}$$

Thus we have

$$h'^{\rho}_{\mu} - \frac{1}{2}h'\delta^{\rho}_{\mu} = h^{\rho}_{\mu} - \frac{1}{2}h\delta^{\rho}_{\mu} - \eta^{\nu\rho}\xi_{\mu,\nu} - \xi^{\rho}_{,\mu} + \xi^{\nu}_{,\nu}\delta^{\rho}_{\mu} \ . \tag{5.69}$$

Differentiating Eq.(5.69) relative to x^{ρ} it is easy to see that the last two terms simply cancel and we have

$$\left(h'^{\rho}_{\mu} - \frac{1}{2}h'\delta^{\rho}_{\mu}\right)_{,\rho} = \left(h^{\rho}_{\mu} - \frac{1}{2}h\delta^{\rho}_{\mu}\right)_{,\rho} - \Box\xi_{\mu} \ . \tag{5.70}$$

Defining the term in the second bracket to be φ^{ρ}_{μ} and choosing ξ_μ to satisfy

$$\Box\xi_\mu = -\varphi'^{\rho}_{\mu,\rho} \ , \tag{5.71}$$

we finally obtain

$$\varphi^{\rho}_{\mu,\rho} = 0 \ . \tag{5.72}$$

The condition (5.71) is reminiscent of the choice of the *Lorentz gauge* for the electromagnetic field, in which the electromagnetic 4-vector potential satisfies the wave equation. Here with the Lorentz gauge condition the gravitational 4-*tensor potential*, $(\varphi_{\mu\nu})$, satisfies the wave equation. It is again a choice of *gauge* that we have made and Eq.(5.66) represents the gauge transformation for gravity according to GR. This is an arbitrary infinitesimal coordinate transformation in 4-d. Thus the gauge group for GR is $GL(4,\mathbb{R})$.

It should be pointed out here that the use of Cartesian coordinates was not crucial, but merely for convenience. With any other coordinates we would have to introduce the corresponding flat spacetime Christoffel symbols. This has been avoided so as not to cause confusion of notation.

We now drop the primes and so obtain

$$R_{\mu\nu} \approx -\frac{1}{2}\eta^{\rho\pi}h_{\mu\nu,\rho\pi} \equiv -\frac{1}{2}\Box h_{\mu\nu} \ . \tag{5.73}$$

We can contract Eq.(5.73) to obtain the Ricci scalar and hence combine them to obtain the Einstein tensor. Thus the general Einstein field equations given by Eqs.(5.24), but without the cosmological constant (i.e. $\Lambda = 0$), become

$$\Box\varphi_{\mu\nu} = -2\kappa T_{\mu\nu} \ . \tag{5.74}$$

In regions where $T_{\mu\nu} = 0$, $\varphi_{\mu\nu}$ satisfies the wave equation. These are gravitational waves. If $T_{\mu\nu} \neq 0$, we have gravitational waves with source.

Linearization has been used in attempts to quantize gravity and unify it with the other forces of nature, by treating $\eta_{\mu\nu}$ as classical and quantizing $h_{\mu\nu}$. Usual quantization in this approach leads to infinities that are non-renormalizable. However, the use of supersymmetry (a postulated symmetry between bosons and fermions, i.e. integral and half-integral spin objects) and replacement of "particles" by "strings" *may* lead to a finite theory. The gauge transformations of GR play a crucial role there. Many relativists believe that this approach will not be successful, as it is non-geometrical, regarding gravity as a spin 2 field (as the potential has two tensor indices) in a Minkowski spacetime. Clearly there remains no justification for the Einstein-Hilbert-Lagrangian, which makes sense only in its geometrical context. Indeed, Superstring theory (the candidate for quantum gravity and unification just mentioned) does require that there be more terms in the Lagrangian, quadratic, cubic, etc. in the curvature tensor. If this is true it is surprising that usual GR works so well. Why should the geometrical Lagrangian bear any resemblance to a physical Lagrangian, let alone give the classical (i.e. non-quantum) limit exactly? If Superstring theory works this question will have to be addressed.

Linearised theory is also extremely useful for making comparisons of GR with other theories — particularly with Newtonian gravity — so as to test between them experimentally and observationally. In fact we need not limit ourselves to a linear theory but extend Eq.(5.61) to the quadratic term. Then all subsequent equations would be correspondingly altered. In particular, we could

use Eq.(5.63) to obtain the motion of particles and compare it with other theories, order by order. This is done in the PPN (parameterised post-Newtonian) formalism. It would not be out of place to mention that all observations and experiments agree with GR (as of even date) and other theories have been excluded unless they be so adjusted as to give results consistent with GR, within experimental (or observational) error. We will discuss gravitational waves further.

5.6 Exact Gravitational Wave Solutions

So far we have obtained the wave equation for gravity by linearizing the Einstein field equations. In principle the solutions so obtained could be exact solutions of the vacuum Einstein field equations. Of course, there could be trivial, static, solutions which effectively satisfy the Laplace equation. Though technically solutions of the wave equation, they do not represent "moving waves". We are, therefore, interested in time dependent exact solutions of Einstein's vacuum equations. One might expect that plane or spherical gravitational wave solutions would have been obtained first. It turns out that spherical gravitational waves *cannot* exist. The reason is that in a vacuum, the general spherically symmetric solution is the Schwarzschild exterior solution. Thus, whatever fluctuations might arise in the region where matter is present, so long as they are spherically symmetric they will not be able to influence the region outside the matter. (However, there *is* an exact solution obtained for a spherical gravitational wave with a cone excised [27].

Einstein and Rosen [28] discovered the first exact gravitational wave solution in 1937, which was for cylindrical waves. The general cylindrically symmetric metric depends on two arbitrary functions, γ and ψ, of the time, t, and the cylindrical radial coordinate, ρ,

$$ds^2 = e^{2(\gamma-\psi)}(c^2dt^2 - d\rho^2) - e^{-2\psi}\rho^2 d\varphi^2 - e^{2\psi}dz^2 \ . \tag{5.75}$$

It may be easily verified that the non-zero Christoffel symbols for this metric are:

$$\left.\begin{array}{l}
\{_0\ ^0\ _0\} = \{_1\ ^0\ _1\} = \{_1\ ^1\ _0\} = \{_0\ ^1\ _1\} = \dot\gamma - \dot\psi \ , \\
\{_0\ ^0\ _1\} = \{_1\ ^0\ _0\} = \{_0\ ^1\ _0\} = \{_1\ ^1\ _1\} = \gamma' - \psi' \ , \\
\{_2\ ^0\ _2\} = -\rho^2\dot\psi e^{-2\gamma} \ , \ \{_3\ ^0\ _3\} = \dot\psi e^{2(2\psi-\gamma)} \ , \\
\{_2\ ^1\ _2\} = \rho\left(\rho\psi' - 1\right)e^{-2\gamma} \ , \ \{_3\ ^1\ _3\} = -\psi' e^{2(2\psi-\gamma)} \ , \\
\{_0\ ^2\ _2\} = \{_2\ ^2\ _0\} = -\dot\psi \ , \ \{_1\ ^2\ _2\} = \{_2\ ^2\ _1\} = -\left(\psi' - 1/\rho\right) \ , \\
\{_0\ ^3\ _3\} = \{_3\ ^3\ _0\} = \dot\psi \ , \ \{_1\ ^3\ _3\} = \{_3\ ^3\ _1\} = \psi' \ , \\
\{_\mu\ ^\mu\ _0\} = \left(\ln\sqrt{|g|}\right)_{,0} = 2\left(\dot\gamma - \dot\psi\right) \ , \\
\{_\mu\ ^\mu\ _1\} = \left(\ln\sqrt{|g|}\right)_{,1} = 2\left(\gamma' - \psi'\right) + 1/\rho \ ,
\end{array}\right\} \tag{5.76}$$

where the prime refers to differentiation with respect to ρ and the dot to differentiation with respect to ct. (To convert it to differentiation with respect to t the term would simply have to be divided by c.) The non-vanishing components of the Ricci tensor are R_{00}, R_{01}, R_{11}, R_{22} and R_{33}. We only need the first four components for our purposes. Using the Christoffel symbols of Eq.(5.76) in Eq.(3.106), to obtain these components of the Ricci tensor and simplifying,

the relevant Einstein equations become

$$R_{00} = -\left(\ddot{\gamma} - \ddot{\psi}\right) + (\gamma'' - \psi'') + \frac{1}{\rho}(\gamma' - \psi') - 2\dot{\psi}^2 = 0 \ , \tag{5.77}$$

$$R_{11} = \left(\ddot{\gamma} - \ddot{\psi}\right) - (\gamma'' - \psi'') + \frac{1}{\rho}(\gamma' + \psi') - 2\psi'^2 = 0 \ , \tag{5.78}$$

$$R_{22} = \rho^2 e^{-2\gamma}\left(-\ddot{\psi} + \psi'' + \frac{1}{\rho}\psi'\right) = 0 \ , \tag{5.79}$$

$$R_{01} = \frac{1}{\rho}\dot{\gamma} - 2\dot{\psi}\psi' = 0 \ . \tag{5.80}$$

Eq.(5.79) is the usual cylindrical form of the wave equation. Being a second order linear differential equation the solution has two arbitrary constants, one corresponding to ingoing cylindrical waves and the other to outgoing. Retaining only the outgoing waves with amplitude A and frequency ω

$$\psi(t,\rho) = A\left[J_0(x)\cos\omega t + N_0(x)\sin\omega t\right] \ , \tag{5.81}$$

where $x = \omega\rho/c$ and J_0 and N_0 are the zero order Bessel and Neumann functions respectively.

To solve the other equations add Eqs.(5.77) and (5.78) to obtain

$$\gamma' = \rho\left(\dot{\psi}^2 + \psi'^2\right) \ . \tag{5.82}$$

Now Eqs.(5.80) and (5.82) give the space and time derivatives of $\gamma(t,\rho)$ in terms of functions that are now known, through Eq.(5.81). All that is required is an integration with respect to each variable. The time integration of Eqs.(5.80) is relatively easy and the space integration, though tedious, is in principle easy (using standard formulae for the integrals of Bessel and Neumann functions). The resulting solution for γ is, then,

$$\left.\begin{aligned}\gamma(t,\rho) =& \frac{1}{2}A^2 x\{J_0(x)J_0'(x) + N_0(x)N_0'(x) \\ &+ x[J_0(x)^2 + N_0(x)^2 + J_0'(x)^2 + N_0'(x)^2] \\ &+ [J_0(x)J_0'(x) - N_0(x)N_0'(x)]\cos 2\omega t \\ &+ [J_0(x)N_0'(x) + J_0'(x)N_0(x)]\sin 2\omega t\} \\ &- \frac{2}{\pi}A^2\omega t \ ,\end{aligned}\right\} \tag{5.83}$$

where prime now refers to differentiation with respect to x, not ρ. It is easy to verify, by subtracting Eq.(5.77) from (5.78), that all four equations are satisfied. Notice that the Bessel functions arise because of cylindrical symmetry. Had there been spherical waves, we would get spherical Bessel functions. Since there are no infinite cylindrical sources, actual sources should lie somewhere between the two. The reader is encouraged to check that $R_{33} = 0$.

We now come to the plane wave solution of Bondi and Robinson [29], discovered in 1957. Here we take a line element which incorporates the symmetries of a plane and represents a wave going in the x-direction, in that all coefficients in the metric are functions of $(ct - x)$, which will be represented by u,

$$ds^2 = e^{2\Omega(u)}\left(c^2dt^2 - dx^2\right) - u^2\left(e^{2\beta(u)}dy^2 + e^{-2\beta(u)}dz^2\right) \ . \tag{5.84}$$

It should be borne in mind that the (x, y, z) are not usual Cartesian coordinates but rather rectangular coordinates in a curved spacetime. (Cartesian coordinates would, of course, be in a flat spacetime only.)

Writing the derivative of Ω (or β) with respect to u as Ω' (or β'), it is clear that $\Omega_{,0} = \Omega' = -\Omega_{,1}$ (and $\beta_{,0} = \beta' = -\beta_{,1}$). Similarly $u_{,0} = 1 = -u_{,1}$. The non-zero Christoffel symbols are:

$$\left.\begin{array}{l} \{_0\ ^0\ _0\} = \{_1\ ^0\ _1\} = \{_0\ ^1\ _1\} = \{_1\ ^1\ _0\} = \Omega' \ , \\ \{_0\ ^0\ _1\} = \{_1\ ^0\ _0\} = \{_0\ ^1\ _0\} = \{_1\ ^1\ _1\} = -\Omega' \ , \\ \{_2\ ^0\ _2\} = \{_2\ ^1\ _2\} = u\,(u\beta' + 1)\, e^{2(\beta-\Omega)} \ , \\ \{_3\ ^0\ _3\} = \{_3\ ^1\ _3\} = u\,(-u\beta' + 1)\, e^{-2(\beta+\Omega)} \ , \\ \{_0\ ^2\ _2\} = \{_2\ ^2\ _0\} = -\{_1\ ^2\ _2\} = \{_2\ ^2\ _1\} = (\beta' - 1/u) \ , \\ \{_0\ ^3\ _3\} = \{_3\ ^3\ _0\} = \{_1\ ^3\ _3\} = \{_3\ ^3\ _1\} = (-\beta' + 1/u) \ . \end{array}\right\} \tag{5.85}$$

The non-trivially zero components of the Ricci tensor are as before for the cylindrical gravitational waves. The reader should verify that, in this case, R_{22} and R_{33} also come out to be zero. Also, that the remaining three Einstein equations just reduce to the same. As such we have only one Einstein equation

$$R_{00} = \{_0\ ^\mu\ _0\}_{,\mu} - \left(\ln\sqrt{|g|}\right)_{,00} + \left(\ln\sqrt{|g|}\right)_{,\mu} \{_0\ ^\mu\ _0\} - \{_\nu\ ^\mu\ _0\}\{_\mu\ ^\nu\ _0\} \ . \tag{5.86}$$

For the above metric we have

$$\begin{aligned} \ln\sqrt{|g|} &= 2\ln u + 2\Omega \ , \\ \left(\ln\sqrt{|g|}\right)' &= 2\,(\Omega' + 1/u) \ , \\ \left(\ln\sqrt{|g|}\right)'' &= 2\left(\Omega'' - 1/u^2\right) \ . \end{aligned} \tag{5.87}$$

Thus we have

$$\begin{aligned} R_{00} &= \{_0\ ^0\ _0\}_{,0} + \{_0\ ^1\ _0\}_{,1} - \left(\ln\sqrt{|g|}\right)_{,00} + \left(\ln\sqrt{|g|}\right)_{,0} \{_0\ ^0\ _0\} \\ &\quad + \left(\ln\sqrt{|g|}\right)_{,1} \{_0\ ^1\ _0\} - \{_0\ ^0\ _0\}^2 - 2\{_1\ ^0\ _0\}\{_0\ ^1\ _0\} - \{_1\ ^1\ _0\}^2 \\ &\quad - \{_2\ ^2\ _0\}^2 - \{_3\ ^3\ _0\}^2 \\ &= 2\Omega'' - 2\left(\Omega'' - 1/u^2\right) + 4\Omega'\,(\Omega' + 1/u) - 4\Omega'^2 \\ &\quad - \left(\beta'^2 + 2\beta/u + 1/u^2\right) - \left(\beta'^2 - 2\beta/u + 1/u^2\right) \\ &= 4\Omega'/u - 2\beta'^2 = 0 \ . \end{aligned} \tag{5.88}$$

Hence any $\beta(u)$ and $\Omega(u)$ which satisfy the relation

$$\Omega'(u) = \frac{1}{2} u\beta'(u)^2 \ , \tag{5.89}$$

is an exact plane gravitational wave solution. This solution will become trivial, in that the spacetime will be Minkowski, if

$$u\beta''(u) + 2\beta'(u) = u^2\beta'(u)^3 \ . \tag{5.90}$$

This equation can be integrated by putting $\beta'(u) = \gamma(u)/u$ to give

$$\beta(u) = \pm \ln \left| \frac{1 + \sqrt{1 - Au^2}}{Bu} \right| , \tag{5.91}$$

where A, B are the integration constants. These values of $\beta(u)$ must be excluded for the gravitational wave solutions.

Eq.(5.84) represents a linearly polarized gravitational wave. The more general, circularly polarized plane-fronted gravitational wave is given by the line element

$$\left.\begin{aligned} ds^2 = \; & e^{2\Omega}\left(c^2dt^2 - dx^2\right) - (u/U)^2[\left(dy^2 + dz^2\right)\cosh 2\beta \\ & + \left(dy^2 - dz^2\right)\sinh 2\beta \cos 2\theta - 2\sinh 2\beta \sin 2\theta dydz] \, , \end{aligned}\right\} \tag{5.92}$$

where U is a constant and Ω, β and θ are arbitrary functions of u related by

$$2\Omega'(u) = u\left[\beta'(u)^2 + \theta'(u)^2 \sinh^2 2\beta(u)\right] \; . \tag{5.93}$$

Clearly, if $\theta = 0$ the above equations reduce to Eqs.(5.84) and (5.89).

5.7 The Interpretation of Gravitational Wave Solutions

The stress-energy tensor is, by definition, zero for the exact gravitational waves solutions of the vacuum field equations. This creates a conceptual problem. How can these solutions of Einstein's field equations represent waves if they carry no energy? Essentially, this problem arises because energy is not generally a well-defined concept in GR. In the context of Classical Mechanics the Hamiltonian in the Poisson bracket acts as a time derivative for a conservative system. Then energy is a conserved, and hence well-defined, quantity. Thus it is clear that for energy to be well-defined in GR, the metric tensor must have a time-like isometry (Killing vector), so as to allow time-translational invariance. This will not generally be true. In fact spacetimes for which it is true are static whereas genuine (non-trivial) gravitational wave solutions *must* be non-static. Thus energy *is not* well-defined for spacetimes containing gravitational waves. The question then is, how can we tell that these solutions really do behave as we would expect of *waves*? The answer can be seen in various ways which we shall now discuss.

One point to note is that the context in which we want the energy to be defined is as "the work the gravitational wave does" in some sense. Thus, at the back of the mind is the idea of a test particle on which this gravitational wave acts and causes a change of its motion. This concept of "causing a change of motion" has implicit in it the idea of a force. In other words, we are thinking of the work done on a test particle by the force in the gravitational wave. We could then look for the wave behaviour of the work done. Now in GR we have eliminated forces and replaced them by geometry. All that is relevant is the equation of the motion of the particle, given by the geodesic equation. There the force term (as for example in the case of the Lorentz force given in Eq.(5.49)) appears on the right hand side of the equation, in place of zero. However, gravity appears on the left side in the Christoffel symbol. One could, then, hope to see the wave-like behaviour in the Christoffel symbol. Since the Christoffel symbol

is observer dependent this way of seeing the effect is not trivial. We will discuss it later in this chapter, very briefly.

A way to visualise gravitational waves is by analogy with electromagnetic waves. Consider a point charge oscillated back and forth with a regular frequency about some mean position. At any observation point the electric field will change cyclically. The change of the position of the source can not have an instantaneous effect at the observation point. Rather, the disturbance will travel at the speed of light, the fastest available speed. This disturbance is an electromagnetic wave (in fact light if the frequency is in the correct range). More generally, in a field theory of electromagnetism accelerating an electric charge gives electromagnetic radiation. Similarly, in a field theory of gravity oscillating (or more generally accelerating) a "gravitational charge", i.e a mass, gives gravitational waves (or gravitational radiation). This argument by analogy does not address the problem of the reality of exact gravitational wave solutions, nor does it distinguish between scalar, vector and tensor waves. The problem of the reality of gravitational waves stems from the 4-d way of thinking of the whole history "frozen in", instead of *happening*. In this way of thinking the gravitational wave will be frozen in and test particles lying on it will merely *appear* to oscillate. So the waves are as "real" as anything else.

A more precise way to deal with the problem is to consider the effect of the gravitational wave pulse (or impulse) on a test particle, initially at rest in some frame, and ask what its momentum will be in that frame, after the pulse (or impulse) has passed. It was found that the plane (impulse) wave imparts a constant momentum to the particle, while the cylindrical pulse gives the expected wave behaviour. Thus the test particle gets energy from the gravitational wave. This is the counterpart of the accelerating particle radiating. Here the radiation accelerates the particle. To this extent the reality of the gravitational wave has been demonstrated. Replacing the single test particle by a sphere of test particles we can calculate its deformation due to the passage of the wave. This displays the *tensor* nature of the gravitational wave. We will not go into the details here.

Where does the energy extracted from the gravitational wave by the test particle come from? The energy given by the force defined by the geodesic equation (mentioned earliar) depends on the choice of the reference frame. The reference frame required for this purpose may be incompatible with the choice of guage in which the Einstein equations reduce to a wave equation. In the *relevant* frame, where does the energy come from? It can be (and had been) argued that the very process of the linearization provides the energy. The point is that if the $\Box\varphi_{\mu\nu} = 0$ exactly, then $R_{\mu\nu} \approx 0$ to order h but $R_{\mu\nu} \neq 0$ to higher orders in h. Thus $T_{\mu\nu} \approx 0$ only to order h, but $T_{\mu\nu} \neq 0$ generally. Conversely, if $T_{\mu\nu} = 0$, and hence $R_{\mu\nu} = 0$ exactly then $\Box\varphi_{\mu\nu} \approx 0$ only to order h. Thus we can expand $R_{\mu\nu}$ in powers of h, retaining the linear terms on the left side of the equation and transposing all higher powers to the right side. These higher order terms become an effective stress-energy tensor, called the stress-energy *pseudo-tensor* and the linearized Einstein equations give the gravitational wave. However this does not correspond to the momentum (energy) imparted to test particles.

5.8 The $(3+1)$ Split of Spacetime

A fundamental difference between Newtonian Mechanics and Maxwell's Electromagnetism is that the latter is a field theory while the former is not. Thus, in Newtonian mechanics one has a clear concept of a "cause" and its "effects". In the field theory the entire past is a "cause" for the future. One can not isolate separate "causes". Special Relativity (which is implicit in the field theory of Electromagnetism) limits the "cause" for any event to the past light cone of the event, rather than the entire past. While linearization provided a "Newtonian understanding" of GR in the sense that non-linear effects were regarded as "corrections" to a linear theory, it did not provide a "cause and effect" Newtonian understanding. Also, it could not deal with strong gravitational fields.

General Relativity is very much a field theory. Its very basis is the unification of the 3-d Newtonian *space* with the 1-d Newtonian *time* to form a single 4-d *spacetime* manifold. As just mentioned, in this 4-d arena things do not *happen*, they just *are*. To obtain the "cause and effect" Newtonian understanding the unification would have to be reversed and spacetime split back into space and time. One may feel that "that which Einstein hath joined together in Holy Unification let no mere mortal rend asunder". However, there is no reason to close one's eyes to the insights that can be obtained by this split. The point of the split is *not* to claim, in any way, that spacetime is unnecessary and only space and time are relevant, but to obtain a more complete comprehension of the whole by seeing its parts. Seeing the "heads" and "tails" faces of a coin helps comprehend the coin as a whole without casting doubts on its existence.

There is no unique $(3+1)$-split of spacetime. As an analogy, there are many ways to slice an apple (the traditional symbol for gravity since the time of the Newton and of Relativists since, at least, the publication of MTW [15]). For a splitting procedure to be valid one must ensure that *every* spacetime event lie on a *unique* spacelike hypersurface at a *unique* time. In terms of our analogy, the *whole* apple must be sliced and no part of it can lie on two separate slices. For our purposes another analogy is more useful. Imagine a big, thick, heavy book (MTW is an excellent example) lying on a table with its cover with the title, face down. The leaves of the book are so numerous and thin that they may be regarded as infinitely many. Clearly, no two leaves of the book intersect and, due to its weight, there are no gaps between them. Any point in the region occupied by the book can be specified by its page number and location on the page. This is a $(2+1)$-split of the 3-dimensional space occupied by the book, with the page numbers playing the role of a discrete time. One can now consider the continuum limit of this procedure applied to breaking spacetime into space and time. (To some people it feels that the book goes on and on continuously.) This procedure is called a *foliation* (from the Latin word for leaves, *folia*, which gives the common word *foliage*). Since it specifies space-like hypersurfaces at every instant of time, it provides a *frame of reference.* (The term is also applied for other splits of spaces or spacetimes, but such splits would *not* correspond to frames of reference.)

The simplest spacetime to discuss is, of course, Minkowski. The simplest foliation of this spacetime, according to some inertial observer, O, is by $t = const.$ (say t_0) hypersurfaces, as shown in Fig. 5.1. For convenience we will consider positions on the hypersurfaces given by Cartesian coordinates and suppress two directions and use r. Now consider an observer, O', moving with a constant

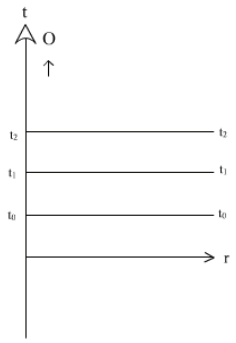


Figure 5.1: Time slicing the Minkowski spacetime so that at each value of t we get the complete hypersurface as seen in our arbitrarily chosen rest-frame. Visualized as a book, each slice is a page of the book. Pages are called "folia", by analogy with leaves. Consequently, this type of slicing got called "foliation".

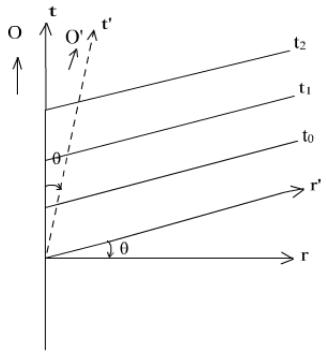


Figure 5.2: Time slicing Minkowski spacetime from the point of view of an observer moving at constant speed relative to the previous observer.

speed v in the r-direction relative to O. This observer would foliate the spacetime by $t' = t_0$ hypersurfaces, which would look like Fig. 5.1 in his frame. In the frame of O these hypersurfaces can be obtained by using the usual Lorentz transformations, to be

$$t = t_0\sqrt{1 - v^2/c^2} + rv/c^2 \ . \tag{5.94}$$

These hypersurfaces are shown, in Fig. 5.2 from the point of view of O and the hypersurfaces of O from the point of view of O' are shown in Fig. 5.3. Clearly the foliation procedure is non-unique.

Now consider a uniformly linearly accelerated observer, $\overline{O}$, with acceleration a relative to the inertial observer O and O'. The world line of $\overline{O}$ is shown in Fig. 5.4. Clearly this observer can not see the whole Minkowski space but only a quarter of it (called the Rindler wedge) whose boundary is asymptotic to his world line. The hypersurfaces $\bar{t} = t_0$ will *not* foliate the spacetime as the region outside the Rindler wedge is not accessible to $\overline{O}$. This is not to say that foliations of Minkowski spacetime can only correspond to inertial frames as the following example shows.

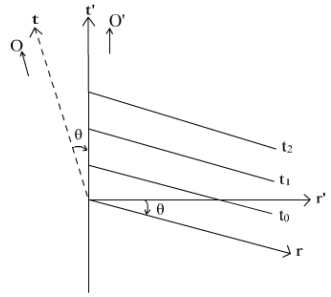


Figure 5.3: As the choice of frame is arbitrary, we could make a new diagram with (t', r') coordinates drawn in place of (t, r). Now the new (t, r) coordinates would be an oblique frame to either side of the (t', r') coordinates.

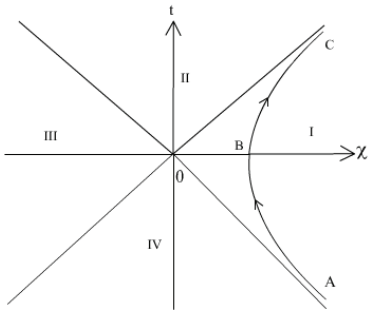


Figure 5.4: The Minkowski spacetime for an accelerated observer. This observer can only see the right wedge and not the rest of the spacetime. The part that is visible is called the *Rindler wedge*.

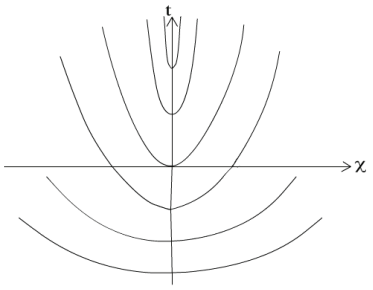


Figure 5.5: Time slicing by a sequence of parabolic hyper-cylinders $t = t_0 + (\tanh(t_0/T))\,(x/X)^2$, where T, X are arbitrary constants and increasing values of t_0 move up in the foliation.

Consider the family of hypersurfaces given by the parabolic hyper-cylinders in 4-d

$$t = t_0 + T \tanh(t_0/T)(x/X)^2 \ , \tag{5.95}$$

where T, X are constants with units of time and distance, respectively. For any $t_0 \in (-\infty, +\infty)$ there is a unique t for every x. Conversely, given any t and x,

it is easily verified that there is a unique t_0. Hence every event lies on some hypersurface and none lies on two. Thus we have a foliation (shown in Fig. 5.5).

To work out the frame corresponding to this foliation we first construct the unit tangent to this spacelike hypersurface, T^μ. Now identify T^0 with $c\delta t/\delta s$ and T^1 with $\delta x/\delta s$, taking the vector to lie in the t-x plane for convenience, the requirement that it is a unit space-like vector gives

$$(\delta x/\delta s)^2 - c^2\,(\delta t/\delta s)^2 = 1\ . \tag{5.96}$$

Differentiating Eq.(5.95) relative to x we get

$$c\delta t/\delta x = 2(x/X)\tanh(t_0/T)\ . \tag{5.97}$$

Solving Eqs.(5.96) and (5.97) simultaneously, using Eq.(5.95) to eliminate $\tanh t_0$ and inserting the values of T^0 and T^1, we have

$$T^\mu = \frac{\pm 1}{\sqrt{1-\left(2c\frac{t-t_0}{x}\right)^2}}\left(2c\frac{t-t_0}{x},1,0,0\right)\ . \tag{5.98}$$

The unit normal, N^μ, is then given by the requirements

$$N^\mu T^\nu g_{\mu\nu} = 0\ ,\quad N^\mu N^\nu g_{\mu\nu} = 1\ . \tag{5.99}$$

The above equations are easily solved to yield

$$N^\mu = \frac{\pm 1}{\sqrt{1-\left(2c\frac{t-t_0}{x}\right)^2}}\left(1,2c\frac{t-t_0}{x},0,0\right)\ . \tag{5.100}$$

Writing N^μ as the derivatives of the world line coordinates along the curve, we see that

$$dx/dt = 2c^2\,(t-t_0)\,/x\ . \tag{5.101}$$

Integrating Eq.(5.101) we obtain the equation for the world line orthogonal to the hypersurface, at some t and x, to be

$$x^2 = c^2\left(t^2 - 2tt_0 + \alpha\right)\ , \tag{5.102}$$

where α is a constant with units of time squared to be determined by the requirement that some t and x satisfy Eqs.(5.95) and (5.102). Taking the square root of Eq.(5.102) and differentiating twice relative to t gives the acceleration

$$a(t) = \frac{d^2x}{dt^2} = \pm\frac{c\left(\alpha - t_0^2\right)}{\left(t^2 - 2tt_0 + \alpha\right)^{3/2}}\ . \tag{5.103}$$

We have been working in 2-d, one time and one space. Let us see an example that involves all 4 dimensions. Our spacetime could be visualised as consisting of a sequence of hyper-cylinders with only one spacial dimension curved and the other two flat. They effectively reduce two dimensions, because they only enter into the metric trivially. To see the effect of the other two spacial dimensions, consider the hyper-cylinders replaced by hyper-paraboloids, given by

$$t = t_0 + \tanh(t_0/T)(x^2+y^2+z^2)/X^2\ , \tag{5.104}$$

the above analysis, with x replaced by the spherical radial coordinate r, would hold. Similarly, if we had hyper-parabolic cylinders, given by

$$t = t_0 + \tanh(t_0/T)(x^2 + y^2)/X^2 \ , \tag{5.105}$$

we would replace x by the cylindrical polar radial coordinate ρ. We would then have a radial acceleration.

A more complicated situation can arise if we have circular motion, so that the acceleration and velocity are in different directions. For motion in a circle of radius a with angular speed ω (taken to be constant here), the instantaneous Lorentz transformations in cylindrical polar coordinates are (see SR)

$$\left.\begin{array}{lll} t' = \gamma t \ , & \rho' = a = \rho \ , & \varphi' = \varphi - \gamma\omega t \ , \\ z' = z \ , & \gamma = \left(1 - a^2\omega^2/c^2\right)^{-1/2} \ . & \end{array}\right\} \tag{5.106}$$

(Incidentally there are two typographical errors in SR in these equations: γ is written as multiplying $(\varphi - \omega t)$ instead of only ωt and the fact that $\rho = a$ is missed.) The time period is $2\pi\gamma/\omega$ instead of $2\pi/\omega$, i.e. it is a proper period in t' instead of in t. The inverse transformation is

$$t = \gamma^{-1}t' \ , \quad \rho = a = \rho' \ , \quad \varphi = \varphi' + \gamma\omega t \ , \quad z = z' \ . \tag{5.107}$$

The problem here is that the coordinates cannot be extended off the circle while maintaining the matching of $\varphi = 0$ with $\varphi = 2\pi$ for $\rho \neq a \neq \rho'$. We will not discuss rotating frames here. (An excellent discussion is in [30].)

Apart from obtaining a more intuitive understanding of General Relativity (about which there will be more in the next section), there is another, more compelling, reason for the $(3 + 1)$-split. Quantum Theory regards space and time in entirely different ways. Mathematically, the position measuring operator is well defined as a *linear* operator on a Hilbert space. If a time measuring operator were to be defined it would be anti-linear. Again, canonical quantization of a classical Field Theory requires the definition of equal *time* commutation relations for spatially separated (local) operators. There is no corresponding *spatial* requirement for operators at different times. There are various other points relevant for a discussion of the difference between the status of time and space in Quantum Theory, but they will not be elaborated here. Taking this point of view, Arnowitt, Deser and Misner (ADM) developed a $(3+1)$-splitting for the purpose of *canonical quantization of gravity* [31]. A detailed discussion of this split is given in [15]. This attempt, like all other attempts to formulate a consistent theory of Quantum Gravity to date, was not successful. Essentially because there was too much freedom of choice still for the frame still left. In an attempt to avoid this problem Wheeler (with C.M. De Witt) proposed an analogue of the Schrödinger equation (called the Wheeler-De Witt equation) [36]. However, this was too complicated to provide meaningful attempts at quantization. It should be mentioned that there are also *covariant quantization* attempts (see e.g. B.S. De Witt's attempt [32]) that do not require a $(3 + 1)$-split, but they have other drawbacks. There is also the *twistor quantization* attempt that avoids any such split, and quantizes the geometry itself [22, 34, 44] but it is very complicated. These issues will be briefly discussed further in the last section of the book.

5.9 General Relativity in Terms of Forces

Following the spirit of the $(3 + 1)$-split one can go further. For an intuitive understanding of the results of General Relativity it would be useful to know how "Einstein's Law of Gravitation" differs from Newton's. Of course, Einstein's theory replaces the concept of "gravitational force" by Geometry. Thus the bending of the path of the test particle in a gravitational field is due, not to a force, but to the curvature of the spacetime. One could state Einstein's "law of motion" as: test particles move along the straightest *available* path unless an external (i.e. non-gravitational) force acts on it. If the external force is given by $\mathcal{F}^\mu$ and the mass of the test particle is m, the equation of motion is

$$\ddot{x}^\mu + \{_\nu \ ^\mu \ _\rho\} \dot{x}^\nu \dot{x}^\rho = \mathcal{F}^\mu / m \ . \tag{5.108}$$

The "availability" of the path is given by the Einstein field equations (Eqs. (4.29) or (4.30)). These equations replace Newton's law of gravitation.

It is far from satisfying to replace Newton's law by a set of coupled, non-linear, partial differential equations. One wants to know what they mean — to understand them in terms of the "gravitational force". In effect, in the absence of $\mathcal{F}^\mu$, one is asking for the "force" in a Minkowski space which will bend the straight line path of the test particle to give the General Relativistic path. This force, called the *pseudo-Newtonian* (ψ-N) *force*, is the GR generalization of the Newtonian gravitational force. Notice that this approach is radically different from linearization as strong gravitational field effects can be looked at in Newtonian terms.

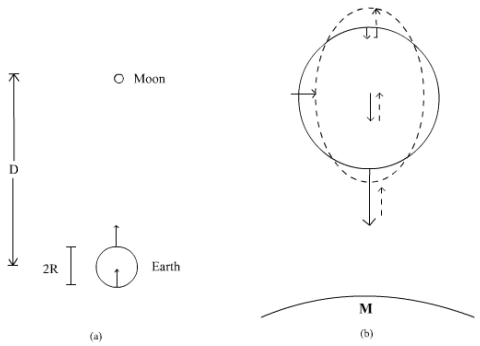


Figure 5.6: Newtonian tidal forces are exemplified by the pull of the Moon on opposite sides of the Earth. The closer part is pulled more than that further away. Subtracting off the pull at the centre of mass, the closer side is pulled *towards* the Moon and the further one pushed *away* from the Moon. Thus the sphere is distorted into an ellipse. This results in a high tide when the Moon is overhead and when it is underfoot.

The precise formulation of the ψ-N force is based on an analysis of Newtonian tidal forces. The tides on Earth are caused by the Moon's gravitational pull. Let the Moon's mass be denoted by m_M and consider two test particles on the Earth's surface of mass μ each, see Fig. 5.6. If the radius of the Earth is R and the distance from the Earth's centre to the Moon's centre is D, the difference between the gravitational pulls on the two test particles, placed so that they lie

in a straight line with the Moon, is

$$F_T = -Gm_M\mu\left(\frac{1}{(D+R)^2} - \frac{1}{(D-R)^2}\right) . \tag{5.109}$$

Since $D \gg R$ we can approximate this *tidal force*, using the binomial expansion for the two terms, by

$$F_T \approx \frac{2Gm_M\mu}{D^2}\frac{2R}{D} + \mathcal{O}\left(\frac{R^3}{D^3}\right) . \tag{5.110}$$

Whereas solids resist this tidal force water flows to accommodate it. Thus the sea closer to the Moon is pulled *more* than the centre of the Earth and *rises*, while the sea on the other side of the Earth is pulled less than the centre, and hence also *rises* relative to the Earth's centre (though it *falls* relative to the Moon). Further, the sea at the right angles to the line joining the centres of the Earth and Moon has to supply the extra water and, consequently, *falls* relative to the Earth's centre.

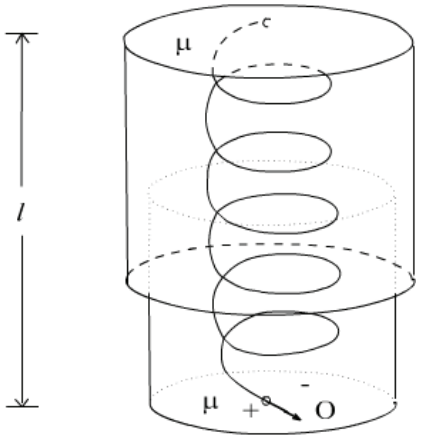


Figure 5.7: An instrument to measure the acceleration experienced by a body in a freely falling frame. By analogy with the speedometer, this could be called an accelerometer. It consists of two frictionless cylinders, one fitting snugly in the other. Each has a massive disc at its end and the two are connected by a spring. One of the discs has a dial on marked on it and the spring ends in a needle that is pivoted on the disc at O. As the spring gets stretched the needle moves to one side, which will be marked $+ve$. If it is squashed, the needle moves to the other, $-ve.$, side.

General Relativistically we have a Schwarzschild metric with two test masses in its field. Imagine these test masses connected by an ideal spring. The masses may be thought of as discs, with one of them marked off as a dial. The spring is attached to this dial so that its end can swivel on it like a needle, see Fig 5.7. If the length of this device is l, the mass of the gravitational source m and the distance from it to the centre of the device r, the tidal force measured on the dial will be that given by Eq.(5.110), with the replacement of m_M by m, $2R$ by l and D by r. Relativistically, this device measures the acceleration vector of

Eqs.(3.151) and (4.24). The placement of the device gives the separation vector, p^μ, in its rest-frame. Maximizing the dial reading places the device in the radial direction. Thus $p^\mu = l\delta^\mu_1 \left(-g^{11}\right)^{1/2}$. Further, since t^μ is the unit timelike vector $t^\mu = c\delta^\mu_0 \left(g^{00}\right)^{1/2}$. Defining the tidal force as $\mu\mathcal{A}^\mu$, Eq.(3.161) gives

$$\overline{F}^\alpha_T = -c^2\mu R^\alpha{}_{010}\, l \left(g^{00}\right)\left(-g^{11}\right)^{1/2} \ , \tag{5.111}$$

where $\overline{F}^\mu_T$ is the maximum tidal force.

Let us calculate this value, using the fact that $g_{00} = e^v = -1/g_{11}$, and the Christoffel symbols given in Eqs.(4.65) and (4.66).

$$\begin{aligned} R^\alpha{}_{010} &= \left\{0\ {}^\alpha\ 0\right\}_{,1} - \left\{0\ {}^\alpha\ 1\right\}_{,0} + \left\{\beta\ {}^\alpha\ 1\right\}\left\{0\ {}^\beta\ 0\right\} - \left\{\beta\ {}^\alpha\ 0\right\}\left\{0\ {}^\beta\ 1\right\} \\ &= \delta^\alpha_1\left[\left\{0\ {}^1\ 0\right\}_{,1} + \left\{1\ {}^1\ 1\right\}\left\{0\ {}^1\ 0\right\} - \left\{0\ {}^1\ 0\right\}\left\{0\ {}^0\ 1\right\}\right] \\ &= \delta^\alpha_1\left[\frac{1}{2}\left(v'e^{2v}\right)'' - \frac{1}{4}v'^2e^{2v} - \frac{1}{4}v'^2e^{2v}\right] \\ &= \frac{1}{2}e^v\left(e^v\right)''\delta^\alpha_1 \ . \end{aligned} \tag{5.112}$$

Inserting the value of e^v from Eqs.(4.72) and (4.75)

$$\begin{aligned} F^\alpha_T &= -\frac{1}{2}\mu l\sqrt{1-2Gm/c^2r}c^2\left(1-2Gm/c^2r\right)''\delta^\alpha_1 \\ &\approx \left[\frac{2Gm\mu}{r^2}\frac{l}{r} + \mathcal{O}\left(\frac{l^3}{r^3}\right)\right]\delta^\alpha_1 \ . \end{aligned} \tag{5.113}$$

Hence, the tidal force in both cases is the same (to order l^3/r^3 in any case). Using the same argument for other metrics we could obtain the tidal force for any general metric and maximise it.

This tidal force is *not* the ψ-N force. In Newtonian gravity

$$\overline{\mathbf{F}}_T = (\mathbf{l}\cdot\nabla)\,\mathbf{F} \ , \tag{5.114}$$

where $\mathbf{F}$ is the Newtonian gravitational force and $\mathbf{l}$ is the 3-d separation vector. Since $\mathbf{F}$ is central $\mathbf{l}\cdot\nabla = l^1\partial/\partial r$. Thus we get $\mathbf{F}_T$ in the radial direction and l^1 (the radial component of $\mathbf{l}$) is all that is relevant, with $\overline{\mathbf{F}}_T$ being the maximal value of the tidal force. Generally

$$F^\alpha_T = -\mu c^2\left(g_{00}\right)^{-1}R^\alpha_{0\beta0}l^\beta \ . \tag{5.115}$$

The extremal value will be obtained by solving the eigenvalue equation for $R^\alpha_{0\beta0}$. In that case $\overline{F}^\alpha_T \propto l^\alpha$. Thus $\overline{F}^0_T$ will be zero and $\overline{F}^i_T \propto l^i$. For a static spacetime we have $\left\{0\ {}^\alpha\ \beta\right\}_{,0} = 0$ and using Riemann normal coordinates, products of Christoffel symbols can be neglected. Thus

$$\begin{aligned} R^\alpha{}_{0\beta0} &= \left\{0\ {}^\alpha\ 0\right\}_{,\beta} - \left\{0\ {}^\alpha\ \beta\right\}_{,0} + \left\{\beta\ {}^\alpha\ \gamma\right\}\left\{0\ {}^\gamma\ 0\right\} - \left\{0\ {}^\alpha\ \gamma\right\}\left\{0\ {}^\gamma\ \beta\right\} \\ &= \left\{0\ {}^\alpha\ 0\right\}_{,\beta} \ . \end{aligned} \tag{5.116}$$

Comparing Eqs.(5.114) and (5.116) we see that

$$F^i = -\mu c^2\left(g_{00}\right)^{-1}\left\{0\ {}^i\ 0\right\} \tag{5.117}$$

is the ψ-N force.

Defining the potential as the quantity whose gradient is the negative of the force, i.e.

$$F_i = -\Phi_{,i} \ , \tag{5.118}$$

the ψ-N potential is simply

$$\Phi = \mu c^2 \ln \sqrt{g_{00}} \ , \tag{5.119}$$

where it is implicit in this derivation that $g_{0i} = 0$ and there is no time dependence of the metric coefficients, in the sense of the time coordinate being used. Let us consider what these quantities are for some other metrics.

For the Reissner-Nordström metric Eqs.(5.111) and (5.112) give the maximal tidal force due to a massive point charge, with e^v given by Eq.(5.45) and α given by Eq.(4.75). Thus

$$\begin{aligned} F_T^\alpha &= -\frac{1}{2}\mu c^2 l \sqrt{1 - \frac{2Gm}{c^2 r} + \frac{GQ^2}{c^4 r^2}} \left(1 - \frac{2Gm}{c^2 r} + \frac{GQ^2}{c^4 r^2}\right)'' \delta_1^\alpha \\ &\approx \left[\left(\frac{2Gm\mu}{r^2} - \frac{3GQ^2\mu}{c^2 r^3}\right) + \mathcal{O}\left(\frac{l^2}{r^2}\right)\right] \frac{l}{r} \delta_1^\alpha \ . \end{aligned} \tag{5.120}$$

Thus we see the Relativistic "correction" to Newton's expression for the tidal force. To understand what this means it is necessary to consider another thought experiment.

In the measurement of the tidal force due to a point mass if we replace the point mass by a point charge and the test masses by test charges of the same sign as each other and opposite that of the source, we would measure the electrostatic attraction, instead of the gravitational attraction. Now let the test charges have the *same* sign as the source. This would give the tidal force due to *repulsion*. Here the closer charge would be repelled more than the further one and so we would have a *squeezing* of the spring instead of a stretching. In this case the needle would turn in the *opposite* direction. Thus the negative part of the tidal force corresponds to a repulsion. This repulsion is of a *neutral test particle due to charge on the gravitational source*! Here we see a "mixing" of electric and gravitational effects which could be regarded, in some ways, as a unification of the forces in GR. (There are other aspects of unification which are not so clearly satisfied here.)

The ψ-N force and potential here are

$$\begin{aligned} F^i &= -\mu c^2 \left(e^{-v}\right) \left(v' e^{2v}\right) \delta_1^i \\ &= -\mu c^2 \left(e^v\right)' \delta_1^i \\ &= -\left(\frac{Gm\mu}{r^2} - \frac{GQ^2\mu}{c^2 r^3}\right) \delta_1^i \ . \end{aligned} \tag{5.121}$$

$$\begin{aligned} \Phi &= \mu c^2 \ln \sqrt{1 - \frac{2Gm}{c^2 r} + \frac{GQ^2}{c^4 r^2}} \\ &\approx \frac{Gm\mu}{r} - \frac{GQ^2\mu}{2c^2 r^2} \ . \end{aligned} \tag{5.122}$$

Here we see the Reissner-Nordström repulsion very clearly, specially in Eq. (5.121). Putting $Q = 0$ we recover the usual Newtonian gravitational force

and potential. We can now answer, in this case, how Einstein's law differs from Newton's law of gravitation! We will return to this point later.

Now consider the Kerr metric given by Eqs.(5.50)-(5.52). Here $g_{0i} \neq 0$. As such $(g_{00})^{-1} \neq g^{00}$. Thus Φ can not be expressed as the natural logarithm of g_{00}, multiplied by $\frac{1}{2}\mu c^2$. However, in this case we can approximate it by

$$\begin{aligned} \Phi &\approx \frac{1}{2}\mu c^2 \left(g_{00} - 1\right) \\ &= -\frac{Gm\mu r}{r^2 + a^2\cos^2\theta/c^2} \ . \end{aligned} \tag{5.123}$$

This, being a function of r and θ, the ψ-N force is *not central*! The ψ-N force then becomes

$$F^r = -\frac{Gm\mu}{r^2}\frac{\left(1 - a^2\cos^2\theta/r^2c^2\right)}{\left(1 + a^2\cos^2\theta/r^2c^2\right)\left(1 - 2Gm/c^2r + a^2/r^2c^2\right)} \ , \tag{5.124}$$

$$F^\theta = -\frac{Gm\mu}{r^2}\frac{2a^2\sin\theta\cos\theta/r^2c^2}{r\left(1 + a^2\cos^2\theta/r^2c^2\right)^3} \ . \tag{5.125}$$

Here the polar component with upper index has units of length inverse times the radial component while the lower index has units of length times the radial component. The Einstein correction is dramatic here, introducing a polar component and polar dependence of the force.

The modifications mentioned here will be discussed more thoroughly in the next chapter. Here it should just be mentioned that the ψ-N force is obtained, essentially, by "integrating the tidal force" in some sense. As such there is an "integration constant", which is actually a function annihilated by $(\mathbf{l}\cdot\nabla)$. This is set by requiring that Minkowski space (in any coordinates) have no ψ-N force. Again the ψ-N potential is obtained by integrating the ψ-N force. We require, again, that the ψ-N potential of Minkowski space be zero. Details of this formalism are given in [37]. It is worth mentioning, here, that the ψ-N frame is a special case of *Fermi normal coordinates* which are attached to the "freely falling rest frame", by which is meant the frame of reference of an observer falling freely from infinity. Details of Fermi normal coordinates may be found in [15].

The ψ-N formalism has been modified and extended to spacetimes that are time-varying. This $e\psi$-N formalism does *not* use Fermi normal coordinates, but requires a synchronisation like condition, whereby $g_{0i} = 0$ and the time-like isometry is replaced by the unit tangent vector to the geodesic. Now we require both Riemann normal coordinates and $g_{0i} = 0$. It is not clear whether both conditions are consistent in all spacetimes. If they are, it is not clear that they provide a unique coordinate system. However, when used for cylindrical and plane gravitational waves the $e\psi$-N formalism does give the momentum (or energy) imparted to test particles. This work is of a more speculative nature and is given in [38].

5.10 Exercises

1. Assume a plane symmetric, static metric and a pure electric field (with no source. Construct the Einstein field equations for these requirements. Try to solve it. Can we have a pure magnetic field with plane symmetry?

2. Assume an axially symmetric stationary spacetime and a pure magnetic field in the axial direction. Construct the Einstein field equations in this case and try to solve it.

3. The cylindrical gravitational wave solution given by Eqs.(5.75), (5.81) and (5.83), is exact. Find the gauge which converts the nonlinear vacuum equations in the given coordinates into the wave equation. Determine the remnant terms as the "pseudo-tensor" giving the energy, momentum and stresses in classical terms.

4. The linearly polarized exact plane gravitational wave solution given by Eqs.(5.84) and (5.89) involves an arbitrary choice of the function $\beta(u)$. Make any choice other than Eq.(5.91) and hence obtain $\Omega(u)$ to get the full solution. now work out the pseudo-tensor for this choice.

5. Instead use the circularly polarized exact plane gravitational wave solution given by Eqs.(5.84) and (5.92) involve two arbitrary functions. Determine what choice of $\theta(u)$ reduces to the linearly polarized case. Avoiding that choice, make any other choice of $\theta(u)$ and $\beta(u)$ and now determine the pseudo-tensor giving its interpretation.

6. For the Schwarzschild metric, call the unit tangent vector for a freely falling observer, $\mathbf{n}(s)$ and denote the unit tangent vector to the hypersurface orthogonal to this vector with constant (θ, ϕ) by $\mathbf{t}$. Determine these vectors and hence obtain an equation for the time coordinate, t, in terms of the radial coordinate, r. Use this equation to obtain the hypersurfaces. Since this observer sees a Minkowski spacetime these hypersurfaces satisfy the flatness condition, $R^{\mu}_{\nu\rho\pi} = 0$. Verify that this condition is satisfied. These coordinates only work outside the horizon. Convert the foliation to Kruskal coordinates, so that you can penetrate it. Now convert to the compactified coordinates so that you can get a complete picture of the foliated spacetime. Repeat the above for the Reissner-Nordström metric.

7. What is the problem faced in determining the ψ-N force, and hence potential, for cylindrical gravitational waves? Try to find a way around the problem along the lines indicated at the end of the chapter and given in [38]. Compare the energy and momentum so obtained with that of the pseudo-tensor. Why is there a difference? Repeat the process above for the linearly polarized and circularly polarized plane gravitational wave solutions.

Chapter 6

Black Holes

One of the predictions of General Relativity that has most captured the public imagination is the existence of "black holes" — objects which can not be seen but can pull things in. They have been variously described as exciting, awe-inspiring, weird, terrifying and other similar adjectives. They seem to be the very stuff of which science fiction is made. In this chapter we will discuss what they are, their mathematical treatment, their physical properties and their current observational status. Before going on to the Relativistic discussion we will first consider them classically, i.e. in terms of Newton's theory of gravitation.

6.1 The Classical Black Hole

Consider a test particle of mass μ moving in the gravitational field of a body of mass m with an instantaneous speed v and radial distance from the centre of the mass of the body, r. It will have kinetic and potential energies

$$KE = \frac{1}{2}\mu v^2 \ , \ PE = Gm\mu/r \ . \tag{6.1}$$

If $PE > KE$ the particle is "bound" to move in a closed (elliptical) orbit. If, on the other hand, $PE < KE$ it is "free" to move in an open (hyperbolic) orbit. At the boundary between the two cases it moves in a parabolic orbit and is said to be at *escape velocity*, v_{esc}. Thus by definition,

$$\frac{1}{2}\mu v_{esc}^2 = Gm\mu/r \ , \tag{6.2}$$

and hence the escape velocity is given by

$$v_{esc} = \sqrt{2Gm/r} \ . \tag{6.3}$$

An object for which $v_{esc} \geq c$ at its surfaces would be totally dark, but would continue to pull things into itself due to its gravitational attraction. On account of these properties John Wheeler dubbed it a *black hole* and the name has stuck.

The earliest suggestion of such objects not only pre-dates Wheeler, it pre-dates General Relativity, coming from Michell (1783) and Laplace (1796). According to Laplace's theory of planetary formation, stellar systems evolved from enormous swirling gas clouds. These clouds underwent gravitational collapse

and half the kinetic energy of collapse went into heating the clouds (the other half, according to the virial theorem of Classical Mechanics, going into balancing the potential energy). Due to the conservation of angular momentum the collapse would be accompanied by an enhancement of the rotation rate. If the density of the cloud, ρ, was approximately constant and rotated rigidly with an angular speed ω, then

$$\omega a^2 = \text{constant} \ , \tag{6.4}$$

where a is the radius of the cloud. At some stage it will be energetically favourable for a ring, carrying a large angular momentum, to be left in place while the rest of the cloud continued its collapse. The matter in the ring could coalesce to form a planet and the cloud would form the star. This model is consistent with the fact that Jupiter carries about 60% of the total angular momentum of the Solar system, while the Sun has a meagre 2%. Of course, it is now known that the kinetic energy of collapse is inadequate for the Sun to have been burning for about 4.6 billion years (as it has done). The main source of energy for the Sun is *not* gravitational collapse but nuclear fusion.

Laplace further argued that as the star continued its gravitational collapse it would become denser. At some critical stage it would become so dense that the escape velocity at its surface would exceed the speed of light. Hence it would disappear from view. From Eq.(6.2) this would occur when

$$a < 2Gm/c^2 \ . \tag{6.5}$$

Writing m in terms of ρ and a, this condition becomes

$$\rho > 3c^2/8\pi G a^2 \ . \tag{6.6}$$

For a solar *mass* ($\sim 2 \times 10^{30}$kg) object its radius would be about ~ 3 km. For a solar *size*, ($\sim 7 \times 10^8$m), object the mass would be ($\sim 5 \times 10^{35}$ kg), roughly that of a globular cluster of stars. In the former case the density would be nearly 2×10^{20} kg/m^3 and in latter case the mass would be about 4×10^9 kg/m^3.

The assumption implicit in Laplace's argument is that light would be affected by gravity. Though this view had been propounded by Newton, it was not generally accepted. The view seems consistent with a corpuscular theory of light, but it would not be apparent (in Newtonian terms) why a *wave* should need an escape velocity. Thus the argument of Laplace (and some others) was forgotten and the premature birth of black holes was aborted. Nevertheless, it is worth remembering that it was Michell and Laplace who first not only proposed the existence of black holes but also of *dark matter* (which is much discussed these days) in the Universe.

6.2 The Escape Velocity for the Schwarzschild Metric

An object at escape velocity does not feel the effect of gravity, and is hence, in free-fall. As such, for the relativistic analysis we will use the freely falling frame. We can use the same calculations as were used in SR when deriving the Schwarzschild metric. In the instantaneous Lorentz frame, at a radial distance r from the mass m, there is a time dilation, a length contraction in the radial direction (but no change in the radial parameter.) and no change orthogonal to

it, so that

$$\left.\begin{aligned} dt' &= \left(1-2Gm/c^2r\right)^{1/2} dt \ , \\ dr' &= \left(1-2Gm/c^2r\right)^{-1/2} dr \ , \\ d\theta' &= d\theta \ , \qquad d\varphi' = d\varphi \ . \end{aligned}\right\} \tag{6.7}$$

For convenience we define two new spherical polar angles (ψ, χ) instead of (θ, φ), so as to avoid confusion. The choice is such that the new z-axis lies along the radial direction. We can now write the velocity components in the new xyz frame using Eq.(6.7), see Fig. 6.1.

$$\left.\begin{aligned} v\cos\psi &= (dr'/dt') \\ &= \left(1-2Gm/c^2r\right)^{-1} (dr/dt) \ , \\ v\sin\psi\cos\chi &= r\sin\theta (d\varphi'/dt') \\ &= \left(1-2Gm/c^2r\right)^{-1/2} r\sin\theta (d\varphi/dt) \ , \\ v\sin\psi\sin\chi &= r(d\theta'/dt') \\ &= \left(1-2Gm/c^2r\right)^{-1/2} r(d\theta/dt) \ . \end{aligned}\right\} \tag{6.8}$$

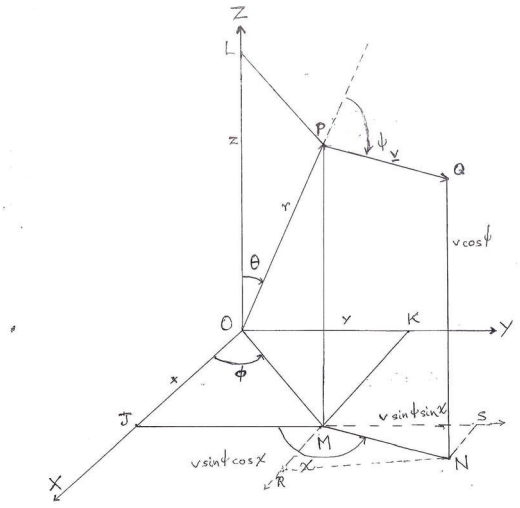


Figure 6.1: The point P has Cartesian coordinates (x, y, z) and spherical coordinates (r, θ, ϕ). For a vector $\mathbf{v}$, we define ψ as the angle between $\mathbf{v}$ and the extension of $\overrightarrow{OP}$. Thus the projection of $\mathbf{v}$ on the XY-plane has magnitude $v\sin\psi$. The angle this projection makes with the Y-axis is χ.

We now consider the Schwarzschild metric given by Eqs.(4.74, 4.75) and

divide it through by $(dt)^2$. Using Eq.(4.91) for $\dot{t}^2$ we have

$$\begin{aligned}\left(1-2Gm/c^2r\right)^2 &= \left(1-2Gm/c^2r\right) - \left(1-2Gm/c^2r\right) v^2\cos^2\psi/c^2 \\ &\quad - \left(1-2Gm/c^2r\right) v^2\sin^2\psi\sin^2\chi/c^2 \\ &\quad - \left(1-2Gm/c^2r\right) v^2\sin^2\psi\cos^2\chi/c^2 \\ &= \left(1-2Gm/c^2r\right)\left(1-v^2/c^2\right) . \end{aligned} \tag{6.9}$$

Simplifying Eq.(6.9) yields the speed of the free-fall frame exactly as given by Eq.(6.2)

$$v_{esc}^2 = 2Gm/r . \tag{6.10}$$

Thus Laplace's argument is relativistically valid, with the additional support that light is affected by gravitation according to General Relativity. Further, since the speed of light is the limit for physically attainable speeds, *nothing* can emerge from less than the radial distance

$$r_s = 2Gm/c^2 . \tag{6.11}$$

This distance is called the *Schwarzschild radius*. For example, for the Sun the Schwarzschild radius is about 2.953 km. For the Earth, with a mass of nearly 6×10^{24} kg it is about 0.4438 cm. Our galaxy, the Milky Way is believed to have a mass of about 1.8×10^{11} solar masses and hence its Schwarzschild radius would be about 2.7×10^{11} km, which is about 9 light days. Clearly, none of the above objects is a black hole as their actual sizes are incomparably larger than their Schwarzschild radii.

6.3 The Black Hole Horizon

Many strange things occur at the Schwarzschild radius $r_s = 2Gm/c^2$, in the Schwarz- schild metric, given by Eqs.(4.74, 4.75). The most obvious is that g_{00} is zero and g_{11} is infinite there. Thus the metric appears to be singular. Is this singularity an indication of some genuine physical feature or just a bad choice of coordinates? If the latter is the case the argument regarding the Schwarzschild radius being the place where the escape velocity is the speed of light, could be erroneous. It is *important*, therefore, to answer this question. Before proceeding it is necessary to consider what would be meant by "physical features". Bad coordinates can be like $r = 0$ on the plane in polar coordinates or due to a bad choice of frame of reference. A change of frame of reference can introduce fictitious forces, which are physical in some ways but not in others. We are interested in non-observer dependent features.

We first check whether there is an essential singularity where the curvature becomes infinite. By Eq.(4.24) the acceleration would become infinite where the curvature is infinite. This is, consequently, called a *crushing singularity* also. Since we want to avoid observer dependent effects we can not simply look at the metric or curvature tensor components, which are coordinate dependent quantities, but need to use curvature invariants. In 4-dimensions there are 20 independent non-zero components of the Riemann tensor and 10 of the Ricci tensor, in general. There being some gauge (coordinates) freedom in their specification, there are fewer independent curvature invariants. It turns out [22]

that, in general, there are 14 invariants. For a vacuum 10 of these automatically disappear as $R_{\mu\nu} = 0$. Thus we are only left with 4. Two of these are constructed from the single and two sided duals of the Riemann tensor

$$ {}^*R^{\alpha\beta}{}_{\rho\pi} = \tfrac{1}{2} e^{\alpha\beta\gamma\delta} R_{\gamma\delta\rho\pi} \quad , \quad {}^*R^{*\alpha\beta\mu\nu} = \tfrac{1}{4} e^{\alpha\beta\gamma\delta} R_{\gamma\delta\rho\pi} e^{\rho\pi\mu\nu} \ . \tag{6.12} $$

The former turns out to be orthogonal to the Riemann tensor in this case while the latter is simply its negative. As such we only need to check the two curvature invariants $\mathcal{R}_2$ and $\mathcal{R}_3$ given by Eqs.(3.123 and 3.124).

Using Eqs.(4.66) for the non-zero Christoffel symbols of the Schwarzschild metric, the non-zero independent components of the Riemann tensor for this metric are

$$ \left.\begin{aligned} & R^0{}_{101} = -\tfrac{1}{2}\nu'' + \tfrac{1}{4}\nu'\lambda' - \tfrac{1}{4}\nu'^2 \ , \\ & R^0{}_{202} = -\tfrac{1}{2} r\nu' e^{-\lambda} \quad , \quad R^0{}_{303} = R^0{}_{202} \sin^2\theta \ , \\ & R^1{}_{212} = \tfrac{1}{2} r\lambda' e^{-\lambda} \quad , \quad R^1{}_{313} = R^1{}_{212} \sin^2\theta \ , \\ & R^2{}_{323} = (1 - e^{-\lambda}) \sin^2\theta \ . \end{aligned} \right\} \tag{6.13} $$

The Riemann tensor components can be written more conveniently in the form $R^{\alpha\beta}{}_{\gamma\delta}$. Using the fact that $\lambda(r) = -\nu(r)$ and Eqs.(4.91 and 4.94) we have

$$ \left.\begin{aligned} & R^{01}{}_{01} = R^{23}{}_{23} = -2Gm/c^2 r^3 \ , \\ & R^{02}{}_{02} = R^{03}{}_{03} = -R^{12}{}_{12} = -R^{13}{}_{13} = Gm/c^2 r^3 \ . \end{aligned} \right\} \tag{6.14} $$

By the symmetry properties of the Riemann tensor embodied in the first Bianchi identities, given by Eqs.(3.114 and 3.115) each term in the expression for $\mathcal{R}_2$ appears 4 times and $\mathcal{R}_3$ appears 8 times. Thus

$$ \left.\begin{aligned} \mathcal{R}_2 &= 4\left[(R^{01}{}_{01})^2 + (R^{02}{}_{02})^2 + (R^{03}{}_{03})^2 \right. \\ &\quad \left. + (R^{12}{}_{12})^2 + (R^{13}{}_{13})^2 + (R^{23}{}_{23})^2\right] \\ &= 48 G^2 m^2/c^4 r^6 \ , \\ \mathcal{R}_3 &= 64 G^3 m^3/c^6 r^9 \ . \end{aligned} \right\} \tag{6.15} $$

Hence there is a physical, essential (crushing), singularity at $r = 0$, but there is none at $r = r_s$ where $\mathcal{R}_2 = 3c^8/4G^4m^4$ and $\mathcal{R}_3 = 5c^{12}/8G^6m^6$. Thus $r = r_s$ is only a coordinate singularity! However, this does not necessarily imply that there are no physical features there. We continue our search for other physical features.

From Eq.(4.102) light of finite frequency, ν_A, emitted from $r_A = r_s$, will be infinitely red-shifted in Schwarzschild coordinates, which are the relevant coordinates for an observer at rest at infinity. Thus $r = r_s$ is a *red-shift horizon* and the energy of light emitted from there would appear to be zero at infinity. However, as will become clearer later, a freely falling observer would see no such red-shift. Observers falling non-freely would see some red-shift and the red-shift horizon would appear for $r < r_s$. Thus this feature is observer dependent. Our search must continue.

A very crude analysis, which will be refined later, shows that $r = r_s$ is a *null hypersurface*, i.e. vectors on this hypersurface are null vectors, like the

null cone. Ofcourse here one is thinking of future directed vectors. Consider a velocity vector for light emitted radially outwards from $r = r_s$,

$$v^\alpha = (v^t,\, v^r,\, 0,\, 0)\ , \tag{6.16}$$

and hence

$$v^\alpha v^\beta g_{\alpha\beta} = 0 = e^\nu (v^t)^2 - e^{-\nu}(v^r)^2\ . \tag{6.17}$$

Thus we see that

$$\frac{dr}{dt} = \frac{dr/d\tau}{dt/d\tau} = \frac{v^r}{v^t} = e^\nu\ , \tag{6.18}$$

the negative root being excluded as the motion (if any) is radially outwards. At $r = r_s$ we have $e^\nu = 0$ and hence dr/dt or $dr/d\tau$ are zero and thus r remains at r_s. Therefore light remains on the hypersurface $r = r_s$, which is then a null hypersurface. This is an observer independent, genuinely physical feature. Our search has ended. The answer to the question is that though the singularity is not physical there *is* an important physical feature there!

We see that no geodesic, starting inside the red-shift horizon $r = r_s$, can emerge outside it. Thus the spacetime is broken into two disjoint pieces by this hypersurface. This type of hypersurface is called a *trapped surface*. Generally a null trapped surface, which is also an infinite red-shift horizon, is called an *event horizon.*We will see later that other geometries have event horizons and similar (though distinct) features. A *black hole*, generally is a spacetime region clothed by an event horizon.

6.4 Convenient Coordinates to Study the Black Hole

We need to construct coordinates that can cross the event horizon smoothly, so that its physical and geometrical features can be better investigated. Before doing so, we will study the spacetime geometry in Schwarzschild coordinates, to the extent it can be studied, so as to be in a position to understand the other coordinates and the geometry in these terms.

To start with, consider a spacetime diagram with two spatial dimensions suppressed, so that the time parameter, t, varies along the vertical axis and the radial parameter, r, along the horizontal axis. A light ray emitted radially outwards from $r = r_s + \varepsilon$ ($\varepsilon > 0$), according to Eq.(6.18), would have the speed

$$v = [\varepsilon/(r_s + \varepsilon)]c\ . \tag{6.19}$$

Thus, $v \approx \varepsilon c/r_s$ for small ε ($\varepsilon \ll r_s$) and $v \to c$ as $\varepsilon \to \infty$. In units with $c = 1$ the tangent to the path of the light ray is, asymptotically, the line at 45°. As ε is decreases from there the angle of the tangent to the vertical decreases, becoming nearly zero as $\varepsilon \to 0$. The light ray emitted radially inwards from the same point has an equal but opposite velocity. Thus the light ray paths and their time reversed images specify the null cones, as shown in Fig. 6.2. Introducing one more dimension so that only the z- axis is suppressed, the horizon is represented by a vertical cylinder and the singularity at $r = 0$ by a vertical line at its axis. The light cones appear in a more familiar representation in this way, see Fig. 6.3.

Since $v \approx c$ for $r = \infty$ the null cones at infinity are the same as in Minkowski spacetime. As r is decreased from infinity v decreases and hence the null cones

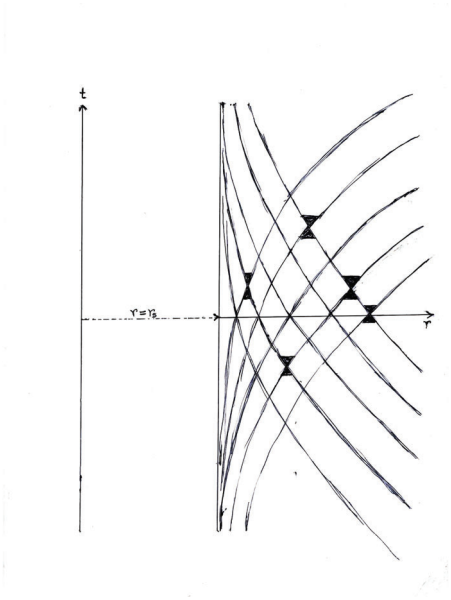


Figure 6.2: The Schwarzschild spacetime in Schwarzschild coordinates shown with two dimensions suppressed. These coordinates do not penetrate to the region $0 \leq r < r_s$; inside the black hole. Notice how the "light cones" squeeze up near the horizon

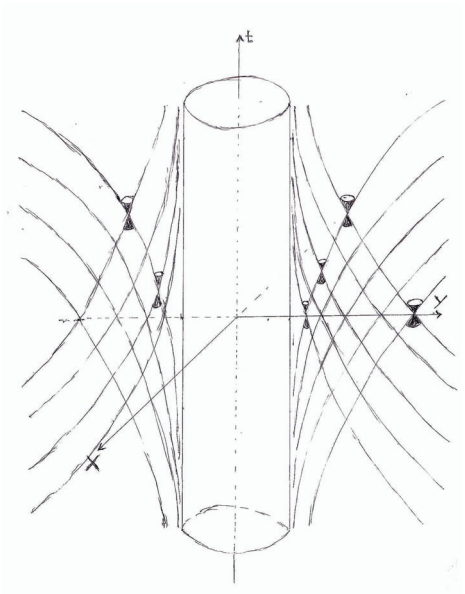


Figure 6.3: The Schwarzschild spacetime in Schwarzschild coordinates with only one dimension suppressed. The light cones appear in a more recognisable form.

start "squeezing up" as the event horizon is approached. At the cylinder $r = r_s$, the event horizon, they have squeezed up totally to a line going up along the cylinder. This is why the event horizon is a null hypersurface. What happens beyond this value can not be validly expressed in Schwarzschild coordinates. However, we continue our investigation at present *as if* they could be used.

Put $r < r_s$ naively into the metric given by Eqs.(4.74 and 4.75). The coefficient of dt^2 becomes negative and of dr^2 positive. Thus t becomes a *spacelike* and r a *timelike* parameter here. While r passes smoothly through the horizon, at r_s, and goes on to zero, $t \to \infty$ as $r \to r_s$ from above and comes down from ∞ as $r \to 0$. Thus the t parameter inside the horizon bears no relation to that outside. Drawing the light cone in these coordinates, as we did outside the horizon, the cone starts out spread over the whole of the cylinder (looking locally like a disc) at $r = r_s$ and then squeezes up as r decreases, approaching a line again, at $r = 0$, as shown in Fig. 6.4. The significance attaching to this observation is not clear as the coordinates used are not the usual ones, even though the same symbols are used. An alternative way of depicting the null cone structure is given in Fig. 6.5.

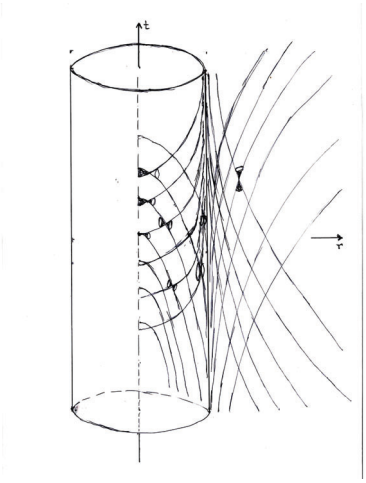


Figure 6.4: The interior of the black hole in "Schwarzschild-like" coordinates. Notice that these coordinates do not connect with the coordinates outside the black hole and that the cones have a sense orthogonal to those outside. Also notice how they start near the horizon really "flattened out" and end up at the singularity "squeezed" into a line.

Before proceeding to the construction of the new coordinate system for the Schwarz- schild spacetime, it is instructive to first see the procedure used in its embryonic form, without the complications introduced by the spacetime curvature. This procedure is to use *null coordinates*

$$v = \tfrac{1}{\sqrt{2}}(ct + r) \quad , \quad u = \tfrac{1}{\sqrt{2}}(ct - r) \ , \tag{6.20}$$

so that

$$ct = \tfrac{1}{\sqrt{2}}(v + u) \quad , \quad r = \tfrac{1}{\sqrt{2}}(v - u) \ , \tag{6.21}$$

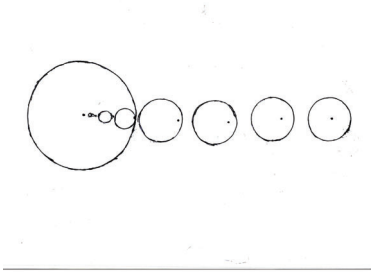


Figure 6.5: A top view of the sequence of light cones entering a black hole. The point is the vertex of the cone "seen from above".

and the Minkowski metric takes the form

$$ds^2 = 2dvdu - r^2 d\Omega^2 \ . \tag{6.22}$$

The metric tensor components in these coordinates are

$$\left.\begin{array}{ll} g_{01} = g_{10} = 1, & g_{22} = -r^2, \\ g_{33} = -r^2 \sin^2\theta, & g_{\mu\nu} = 0 \text{ otherwise,} \end{array}\right\} \tag{6.23}$$

and its determinant is $g = -r^4 \sin^2\theta$. The coordinates are called null because $g_{00} = g_{11} = 0$. Here v plays the role of an *advanced* and u of a *retarded* time. The use of null coordinates is not crucial for our purposes but has been explained because they are commonly used.

To avoid the singularity a new radial coordinate can be defined in which the singularity at $r = r_s$ disappears. This is done by taking

$$r^* = \int \frac{dr}{1 - r_s/r} = r + r_s \ln\left|\frac{r}{r_s} - 1\right| \ , \tag{6.24}$$

where the constant of integration is taken to make the argument of the logarithm dimensionless and convenient. We can use this r^* to define advanced and retarded times in Eqs.(6.20). These give

$$dv = \tfrac{1}{\sqrt{2}}(cdt + \tfrac{dr}{1-r/r_s}) \quad , \quad du = \tfrac{1}{\sqrt{2}}(cdt - \tfrac{dr}{1-r_s/r}) \ , \tag{6.25}$$

in terms of which one can write

$$ds^2 = 2(1 - r_s/r)dv^2 - 2\sqrt{2}dvdr - r^2 d\Omega^2 \ , \tag{6.26}$$

$$= 2(1 - r_s/r)du^2 + 2\sqrt{2}dudr - r^2 d\Omega^2 \ . \tag{6.27}$$

In these coordinates, though $g_{00} = 0$ at $r = r_s$ the determinant remains finite there, being $-2r_s \sin^2\theta$. These are called *Eddington-Finkelstein* (advanced and retarted) coordinates. The null cone structure for both systems is given in Fig. 6.6.

The above coordinate systems seem "asymmetrical" in that the diagonal term in the radial part of the metric tensor has been replaced by an off-diagonal term. If we try to use them simultaneously it is easy to verify that the metric becomes

$$ds^2 = 2(1 - r_s/r)dudv - r^2 d\Omega^2 \ , \tag{6.28}$$

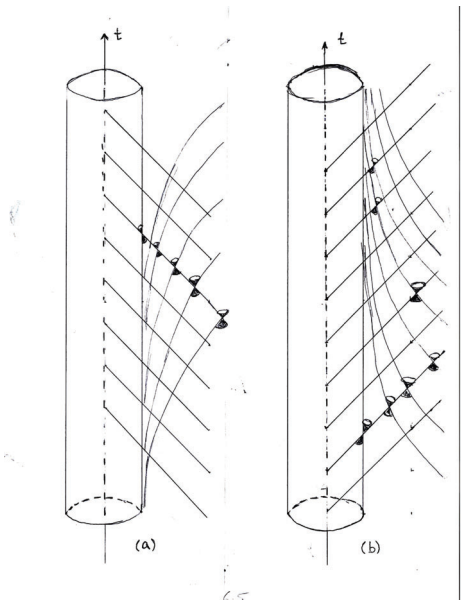


Figure 6.6: The Schwarzschild spacetime in Eddington-Finkelstein: (a) retarded; (b) advanced coordinates. Notice below the light cones tilt up and squeeze as the black hole horizon is approached.

which is singular at $r = r_s$ since $g = 0$ there. For our purposes, we need to construct a new coordinate system which is symmetrical looking and is non-singular.

The singularity arises due to a zero in the metric. As such we may try exponentiating the previous (v, u) coordinates. We introduce two constants to be fixed later for convenience, α and β. Thus

$$V = \alpha e^{v/\beta} \quad , \quad U = -\alpha e^{-u/\beta} \ . \tag{6.29}$$

Hence

$$VU = -\alpha^2 e^{(v-u)/\beta} = -\alpha^2 e^{\sqrt{2}r^*/\beta} \ . \tag{6.30}$$

Using Eq.(6.24) we see that

$$VU = -\alpha^2 \left| \frac{r}{r_s} - 1 \right|^{\sqrt{2}r_s/\beta} e^{\sqrt{2}r/\beta} \ . \tag{6.31}$$

To reduce the above function to its simplest form it is obviously convenient to choose $\beta = \sqrt{2}r_s$. Now taking differentials of Eqs.(6.30), with this choice of β, and multiplying them gives

$$dudv = 2(r_s/\alpha)^2 (\frac{r}{r_s} - 1)^{-1} e^{-r/r_s} dU dV \ . \tag{6.32}$$

Inserting this value into Eq.(6.28) we see that

$$\begin{aligned} ds^2 &= \frac{4r_s^2}{\alpha^2} \frac{1 - r_s/r}{r_s/r - 1} e^{-r/r_s} dU dV - r^2 d\Omega^2 \\ &= \frac{4r_s^3}{\alpha^2 r} e^{-r/r_s} dU dV - r^2 d\Omega^2 \ . \end{aligned} \tag{6.33}$$

Now there is no singularity at $r = r_s$ as $g = -(16r_s^4 \sin\theta/\alpha^2 e)^2 \neq 0$. These coordinates are called *Kruskal coordinates*. Normally α is chosen to be unity but if we want U, V to have units of length it is more convenient to take $\alpha = 2r_s$, in which case $g_{01} = \frac{r_s}{r} e^{-r_s/r} = g_{10}$.

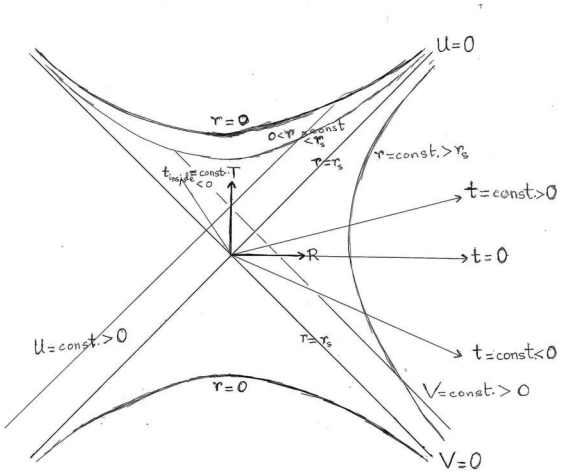


Figure 6.7: The Kruskal picture of the Schwarzschild spacetime. Notice that one coordinate system covers the entire spacetime. Also notice that "mirror images" of regions I and II had to be introduced. Without this introduction the maximal extension of the spacetime is not achieved. Now the null coordinates U, V have the ranges $(-\infty, +\infty)$.

The entire spacetime is covered by a single coordinate patch in Kruskal coordinates. The spacetime picture in these coordinates, called the Kruskal diagram, is shown in Fig. 6.7. The region (I) $V > 0 > U$ is clearly the usual spacetime outside the black hole event horizon; and (II) U, $V < 0$ is the black hole in retarded time coordinates. The second part would not appear to be allowed by Eq.(6.29) but is the extension of the Schwarzschild geometry through the horizon into the black hole interior, where the previous coordinates broke down. As such a negative sign would have to be put in by hand and used in Eq.(6.29). We can also get positive values of U by this procedure. These give region (III) $U > 0 > V$ which is a mirror image of region (I); and (IV) U, $V > 0$ which is a mirror image of (II). The lower region, (IV), gives the black hole interior in advanced time coordinates. Thus we have two separate black hole interiors — the "past black hole" lying behind the past horizon and the "future black hole" hidden by the future horizon. The two spacetime regions in Schwarzschild coordinates became two in advanced Eddington-Finklestein coordinates and two in the retarded so that the total came to three regions. In Kruskal coordinates a fourth, apparently spurious, region called the *maximal extension* is introduced. The complete Kruskal diagram is said to give the maximally extended spacetime.

In the Kruskal diagram $t = const.$ is represented by a straight line emerging from the centre of the diagram. On the right side (region I) the line at 45° going down is $t = -\infty$ and 45° going up is $+\infty$. Thus time goes *upwards* there.

Also, r = constant is represented by a hyperbola to which the 45° lines are asymptotic. In region (I) the hyperbolae further to the right have larger r and those closer to the left have smaller r. The 45° lines are $r = r_s$, being the limit of the hyperbolae as r is decreased to r_s. Though $t = \pm\infty$ here, there is no singularity in (V, U) coordinates. From Eq.(6.31), when $r = r_s$, $UV = 0$. Thus, these are the coordinate axes $U = 0$ (the V-axis) and $V = 0$ (the U-axis). Since these are null coordinates $r = r_s$ is a null hypersurface. This is the proof of the statement referred to earlier. Clearly the vertical hyperbola is a timelike world line in regions (I) and (III) and the horizontal straight line emerging from the origin is the spacelike hypersurface $t = 0$. We could choose other spacelike and timelike hypersurfaces and world lines respectively. In regions (I) and (IV) it is the vertical straight lines that are timelike and the horizontal hyperbolae that give the spacelike hypersurface. This is the observation made earlier, that r is a timelike parameter and t a spacelike parameter inside the black hole. The horizontal hyperbolic end at $r = 0$, which is an essential singularity as is clear from Eq.(6.33). Thus the diagram ends at this hyperbola.

A convenient coordinate system which can be obtained from the Kruskal coordinates, called the *Kruskal-Szekeres coordinates*, has a timelike coordinate, T, and a spacelike coordinate, R, defined by

$$T = \tfrac{1}{\sqrt{2}}(V - U) \quad , \quad R = \tfrac{1}{\sqrt{2}}(V + U) \ , \tag{6.34}$$

by analogy with the Minkowski (t, r) coordinates. Thus, with $\alpha = 2r_s$,

$$ds^2 = \frac{2r_s}{r} e^{-r/r_s}(dT^2 - dR^2) - r^2 d\Omega^2 \ . \tag{6.35}$$

Notice that T has units of R here and not R/c. Also, here

$$\left.\begin{aligned} T &= \sqrt{2}\left|\tfrac{r}{r_s} - 1\right|^{1/2} e^{r/2r_s} \sinh(t/2r_s) \ , \\ R &= \sqrt{2}\left|\tfrac{r}{r_s} - 1\right|^{1/2} e^{r/2r_s} \cosh(t/2r_s) \ , \end{aligned}\right\} \tag{6.36}$$

by which T varies from $-\infty$ to $+\infty$ and R from 0 to ∞. Clearly

$$R^2 - T^2 = 2(\tfrac{r}{r_s} - 1)e^{r/r_s} \quad , \quad T/R = \tanh(t/2r_s) \ . \tag{6.37}$$

This applies to region (I) only. The conversion, of course, *must* break down at the horizon as the (t, r) coordinates do not remain valid. In region (III), the maximal extension, $R \leq 0$. In regions (II) and (IV) the functions get reversed so that T is defined by a "cosh" and R by a "sinh". Also, for continuity, to have R positive when T is negative, in region (IV) there is a negative sign with R and T while in region (II) both coordinates remain positive and have the same signs as above. For continuity, the T must acquire a negative sign in definition, in region (III). Hence we have a usual rectangular grid, that looks like the Cartesian grid, for the Schwarzschild geometry. Notice that at $R = 0$, T^2 lies between -2 and $+2$ and can increase hyperbolically with the increase in R^2.

It is often convenient to use coordinates that have a finite range of values, called *compactified coordinates*, instead of (V, U) or (T, R) which go from $-\infty$ to $+\infty$. For example, if we define

$$\bar{v} = \tan^{-1} V \qquad \text{and} \qquad \bar{u} = \tan^{-1} U \ , \tag{6.38}$$

then $-\infty$ gets mapped on to $-\pi/2$ and $+\infty$ on to $+\pi/2$. Using Eqs.(6.38) in Eq.(6.33) we see that the metric becomes

$$ds^2 = \frac{4r_s^3 e^{-r/r_s}}{r\cos^2\bar{v}\cos^2\bar{u}} d\bar{v}d\bar{u} - r^2 d\Omega^2 \ , \tag{6.39}$$

which is singular at the crushing singularity $r = 0$ and at the coordinate singularities $\bar{v}, \bar{u} = \pm\pi/2$. These new null coordinates have the range $(-\pi/2, \pi/2)$ and can be compactified to $[-\pi/2, \pi/2]$ by using the extended real number system $\overline{\mathbb{R}} = \{-\infty\} \cup \mathbb{R} \cup \{+\infty\}$.

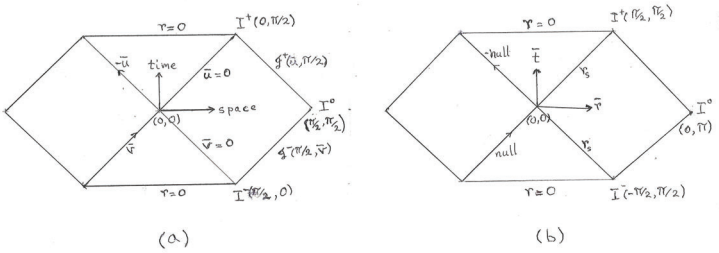


Figure 6.8: (a) The Carter-Penrose diagram in compactified null coordinates, $(\overline{u}, \overline{v})$; (b) in compactified Kruskal-Szekres coordinates $(\overline{t}, \overline{r})$.

One could, similarly, compactify the Kruskal-Szekeres coordinates. This is achieved, for example, by defining

$$\left.\begin{aligned} \bar{t} &= \tan^{-1}(T+R) + \tan^{-1}(T-R) \ , \\ \bar{r} &= \tan^{-1}(T+R) - \tan^{-1}(T-R) \ . \end{aligned}\right\} \tag{6.40}$$

The Carter-Penrose diagram for the compactified Schwarzschild spacetime in compactified null coordinates is given in Fig. 6.8a and the same diagram compactified Kruskal-Szekres coordinates in Fig. 6.8b. The boundaries of this diagram above and below are $r = 0$ and the lower right edge is past null infinity, $\mathcal{I}^-$, while the upper right edge is future null infinity, $\mathcal{I}^+$, read "scri minus" and "scri plus" (which is short for "script I"). The former is reached by past directed null lines while the latter is reached by future directed null lines. The bottom right vertex is past timelike infinity, I^-, and the upper right vertex is future timelike infinity, I^+. These are reached by timelike lines going to the past or future, which do not come from or go into $r = 0$. Finally the right vertex is spacelike infinity, I^0, reached by spacelike lines. A $t = const.$ line in region (I) is a curve going from I^- to I^+. An $r = const.$ line in this region goes from the centre of the diagram, at $r = r_s$, to I^0. Similarly, for the other regions. This diagram is called the *Penrose* (or *Carter-Penrose*) *diagram*.

6.5 Physics Near and Inside a Black Hole

The very strong gravitational field of the black hole leads to the extra term in Eq.(4.110) becoming significant. Let us investigate the consequences of the gravitational red-shift, deflection of light, time delay and tidal forces near (or inside) the black hole, by watching some (imaginary) observers falling freely into it.

First imagine an intrepid explorer who agrees to sacrifice his life investigating a black hole interior for us. At regular intervals he has to send signals to both of us as he falls straight into the hole. As he gets close to it we see him getting redder (due to the gravitational red-shift). His path and your line of sight to the black hole are in co-incidence so that he sends his signals to you opposite to his line of flight, see Fig. 6.9. Clearly the time intervals between the signals increases steadily. This effect is an example of what is generically called "time delay" in the vicinity of a gravitational source. It has been verified in solar system experiments by having signals sent from spacecraft on the other side of the Sun. The time delay is exactly in agreement with the predictions of Relativity.

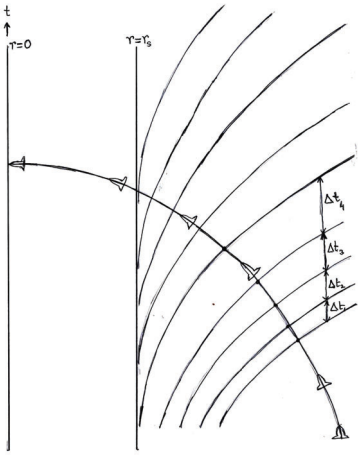


Figure 6.9: The intrepid observer in his rocket ship goes into a black hole and sends us regular signals informing us of his progress as he goes in. Due to the curving of null geodesics, though he sends them at regular intervals, we receive them at ever increasing intervals till they cease altogether.

I am not in the same line as our explorer's motion. As such he has to send his signals at some angle. Due to the gravitational deflection of light his signals to me get diverted and never reach me, see Fig. 6.10. Instead I start receiving the signals he is sending to his girl-friend. Due to the gravitational red-shift there is a loss of energy of the photons emitted by him and so we see him fade away from view as he hovers above the horizon and we never get to see him penetrate it.

Chagrined at our failure to explore the black hole interior from the safety of "infinity" we decide to sacrifice another explorer, but this time follow along behind her. To avoid being swallowed up by the black hole we resolve to accelerate

away from it after coming fairly close and keep in touch with our explorer as long as we can. We see her reddening, but less than before (because the relative red-shift is less) and see less of an increase in the time interval between the signals (for the same reason). She also seems to fade less. However, we now start feeling ourselves being stretched out a bit (due to the tidal force). The larger is the black hole the greater is its area and hence the less is its surface gravity. As we quickly accelerate away we see our observer rapidly redden and fade away from view *outside the horizon*, see Fig.6.11 (drawn in a Carter-Penrose diagram).

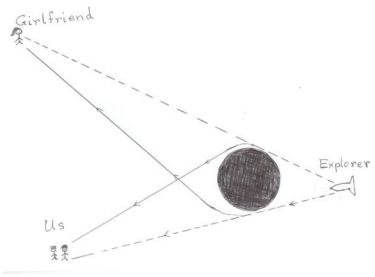


Figure 6.10: The signals sent to us from far away get to us. However, as he nears the hole, the signals beamed to us go astray and we get those beamed to his girlfriend.

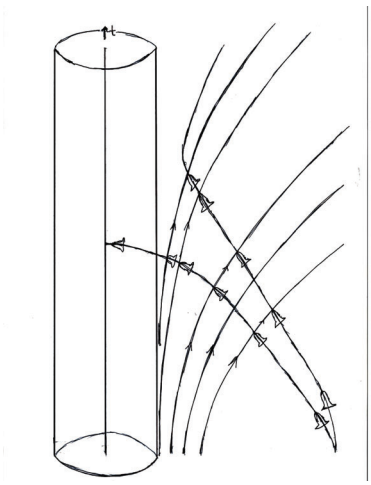


Figure 6.11: Another intrepid observer exploring the black hole interior is followed in by us. We continue to receive signals sent by her but with less of a time delay and red-shift than before till we accelerate away from the black hole and very rapidly lose sight of her due to both effects.

Desperate, by now, for a glimpse of " the promised land" inside the black hole, you persuade me to fall freely into the black hole while you follow me falling freely, see Fig. 6.12. Neither of us see anything spectacular in our journey but we do feel ourselves being increasingly torn apart (by tidal forces). I have been

sending you signals steadily. Then as you get to $3r_s/2$ you see the light coming out of my side window and feel that you have caught up with me. By Eq.(4.111) this is the light that entered a circular orbit. Of course, you had received my signal, informing you that I was at $3r_s/2$, earlier. The signal you received, telling you that I am approaching r_s, reached you because it was beamed straight back at you. After this, my signal informing you where I have reached is received by you barely before you get there. Like Achilles with the tortoise, you never quite overtake me. There is nor much of interest to see in "the promised land" — no intimation of impending doom. For all your cleverness in persuading me to go first, *you never see me hit the singularity — because we hit it simultaneously*!

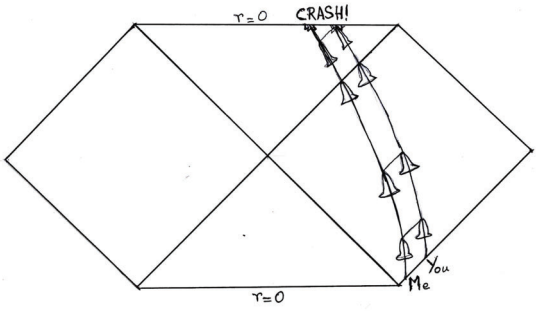


Figure 6.12: A Carter-Penrose representation of me falling freely into the black hole before you do, but you do not see me hit the singularity before you, as we both hit it simultaneously. In fact, you take less proper time to hit it than I do, as evidenced by the shorter path you traverse in the Carter-Penrose diagram. The reason is that I have come from "further away", in the sense of starting at an earlier time.

It is useful to look at some relationships between physical parameters of the black hole. The radius of the black hole is, of course, proportional to its mass, $r_s = 2Gm/c^2$, and so the "area" of the black hole as defined by an observer at infinity is

$$A = 16\pi G^2 m^2/c^4 \tag{6.41}$$

and its "volume" is

$$V = 32\pi G^3 m^3/3c^6 \ . \tag{6.42}$$

Hence the surface area increases as the *square* of the mass of the black hole and its volume as the *cube*. The corresponding "density" of the black hole is, then,

$$\rho = 3c^6/32\pi G^3 m^2 \ , \tag{6.43}$$

i.e. it decreases quadratically with m.

The maximum tidal acceleration of an object of length l, at the surface of the black hole, is

$$\mathcal{A}_T = 2Gml/r_s^3 = lc^6/4G^2m^2 \ , \tag{6.44}$$

which *decreases quadratically* with m. Similarly, the "surface gravity", defined in the sense of the ψN-force, is

$$\mathcal{G} = c^4/4Gm \ , \tag{6.45}$$

which *decreases* with increasing mass. For a solar mass black hole the tidal force on a 10 tonne, 100 metre spaceship would be 10^{23} Newton. However, for a galactic mass black hole it would only be about 10^{-5} Newtons! Our explorer would not be too badly mangled at the surface of a galactic mass black hole, therefore.

6.6 The Charged Black Hole

To be able to see the generic features of black holes we need to investigate other examples than that provided by the Schwarzschild exterior metric for an object with radius less than r_s. The simplest solution after the Schwarzschild is the Reissner-Nordström solution to the source-free Einstein-Maxwell equations, discussed in section 5.3, given by Eq.(5.46). We shall, therefore, study the singularities of that metric.

The Reissner-Nordsrom metric is singular when

$$e^{\nu(r)} = (r^2 - 2Gmr/c^2 + GQ^2/c^4)/r^2 \tag{6.46}$$

is either infinite or zero. This occurs when $r = 0$ or

$$\begin{aligned} r = r_\pm &= \frac{Gm}{c^2} \pm \sqrt{\frac{G^2m^2}{c^4} - \frac{GQ}{c^4}} \\ &= \frac{Gm}{c^2}\left[1 \pm \sqrt{1 - \frac{Q^2}{Gm^2}}\right] . \end{aligned} \tag{6.47}$$

Thus there is a singularity at $r = 0$ and two more if $0 < Q^2 < Gm^2$, one more if $Q^2 = Gm^2$ and no more if $Q^2 > Gm^2$ (as there would then be no real solution of Eq.(6.47)). In the case $Q = 0$ we, of course, revert to the Schwarzschild solution.

The red-shift for the Schwarzschild metric, given by Eq.(4.102), could be extended by the same arguments to the Reissner-Nordström metric, to give

$$\frac{\nu_B}{\nu_A} = \frac{r_B}{r_A}\sqrt{\frac{r_A^2 - 2Gmr_A/c + GQ^2/c^4}{r_B^2 - 2Gmr_B/c + GQ^2/c^4}} . \tag{6.48}$$

Corresponding to Eq.(4.8.3), if we take $r_A = R$ and $r_B = \infty$, we get

$$\begin{aligned} \frac{\delta\nu}{\nu} &\approx -\left(\frac{Gm}{c^2R} - \frac{GQ^2}{2c^4R^2}\right) \\ &= -\frac{Gm}{c^2R}\left(1 - \frac{Q^2}{2mc^2R}\right) . \end{aligned} \tag{6.49}$$

Clearly, $r = r_+$ is a red-shift horizon corresponding to $r = r_s$ for the Schwarzschild black hole. Also, since $c\dot{t} = e^{-\nu(r)}$ in this metric as well, there is an infinite time delay of signals originating from $r = r_+$. As with the Schwarzschild metric, this surface is also a trapped surface and a null surface. What remains to be checked is that $r = r_+$ is not an essential but a coordinate singularity.

To determine the nature of all the singularities we shall repeat the analysis of section 3 for the Reissner-Nordström metric. Using Eqs.(6.13) with $\lambda(r) =$

$-\nu(r)$, which also holds for this metric, we can re-write the non-zero independent Riemann tensor components is a more convenient form as

$$\left.\begin{aligned} R^{01}{}_{01} &= \tfrac{1}{2}e^{\nu}(\nu'' + \nu'^2) = \tfrac{1}{2}(e^{\nu})'' \ , \\ R^{02}{}_{02} &= R^{03}{}_{03} = -R^{12}{}_{12} = -R^{13}{}_{13} = \tfrac{1}{2r}(e^{\nu})' \ , \\ R^{23}{}_{23} &= -\tfrac{1}{r^2}(1 - e^{\nu}) \ . \end{aligned}\right\} \tag{6.50}$$

Replacing e^{ν} in Eqs.(6.50) by the expression for it given by Eq.(6.46) we obtain

$$\left.\begin{aligned} R^{01}{}_{01} &= -\tfrac{2Gm}{c^2r^3} + \tfrac{3GQ^2}{c^4r^4} = -\tfrac{2Gm}{c^2r^3}(1 - \tfrac{3Q^2}{2mc^2r}) \ , \\ R^{02}{}_{02} &= R^{03}{}_{03} = -R^{12}{}_{12} = -R^{13}{}_{13} = \tfrac{Gm}{c^2r^3}(1 - \tfrac{Q^2}{mc^2r}) \ , \\ R^{23}{}_{23} &= - - \tfrac{2Gm}{c^2r}(1 - \tfrac{Q^2}{2mc^2r}) \ . \end{aligned}\right\} \tag{6.51}$$

Thus $\mathcal{R}_1 = R = 0$ as before (since $T^{(em)} = 0$) and from Eq.(6.15)

$$\left.\begin{aligned} \mathcal{R}_2 &= \tfrac{48G^2m^2}{c^4r^6}(1 - \tfrac{2Q^2}{mc^2r} + \tfrac{7Q^4}{6m^2c^4r^2}) \ , \\ \mathcal{R}_3 &= -\tfrac{64G^3m^3}{c^6r^9}(1 - \tfrac{3Q^2}{mc^2r} + \tfrac{7Q^4}{2m^2c^4r^2} - \tfrac{7Q^6}{4m^3c^6r^3}) \ . \end{aligned}\right\} \tag{6.52}$$

It is obvious that both $\mathcal{R}_2$ and $\mathcal{R}_3$ remain finite at $r = r_{\pm}$ and become infinite at $r = 0$. In fact both tend to infinity more sharply as $r \to 0$ in this metric than in the Schwarzschild metric. Hence $r = r_{\pm}$ are coordinate singularities while $r = 0$ is an essential singularity.

6.7 Convenient Coordinates for the Charged Black Hole

As with the Schwarzschild black hole we would like to use coordinates that pass through the horizon at $r = r_{\pm}$ without becoming singular there. With our previous experience to guide us, we can bypass the equivalent of the Eddington-Finkelstein coordinates and proceed directly to the generalised Kruskal coordinates appropriate for the Riemann-Nordström metric. The difference between this black hole and the previous type is that this has two singularities and so a quadratic factor in the denominator has to be integrated to obtain r^*. As such it is not necessary that the previous choice of α and β in Eqs.(6.29) will be appropriate now.

Corresponding to Eq.(6.24) we have, here,

$$r^* = \int \frac{r^2 dr}{(r - r_+)(r - r_-)} = \int [1 + \frac{(r_+ + r_-)r - r_+r_-}{(r - r_+)(r - r_-)}]dr \ . \tag{6.53}$$

Using the method of integration by partial fractions yields

$$r^* = r + \frac{r_+^2}{r_+ - r_-} \ln\left|\frac{r - r_+}{k_1}\right| - \frac{r_-^2}{r_+ - r_-} \ln\left|\frac{r - r_-}{k_2}\right| \ , \tag{6.54}$$

where k_1 and k_2 are arbitrary constants of integration, which are obviously most conveniently chosen to be r_+ and r_- respectively. Using this r^* to define u and v, corresponding to Eq.(6.28), here ds^2 is given by

$$ds^2 = 2(1 - r_+/r)(1 - r_-/r)dudv - r^2 d\Omega^2 \ . \tag{6.55}$$

The definition of the Kruskal coordinates by Eq.(6.29) here yields

$$\left.\begin{aligned} U &= -\alpha e^{-ct/\beta} e^{-r/\beta} \left|\frac{r}{r_+} - 1\right|^{\frac{r_+^2}{\beta(r_+ - r_-)}} \left|\frac{r}{r_-} - 1\right|^{-\frac{r_-^2}{\beta(r_+ - r_-)}} , \\ V &= \alpha e^{ct/\beta} e^{-r/\beta} \left|\frac{r}{r_+} - 1\right|^{\frac{r_+^2}{\beta(r_+ - r_-)}} \left|\frac{r}{r_-} - 1\right|^{-\frac{r_-^2}{\beta(r_+ - r_-)}} , \end{aligned}\right\} \tag{6.56}$$

and hence

$$UV = -\alpha^2 e^{-2r/\beta} \left|\frac{r}{r_+} - 1\right|^{\frac{2r_+^2}{\beta(r_+ - r_-)}} \left|\frac{r}{r_-} - 1\right|^{-\frac{2r_-^2}{\beta(r_+ - r_-)}} . \tag{6.57}$$

Also, corresponding to Eq.(6.32) we have

$$dudv = \frac{\beta^2}{\alpha^2} e^{2r/\beta} \left|\frac{r}{r_+} - 1\right|^{-\frac{2r_+^2}{\beta(r_+ - r_-)}} \left|\frac{r}{r_-} - 1\right|^{\frac{2r_-^2}{\beta(r_+ - r_-)}} dUdV . \tag{6.58}$$

Thus, corresponding to Eq.(6.33) we get

$$\begin{aligned} ds^2 =& \frac{2\beta^2}{\alpha^2}(1 - \frac{r_+}{r})(1 - \frac{r_-}{r}) e^{2r/\beta} \left|\frac{r}{r_+} - 1\right|^{-\frac{2r_+^2}{\beta(r_+ - r_-)}} \\ & \left|\frac{r}{r_-} - 1\right|^{\frac{2r_-^2}{\beta(r_+ - r_-)}} dUdV - r^2 d\Omega^2 . \end{aligned} \tag{6.59}$$

To be able to eliminate the singularity at $r = r_\pm$ we need to be able to cancel the factors containing $(1 - r_+/r)$ and $(1 - r_-/r)$. It is clear that both cannot be cancelled simultaneously. As such we have to cancel one *or* the other with a given set of coordinates. There also appears to be a sign ambiguity. This comes from the definition of the coordinates on either side of the horizon. Thus, for coordinates regular at $r = r_+$ (and hence *not* at $r = r_-$) we take

$$\beta = 2r_+^2/(r_+ - r_-) , \tag{6.60}$$

and write $(1 - r_+/r)$ as $(r_+/r)(r/r_+ - 1)$. Since $r > r_-$, we get

$$ds^2 = \frac{2r_+^2}{r^2}(\frac{r}{r_-} - 1)^{\frac{r_+^2 + r_-^2}{r_+^2}} e^{-\frac{r(r_+ - r_-)}{r_+^2}} dU_1 dV_1 - r^2 d\Omega^2 , \tag{6.61}$$

having chosen (for convenience)

$$\alpha = 2r_+ r_-/(r_+ - r_-) . \tag{6.62}$$

Clearly, the metric is non-singular at $r = r_+$ and singular at $r = r_-$.

To avoid the singularity at $r = r_-$ we choose

$$\beta = -2r_-^2/(r_+ - r_-) , \tag{6.63}$$

and write $(1 - r_-/r)$ as $(r_-/r)(r/r_- - 1)$. Here $r < r_+$ and we get

$$ds^2 = \frac{2r_-^2}{r^2}(\frac{r}{r_+} - 1)^{\frac{r_+^2 + r_-^2}{r_-^2}} e^{-\frac{r(r_+ - r_-)}{r_-^2}} dU_2 dV_2 - r^2 d\Omega^2 . \tag{6.64}$$

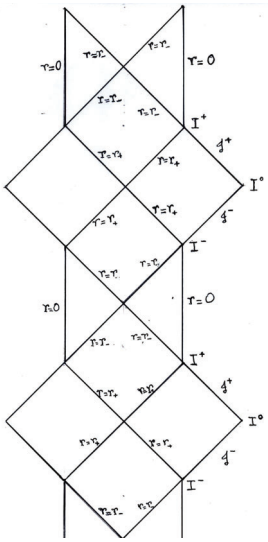


Figure 6.13: The Carter-Penrose diagram for the Reissner-Nordström black hole with $Q^2 < Gm^2$. Notice that the diagram has become non-compact going with infinitely many copies of the basic blocks going from the bottom to the top. Also notice that the singularity is *timelike* in this case.

The two sets of coordinates can be patched together in the region $r_- < r < r_+$, where they are both valid.

We can again construct the generalized Kruskal-Szekres coordinates for the Reissner-Nordström metric, in each patch, using Eq.(6.34). The only difference from the earlier expression for the metric, given by Eq.(6.35),will be due to the extra factor and the replacement of r_s in the exponential. The Kruskal picture is not so easy to draw here because of the two coordinate systems and the fact that one region stretches off to infinity, while the other hits a singularity at a finite coordinate position. Though there is no way to avoid the fact of two coordinate systems which must be patched together, we can make the diagram more tractable by using compactified coordinates, as before. Thus, if we take

$$U = \tan \bar{u} \quad , \quad V = \tan \bar{v} \ , \tag{6.65}$$

for $r > r_-$ we have $-\pi/2 < \bar{u}_1 < \pi/2$ and $-\pi/2 < \bar{v}_1 < \pi/2$. Also, for $r < r_+$ we have the other coordinates $-\pi/2 < \bar{u}_2, \bar{v}_2 < \pi/2$. In the region $r_+ < r < \infty$ we have $U_1, V_1 > 0$. At $r = r_+$, $t \to \infty$ or $t \to -\infty$ just as for the Schwarzschild black hole. For $r_- < r < r_+$ we have, in the other coordinates U_2, $V_2 > 0$, but in the previous coordinates either $U_1 < 0 < V_1$ or $V_1 < 0 < U_1$. We now have a further maximal extension piece U_1, $V_1 < 0$. At $r = r_+$ either $\bar{u}_1$ or $\bar{v}_1 = 0$ (or both $= 0$). At $r = r_-$ either $\bar{u}_2$ or $\bar{v}_2$ (or both) $= 0$. We can "sweep" the difference between the two coordinate systems "under the rug" to draw the Carter-Penrose diagram for the Reissner-Nordström metric, see Fig. 6.13. Notice that the essential singularity has become *timelike* instead of spacelike.

Though we have obtained a single diagram, we need two coordinate patches to cover the basic blocks of the diagram, which then repeat infinitely many times. As such, we need infinitely many coordinate patches. When I made this point

to Wheeler, with his typical turn of phrase he said, "You mean that diagram is a lie?". That point bears further pondering, as one's physical intuition says that one patch should be enough. Could it be that the region behind the inner horizon is a "mathematical mirage"?

From Eq.(6.47) it is clear that $r_+ + r_- = r_s$. As $Q \to 0$, $r_+ \to r_s$ and $r_- \to 0$. Thus, in the limit as the Reissner-Nordström black hole tends to become the Schwarzschild black hole, the outer horizon becomes the usual horizon while the inner horizon collapses on to the essential singularity. The non-compact Carter-Penrose diagram for the charged black hole must then collapse to the compact diagram for the uncharged black hole. The analysis above would not be valid in that case as we could not divide by r_-.

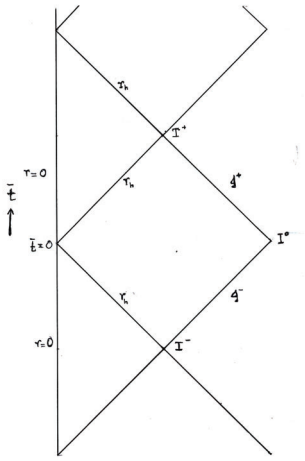


Figure 6.14: The Carter-Penrose diagram for the "extreme" Reissner-Nordström black hole, i.e. with $Q^2 = Gm^2$. While the left side of the diagram has "collapsed", the diagram is still non-compact from bottom to top and the singularity continues timelike.

There is another situation where the above analysis breaks down. That is when $Q^2 = Gm^2$ as there we have $r_+ = r_- := r_h = r_s/2$. This is called the extreme Reissner-Nordström metric. Here we get a double pole in the coordinate singularity in Eq.(6.46). Since there is only one root r^* has to be evaluated as in Eq.(6.24) with r_s replaced by r_+. Proceeding as in section 4 we find that the singularity cannot be removed by taking $\beta = r_h$, but we have to use diagonal coordinates developed by Brandon Carter [76]:

$$\psi = \tan^{-1} v + \cot^{-1} w \ , \quad \xi = \tan^{-1} v - \cot^{-1} w \ ,$$

where $v = ct + r^*$, $w = -ct + r^*$, and $dr^* = (1 - Gm/c^2 r)^2 dr$. In these coordinates the metric becomes

$$ds^2 = \frac{r_h^2 (r - Gm/c^2)^2}{4r^2} \sec^2 \frac{\psi + \xi}{2} \csc^2 \frac{\psi - \xi}{2} (d\psi^2 - d\xi^2) - r^2 d\Omega^2 \ . \qquad (6.66)$$

While the metric is nonsingular at $r = Gm/c^2$, it is not *manifestly* so, in that putting this value into the metric makes the first part of the coefficient

of $(d\psi^2 - d\xi^2)$ zero. However, at this value the second part of the coefficient tends to infinity. Thus it is an indeterminate form of the type $0 \times \infty$, and using L'Hospital's rule it is found that the coordinates are regular (but are "kinky" in a sense that will become clear later in this chapter). Again, the diagram is non-compact and there is a timelike singularity at $r = 0$ (see Fig. 6.14).

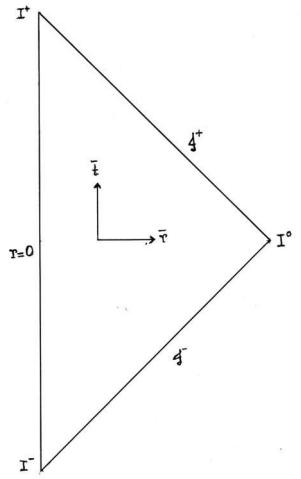


Figure 6.15: The Carter-Penrose diagram for the Reissner-Nordström "naked singularity", with $Q^2 > Gm^2$. This diagram has become compact but the singularity continues timelike. In fact, this diagram is very similar to the Minkowski space Carter-Penrose diagram, but with the coordinate singularity at $r = 0$ of the Minkowski converted to an essential "naked" singularity.

Yet another case is with $Q^2 > Gm^2$. In this case there is no coordinate singularity and so we can continue to use (t, r) coordinates. Here the Carter-Penrose diagram looks the same as that for Minkowski space but the coordinate singularity of Minkowski space at $r = 0$ becomes an essential timelike singularity, see Fig. 6.15. In this case the diagram is compact. When there is an essential singularity that is not "clothed" by an event horizon, it is called a *naked singularity*. (Such objects will be discussed later in this chapter.) It is of interest to note that a charged elementary particle (like an electron) has $Q^2 >> Gm^2$, and so it would be a naked Reissner-Nordström singularity, if one could use GR to describe it. Actually, the electron also has a spin, so the charged Kerr metric (discussed in the next section) would be required for it.

It is instructive to also consider the "inner horizon", $r = r_-$, and compare it with the "outer horizon", $r = r_+$, even though it cannot be seen from far away. The latter is a red-shift and time delay horizon, meaning that light emitted there is infinitely red-shifted and infinitely time delayed, as seen at infinity. In this sense we cannot talk of the inner horizon as a red- (or a blue-) shift horizon, or as a time-delay horizon, as light from it can never reach infinity in any case. However, we *can* consider light from infinity reaching the inner horizon. For this purpose we need to use the outer Kruskal-like coordinates. To see what it implies, re-write Eq.(6.61) using the generalized Kruskal-Szekres coordinates

$$cT_1 = (U_1 + V_1)/2 \quad , \quad R_1 = (V_1 - U_1)/2 \ , \tag{6.67}$$

whence we would get

$$dU_1 dV_1 = c^2 dT_1^2 - dR_1^2 \ . \tag{6.68}$$

Thus the coefficient of dT_1^2 becomes zero at $r = r_-$. Hence light coming from infinity is infinitely blue-shifted as it reaches the inner horizon! In this sense it is a *blue-shift horizon*. This may be compared with the Schwarzschild metric where light would get infinitely blue-shifted as it reached $r = 0$. On this count the inner horizon is more like the crushing singularity than a horizon, as any energy hitting it will become infinite. That infinite energy will have an infinite back-reaction. Richard Matzner picturesquely described it as the energy "shutting the door behind it" as it goes through the inner horizon. In the light of Wheeler's comment mentioned earlier, it may be that it finds "the door locked as it comes to it".

As with the outer horizon, the inner horizon *is* a null surface, but unlike that horizon it is *not* a trapped surface in that light entering $r < r_-$ from $r_- < r < r_+$ could escape back to $r > r_-$. This fact will be discussed further, in the context of the force of gravity in that region, in more detail in section 9. In some ways the inner horizon seems like a "reversed" event horizon. Because of this fact, it is also called an *anti-event horizon*.

6.8 The Kerr Black Hole

For the rotating gravitational source, represented by the Kerr metric, we again need to investigate the essential and coordinate singularities, the red- and blue-shift horizons, the null and the trapped surfaces. Since the metric, given by Eq.(5.50), has off-diagonal terms, simply putting $g_{00} = 0$ will *not* determine the singularities of this metric. To look for these singularities the first step is to consider its determinant

$$g = (g_{00} g_{33} - {g_{03}}^2) g_{11} g_{22} \ . \tag{6.69}$$

Though it is not directly obvious that the theta dependent parts of g_{00}, g_{03} and g_{33} "miraculously" cancel out, it turns out that we are left with only $-\rho^4 \sin^2\theta$, the Δ in the denominator of g_{11} having been "miraculously" cancelled out by the terms in the numerator of $(g_{00} g_{33} - {g_{03}}^2)$. The $\theta = 0, \pi$ singularities are the usual coordinate singularities for spherical coordinates. The cancellation of the Δs here is similar to the cancellation of $(1 - 2Gm/c^2 r + GQ^2/c^4 r^2)s$ in the Reissner-Nordström metric and leads to coordinate singularities at

$$\begin{aligned} r &= r_\pm = \frac{Gm}{c^2} \pm \sqrt{\frac{G^2 m^2}{c^4} - \frac{a^2}{c^2}} \\ &= \frac{Gm}{c^2}\left[1 \pm \sqrt{1 - \frac{a^2 c^2}{G^2 m^2}}\right] \ , \end{aligned} \tag{6.70}$$

again yielding event and anti-event horizons at these values. Corresponding to the $r = 0$ singularity there, we have to replace r^2 by $\rho^2 = r^2 + a^2 \sin^2\theta$. It is not easy to determine the nature of the singularity here as the invariants are quite complicated, and there more functionally independent ones due to loss of spherical symmetry. In general, of the 20 such components only 8 are zero and 12 survive. Bear in mind, also, that there are 12 non-zero, independent

Christoffel symbols of a "diagonal" type and 8 more which are twice as complicated, given by Eq.(5.53). After some very messy calculations, it is found that the singularities given by Eq.(6.70) are coordinate singularities, while there is an essential singularity at $\rho^2 = 0$. This is not zero at $r = 0$ because of the additional term. To make that zero we must also take $\theta = \pi/2$. This is called a *ring singularity*, because no singularity appears at $r = 0$ if it is approached along any $\theta \neq \pi/2$, but is seen if approached along $\theta = \pi/2$, see Fig. 6.16. The "line singularity" of the Schwarzschild metric is seen here as well.

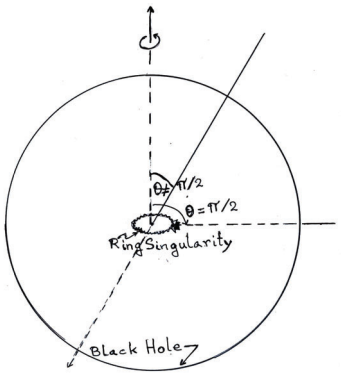


Figure 6.16: The ring singularity of the Kerr metric. Approaching $r = 0$ along $\theta \neq \pi/2$ no singularity is seen, but approaching along $\theta = \pi/2$ there is an essential singularity. This is depicted by a ring in the equatorial plane. However, it must be remembered that the ring is at $r = 0$ and not some other "radius". This counterintuitive description comes because r is a radial *parameter* and not the spherical radial coordinate.

It is possible to construct generalized Kruskal, or Kruskal-Szekres, coordinates here, as we did for the Reissner-Nordström metric, by getting rid of the zeroes of g_{11}. For this purpose we would only need to find r^* by integrating $r^2 dr/\Delta$. Here we can write $\Delta = (r - r_+)(r - r_-)$ with $r_\pm$ given by Eq.(6.70), and use exactly the same formulae as before, i.e. Eqs.(6.61) and (6.64). These can then be compactified to obtain the Penrose diagram given in Fig. 6.17. It needs to be pointed out that since $r = 0$ for $\theta \neq \pi/2$ is not an essential singularity, and r no longer retains the earlier significance of a "radial distance", there is nothing to stop us from extending the metric to negative values of r. The metric is asymptotically flat as $r \to -\infty$.

By constructing the Killing vectors and considering their magnitudes, we can determine many interesting physical properties of spacetimes. This will be discussed in more detail in the last chapter. Suffice it to say, here, that since a Killing vector gives the direction along which the metric tensor is Lie transported, it provides information about the symmetries of the geometry. There exists a timelike Killing vector for the Kerr metric, which is δ^μ_0 in Boyer-Lindquist coordinates. Thus its magnitude squared is g_{00}. This will change sign when $g_{00} = 0$. Thus a feature of the Kerr metric that appears from these considerations is that an observer locally at rest in this frame will appear to be rotating as viewed from infinity, being "dragged along" with the black hole's

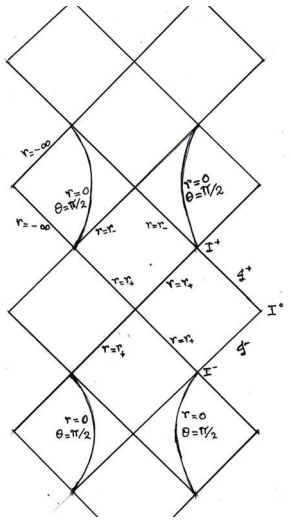


Figure 6.17: The Carter-Penrose diagram for the Kerr black hole with $a^2 < G^2m^2/c^2$. Notice that for this black hole there is a region with r going *negative*. This illustrates, once again, that this r is *not* the spherical radial coordinate but a different parameter denoted by the same symbol. The other two cases are not given here.

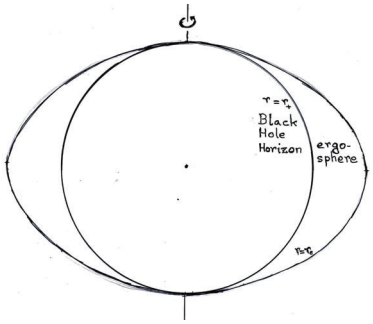


Figure 6.18: The ergosphere of the Kerr black hole, where observers locally at rest appear to be moving faster than light when viewed from infinity. This does not violate the SR requirement that nothing travels faster than light. The effect is due to the dragging of inertial frames.

rotation. For

$$r < r_e := (Gm/c^2)[1 + \sqrt{1 - a^2c^2\cos^2\theta/G^2m^2}] \ , \tag{6.71}$$

the particle locality at rest will appear to be moving superluminally. The region between r_+ and r_e (see Fig. 6.18) is called the *ergosphere*.

For the charged Kerr metric the same considerations apply as for the uncharged case. Here, however,

$$r_\pm = \frac{Gm}{c^2}[1 + \sqrt{1 - \frac{GQ^2 + a^2c^2}{G^2m^2}}], \tag{6.72}$$

and a corresponding modification of Eq.(6.71) replacing "a" by "$a\cos\theta$".

As we saw, the most general spherically symmetric, static (i.e. time independent) solution of the Einstein vacuum equations is the Schwarzschild metric. The most general spherically symmetric, static solution of the source-free Einstein-Maxwell equations is the Reissner-Nordström metric. The former result is known as *Birkhoff's theorem* and the latter as the *generalized Birkhoff theorem*. The most general axi-symmetric, stationary solution of the vacuum Einstein equations is the Kerr metric. This result is known as the *uniqueness theorem* (for axi-symmetric, stationary solutions). It is necessary, here, to distinguish between "static" and "stationary" as used in General Relativity. In both we require that there be time translational invariance according to same class of observers, i.e. that there exist timelike Killing vectors. For the former we further require that there be complete spacelike hypersurfaces orthogonal to the timelike Killing vectors, at all spacetime points. When this requirement is not met we call the spacetime *stationary*. In the Kerr metric the hypersurfaces get "wrapped around" by the rotating gravitational source and do not "match up" over a complete round through 2π.

The generalized uniqueness theorem for the charged Kerr metric as a stationary, axisymmetric solution of the source-free Einstein-Maxwell equations led to the belief that the charged Kerr metric was the most general stationary axisymmetric solution of all source-free coupled Einstein equations, i.e. that black holes could be identified by the three parameters: m, Q and a. This belief was known as the "no-hair conjecture" (that the black hole could be thought of as a bald head with no other parameters sticking out like hairs), and even as the "no-hair theorem". There are now various counter-examples to this conjecture known.

6.9 Naked Singularities and Cosmic Censorship

There is no zero of the function given by Eq.(6.46) if $Q^2 > Gm^2$. This condition could be re-written as

$$\frac{Q^2/r}{Gm^2/r} > 1 \ , \tag{6.73}$$

for arbitrary $r > 0$. In the limit $r \to 0$ the numerator can be interpreted as the electrostatic and the denominator as the gravitational "self-energy". For all charged fundamental particles (like electrons and quarks) or composites (like pions and protons) the above inequality holds. As such this condition is not unlikely to be met for small mass objects though it may be highly unlikely for astrophysical objects (which are more or less charge neutral). Again, for the Kerr metric $\Delta = 0$ has no real root, as is evident from Eq.(6.70), when

$$a^2c^2 > G^2m^2 \ , \tag{6.74}$$

or more generally for the charged Kerr metric, Eq.(6.72), when

$$GQ^2 + a^2c^2 > G^2m^2 \ . \tag{6.75}$$

Even for neutral particles, like the neutrino or the neutron, this inequality holds. Again, for astrophysical objects it may be less likely to occur but the possibility can not be totally ignored.

A problem arises if there is no event horizon blocking the essential singularity from the view of the "outside world". Since the curvature singularity is generally a region of infinite energy density, on account of the Einstein equations, it must lead to a break-down of physical laws there. (This observation holds regardless of whether the cosmological term is included or not.) If any thing can emerge from the singularity, there would be unpredictable effects, everywhere in the future of these singularities. Consequently physical laws would break-down not only on the singularities but in the entire spacetime in their future. Nor is it any use to appeal to the force of gravity to avoid anything emerging from the singularity. Put somewhat facetiously, what ever *goes down* can go up. An infinite energy density may, perhaps, be able to provide the infinite force required to counter the infinite gravitational attraction. More precisely, the geodesic equations are invariant under the reversal, as we get either the second derivative with respect to proper time or the square of the first derivative. Thus $\tau \to -\tau$ makes no difference. As such, there is no *a priori* reason, within GR, why time reversed black holes should not exist. (Such objects have been called *white holes*. It should be mentioned that there is no evidence that there are white holes in nature.) Further, in the Reissner-Nordström or Kerr black holes, timelike geodesics starting at $r < r_-$ can emerge to $r > r_-$. (In fact, in the Kerr metric we can get closed timelike curves in the region $r < 0$.)

Penrose proposed the possibility that Physics was protected against the corrupting influence of such *naked* singularities. Though I can not pretend to remember his exact words, I trust Penrose will allow me the liberty of putting my reconstruction of what he said as a quotation. "It is as if Nature acts as a *Cosmic Censor Board*", he said, "which does not allow naked singularities to appear, but requires that they be *clothed* by an event horizon". This conjecture of Penrose's is called the *cosmic censorship hypothesis*.

There is no proof, or other evidence, for cosmic censorship. One would be happy, therefore, to find an alternative way of "saving the Universe" from naked singularities. One possible idea comes from the investigation of the pseudo-Newtonian force. Let us pursue this idea further.

Remember that the ψN-force on a *neutral* test particle of mass μ, due to a charged gravitational source, given by Eq.(5.120), is

$$F^\mu = -\frac{Gm\mu}{r^2}(1 - \frac{Q^2/r}{mc^2})\delta^\mu_1 \ . \tag{6.76}$$

Since, for a naked singularity $Q^2 > Gm^2$, we can write it as $Q^2 = nGm^2$ with $n > 1$. Thus, the force becomes *negative*, i.e. *repulsive*, for

$$r < nGm/c^2 \ . \tag{6.77}$$

Therefore, despite the lack of a horizon there would be an increasingly greater repulsion of a neutral object by the singularity. If we tried bouncing some neutral probe off the singularity it would turn out to be impossible. The probe would be pushed away before it could hit the singularity. Expressed metaphorically in the context of the injunction for Muslim men, "the beauty of the naked singularity would be protected by the modesty of our vision". In the Greek mythological metaphor, it would be the beauty of the Gorgon that defies anyone gazing directly at it on pain of being converted to stone.

This mechanism applies equally to the, more general, charged Kerr metric with Eq.(6.75) holding. The force is, of course, not given by such a simple expression. It is given by differentiating Eq.(5.123), to yield

$$\left.\begin{aligned} F_r &= -\frac{(Gmr^2 - GQ^2 r/c^2 - Gma^2\cos^2\theta/c^2)}{(r^2 + a^2\cos^2\theta/c^2)^2}\mu \ , \\ F_\theta &= \frac{a^2(2Gmr - GQ^2/c^2)\sin 2\theta}{2(r^2 + a^2\cos^2\theta/c^2)}\mu \ . \end{aligned}\right\} \tag{6.78}$$

We shall, later, be discussing this force in a bit more detail. For the present, suffice it to say that the force can become *repulsive*, instead of attractive, for sufficiently small r. Of course, the distinction between "attractive" and "repulsive' is not so clear because the force is not radial. Nevertheless, we *can* define a rotated frame in which this force appears undoubtedly repulsive.

We have still not "saved the Universe" from the unpredictable influence of the singularity. However, the fact that there is an infinite time delay in any signal emerging from the singularity *does* "save the Universe" for a finite time. Since predictability by physical laws is required only for a finite time the Universe *is* saved. (It must be mentioned that null geodesics *can* hit and emerge from the singularity. However, it is reasonable to suppose that the effective mass of a real photon would cause it to behave like a massive particle. More precisely, one needs to include the photon as a perturbation on the background spacetime and calculate the modification of the geometry on account of it. This is called the "back-reaction" of the photon on the geometry. It would lead to the same effect *as if* the photon had a mass, and hence introduce the "repulsive behaviour of the singularity".) In his inimitable way, Wheeler had talked of "magic without magic". Modifying his phrase, this mechanism of obtaining "censorship without censorship" works for the two simplest singularities that could be considered, the Reissner-Nordström and the charged Kerr [40]. However, there is no clear answer to whether this "censorship without censorship" would hold more generally. There is a serious problem of extending the ψ-N, or $e\psi$-N, arguments to very general spacetimes without symmetries. Nor is it clear that such spacetimes would not contain naked singularities. As such the problem remains wide open and well worth investigating.

Let us go back to the structure of the force around a charged Kerr black hole (with Eq.(6.75) *not* holding). The first point to note is that the force is *not* a central force. This is not actually so surprising. Since the metric has a ring singularity at $r = 0$, $\theta = \pi/2$, it is reasonable that the force should have a polar component. This may be heuristically seen as the force at any $\theta \neq \pi/2$ pulling to (or repelling from) the ring singularity, visualised as being at $r = a\cos\theta$ on the equatorial plane for the same value of φ. Changing θ to $(\pi - \theta)$ and $\dot{\theta}$ to $-\dot{\theta}$ leaves the ψ-N force unaltered. However, for the same $|\dot{\theta}|$, $\mathbf{F}(\theta_0)$ and $\mathbf{F}(\pi - \theta_0)$ are not mirror images. This means that if, somehow, there is a physical mechanism producing the *same* magnitude of $\dot{\theta}$, there will be a net *gravitational* force acting along the axis of rotation. This fact may be crucially important to explain the observed fact that hardly any pulsars (only the two youngest ones) are observed *inside* their parent supernova remnants. The force exhibits some interesting features itself. (These observations may be found in more detail in [41].)

6.10 Foliating the Schwarzschild Spacetime

We now try to understand the geometry of the 4-dimensional Schwarzschild spacetime in terms of a 3+1 splitting. The problem, of course, is that this splitting is non-unique. At first sight the simplest splitting to implement seems to be by taking a constant time. However, the usual time coordinate cannot be used as it becomes infinite at the horizon. Thus we do not obtain a complete foliation of the Schwarzschild spacetime by hypersurfaces of constant t. However, we can obtain it by hypersurfaces of constant T, where T is the Kruskal-Szekres time. This foliation is depicted in Fig. 6.19. Unfortunately, the foliating hypersurfaces hit the singularity for $T^2 \geq 2$. Can we find foliations that avoid the singularity, so that each hypersurface remains non-singular and only some limiting hypersurfaces are singular?

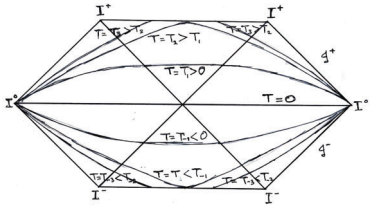


Figure 6.19: Using Kruskal-Szekres coordinates we can break the Schwarzschild spacetime completely into space and time, as shown in this Kruskal diagram. For $T^2 \geq 2$ the spacelike hypersurface hits the singularity at $r = 0$ (passing through the horizon $r = 2m$) but for $T^2 < 2$ it passes smoothly from one region into the other through the two sides of the horizon. Of particular interest are those foliations that avoid hitting the singularity.

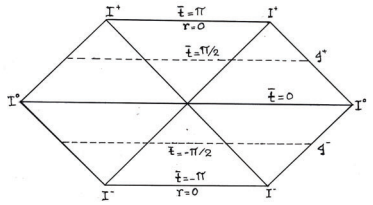


Figure 6.20: The foliation of the spacetime by constant $\bar{t}$ hypersurfaces, where $-\pi/2 < \bar{t} < \pi/2$. They start at $\mathcal{I}^+$ on the right, come in and pass smoothly through the horizon and then out to the maximally extended part of the spacetime. The hypersurface for $\bar{t} = 0$ starts at I^0 on the right, comes in and grazes by the horizon at $\bar{r} = 0$ and then goes out to I^0 on the left.

The simplest complete foliation, satisfying the above requirement, uses the compactified coordinates, $(\bar{t}, \bar{r})$, which are at the heart of the construction of the Penrose diagram. Here the foliating hypersurfaces are constant $\bar{t}$, where $-\pi/2 < \bar{t} < \pi/2$. The limiting hypersurfaces, $\bar{t} = \pm\pi/2$ are the future and past singularities $r = 0$, see Fig. 6.20. These hypersurfaces generically start at infinity at $\mathcal{I}^{\pm}$, come in and pass smoothly through the horizon and then out to the maximally extended part of the spacetime, which are parallel to the

$r = 0$ singularities. There is one hypersurface, $\bar{t} = 0$, which comes in from spacelike infinity, I^0, on the right, "grazes by" the horizon at $\bar{r} = 0$ and ends up at spacelike infinity in the maximal extension. In the previous case *all* of the hypersurfaces would have to end up at spacelike infinity, I^0, as they are spacelike hypersurfaces for some finite T. The $T = 0$ hypersurface is identical with the $\bar{t} = 0$ hypersurface. Neither of these sequences of hypersurfaces is easy to visualise in 3-dimensional terms.

Let us now proceed to physically motivated foliations instead of those that are easy to implement. One such foliation comes from ψ-N considerations and will, therefore, be called the ψ-N foliation. In this formalism the class of preferred observers is those falling freely radially inwards from infinity. Geodesics, given by Eqs.(4.85)-(4.88), must have $\dot{\theta} = \dot{\varphi} = 0$, here, and $\dot{t}$ given by Eq.(4.92) with $k = 1$. Using the metric in this case

$$c^2 d\tau^2 = c^2 e^{\nu} dt^2 - e^{-\nu} dr^2 \ , \tag{6.79}$$

we see that

$$dr/dt = \pm e^{\nu} \sqrt{1 - e^{\nu}}\, c \ . \tag{6.80}$$

For free-*fall* r must decrease and hence we take the negative sign. Inserting the value of e^{ν} $(= 1 - r_s/r)$ and integrating

$$t = t_0 - \frac{2r_s}{c} \left[\sqrt{\frac{r}{r_s}} (1 + \frac{r}{3r_s}) + \ln \sqrt{\left| \frac{\sqrt{r} - \sqrt{r_s}}{\sqrt{r} + \sqrt{r_s}} \right|} \right] \ . \tag{6.81}$$

For a given t_0 we have a particular observer specified. We need to specify the hypersurfaces orthogonal to these world lines. Here the unit vector along the world line is

$$n^{\mu} = (e^{-\nu}, -\sqrt{1 - e^{\nu}}, 0, 0) \ . \tag{6.82}$$

Let the vector tangent to the hypersurface, at the given r, taken along the direction $\dot{\theta} = \dot{\varphi} = 0$, be t^{μ}. Take it to be a unit vector with components t^0 and t^1. Then

$$\left.\begin{aligned} n^{\mu} t^{\nu} g_{\mu\nu} = 0 = t^0 + \sqrt{1 - e^{\nu}}\, e^{-\nu} t^1 \ , \\ t^{\mu} t^{\nu} g_{\mu\nu} = -1 = e^{\nu} (t^0)^2 - e^{-\nu} (t^1)^2 \ , \end{aligned}\right\} \tag{6.83}$$

as t^{μ} is spacelike. This gives us

$$t^{\mu} = \pm (e^{-\nu} \sqrt{1 - e^{\nu}}, -1, 0, 0) \ . \tag{6.84}$$

Now, the differential equation for the hypersurface comes from

$$\frac{dr}{dt} = \frac{ct^1}{t^0} = -\frac{c}{\sqrt{r_s}} \frac{r - r_s}{\sqrt{r}} \ . \tag{6.85}$$

Integrating this equation gives

$$t = t_1 - \frac{2r_s}{c} \left[\sqrt{\frac{r}{r_s}} + \ln \sqrt{\left| \frac{\sqrt{r} - \sqrt{r_s}}{\sqrt{r} + \sqrt{r_s}} \right|} \right] \ . \tag{6.86}$$

For different choices of t_1 we get different hypersurfaces. Here we can allow arbitrary θ and φ and for any given hypersurface (i.e. given t_1) for every r there

will be a unique t. Here t_1 is the ψ-N time. Though the hypersurfaces pass smoothly through the horizon, the coordinates used here break down. We can use Kruskal-Szekres coordinates to obtain an equation for the hypersurface that does not break down at $r = r_s$. This is left as an exercise for the reader. The Carter-Penrose diagram foliation is depicted in Fig. 6.21. The hypersurfaces are flat, i.e. have zero intrinsic curvature ($R^i{}_{jkl} = 0$) and, since the spacetime is asymptotically flat, the extrinsic curvature is zero at infinity and infinite at $r = 0$, see [42].

The above hypersurfaces were appropriate for describing Schwarzschild spacetime as analysed by freely falling observers. In some sense this is a local point of view. This is not the view taken for dealing with the Universe as a whole. Even though in both cases we are dealing with an entire spacelike hypersurface, in the latter we want the Universe to "look the same" to all observers (as we shall be discussing in more detail in the section after next). The appropriate description for such hypersurfaces may be a matter of opinion. One (very strong) point of view is that these should be hypersurfaces of constant mean extrinsic curvature. (We shall be discussing this requirement in some more detail in the last chapter.) For the Schwarzschild spacetime, using the compactified $(\bar{t}, \bar{r})$ coordinates the foliation is depicted by the Carter-Penrose diagram given in Fig. 6.22, see [43].

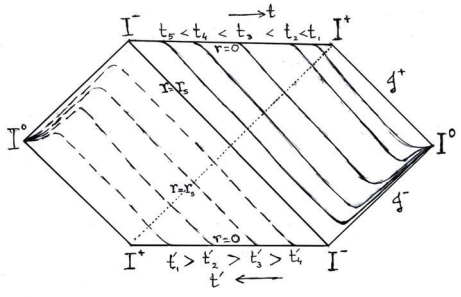


Figure 6.21: Foliation of the Schwarzschild geometry by ψ-N hypersurfaces. They start at I^0 and hit the singularity, $r = 0$. For lower values of t_i the hypersurfaces hit $r = 0$ closer to the left upper corner and so the direction of time along the singularity must be from left to right. This leads to the identification of the upper left corner as I^-. Notice that the foliation is only of the right side of the diagram. The left side would have to be done with time running in the reverse direction. This is obviously "the most classical description" of the black hole.

The foliating hypersurfaces (with one dimension suppressed) may be understood in terms of a sequence of sheets in "embedding diagrams". If the foliation hits I^0 and the hypersurfaces are asymptotically flat (i.e. flat at spacelike infinity), the sheets would be flat far away from $r = 0$ and curve as they approach it. However, if the foliating hypersurfaces do not hit $r = 0$, they must bend away from it without touching and go on into the maximal extension. This will, then, be represented by another asymptotically flat sheet, parallel to the previous one. If the surface passes through the "throat" of the hole (the region inside $r = r_s$) it will form a sort of "bridge" between the two asymptotically flat sheets. This is

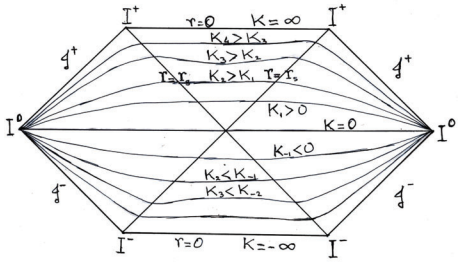


Figure 6.22: Foliation of the Schwarzschild geometry by hypersurfaces of constant mean extrinsic curvature, "K-slicing". The slices start at I^0 on the right and go straight through to the left corner. Since these are surfaces of simultaneity the left vertex is obviously I^0 as well. Here the hypersurfaces go up from bottom to top with increasing K. The hypersurface at $K = 0$ passes through the intersection of the two $r = r_s$ lines and gives a well-defined "present" time. In the lower limit they become the past singularity and past null infinity on the right. As such, the bottom left corner and the bottom left edge must be I^- and $\mathcal{I}^-$ as well. Similarly for the upper limit we get the future singularity, I^+ and $\mathcal{I}^+$. Here the whole diagram is foliated together. This is a truly global description.

called the Einstein-Rosen bridge, see Fig. 6.23. For the hypersurface $\bar{t} = 0$ (or $T = 0$) it will just graze by $r = r_s$. As we increase T (or $\bar{t}$) the hypersurface has a larger section passing through the throat of the hole and the "bridge" becomes longer and narrower. In the free-fall limit the foliating surface will consist of the two sheets joining at a point ($r = 0$), which will be singular. We say the throat has been "pinched off". Observers "thrown in" will reach the singularity faster and the sheets will be joined by a line.

In terms of the ψ-N foliation, the different hypersurfaces correspond to the observers starting at different (finite) times, see Fig. 6.21. In the limit of the observers starting infinitely far in the past, the hypersurface would be $\mathcal{S}_{\psi-N} = \mathcal{I}^- \bigcup I^- \bigcup \{r = r_s\}$. Since this is a hypersurface of simultaneity, the top left corner would also be I^-. Clearly the foliation never enters the maximal extension. This is reasonable, as all freely falling observers hit the singularity. In fact, if they are "thrown in", they will hit it "sooner", i.e. closer to I^- on the top left side. On the other hand, if they fire retro-rockets to slow down their fall, they will hit it "later", closer to I^+. The limiting hypersurface for times infinitely far in the future is simply $\mathcal{I}^+$. The left part of the diagram is got by rotating this diagram and so the arrow of time reverses.

A totally different interpretation appears if we consider foliation by spacelike hypersurfaces of constant mean extrinsic curvature, see Fig. 6.22. In this case hypersurfaces go into the maximal extension and reach I^0 on the left side. To see what this would mean in terms of observers, we make the following construction. At each point on the hypersurface take a timelike vector orthogonal to the hypersurface and do the same for the "next" hypersurface, in that it should be for a very slightly larger value of K. The sequence of timelike vectors at "the same place", will generate a world-line that remains orthogonal to the sequence of foliating hypersurfaces. Now take the collection of all such world lines. They will not be geodesics but will correspond to a class of accelerated observers.

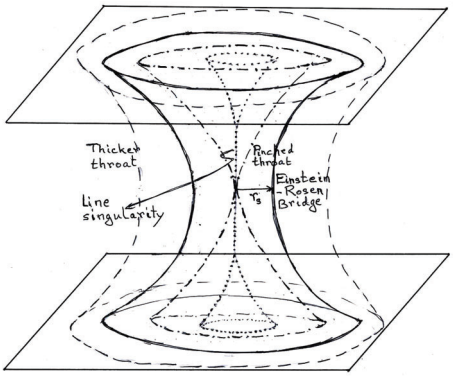


Figure 6.23: The Einstein-Rosen bridge description of the black hole depicted by embedding diagrams. The hard line throat represents the picture as seen by fixed observers. For observers accelerating away from the black hole, the throat becomes thicker, as depicted by the dashed throats. Contrariwise, for observers falling, but with retarded motion, the throat appears thinner. In the limit of free-fall, the throat "pinches off" to a point, as shown with the dot-dash line. For observers accelerating towards the black hole, the throat "pinches off" earlier and the two sheets are connected by the line singularity, as depicted by the dotted line. This may be visualised as a rubber sheet with a ball placed on it. The denser the ball, the deeper the dip. In the limit, it becomes like a needle piercing the sheet.

Only the limiting hypersurfaces for $K = \pm\infty$, $\mathcal{S}_K = \mathcal{I}^{\pm} \bigcup I^{\pm} \bigcup I^{\pm}\mathcal{I}^{\pm}$, hit the singularity $r = 0$ in the future or past. In fact, they incorporate the whole of the singularity. This implies that the left lower and upper corners correspond to the right lower and upper corners, and the left lower and upper null infinities correspond to the right lower and upper null infinities. This is opposite to the ψ-N foliation interpretation. It was big news when Stephen Hawking lost a bet about the "information loss paradox" to Leonard Susskind. This difference plays a critical role in Susskind's argument as will be discussed later.

Wheeler visualized the Einstein-Rosen bridge as connecting two opposite sides of an apple, the hole having been made by a worm eating through the apple, see Fig. 6.24a. While, in a local sense, the two sheets may be regarded as "maximal extensions" of each other, for the apple as a whole the sheets are connected along the surface.He called it a "wormhole" for obvious reasons. A topologically equivalent picture is depicted in Fig. 6.24b. Clearly, no time passes in traversing a wormhole (if it exists) and so the space is multiply connected; things going through the wormhole take no time, but those going along the usual path (the surface of the apple), being limited by light speed, would take time for the journey. Interest in wormholes may owe much to their "SF nature".

6.11 Black Hole Thermodynamics

One of the most exciting and rapidly developing branches of the study of black holes is what is called "black hole dynamics" or "thermodynamics". To be able to discuss it meaningfully I must refer to concepts of thermal physics. As those with

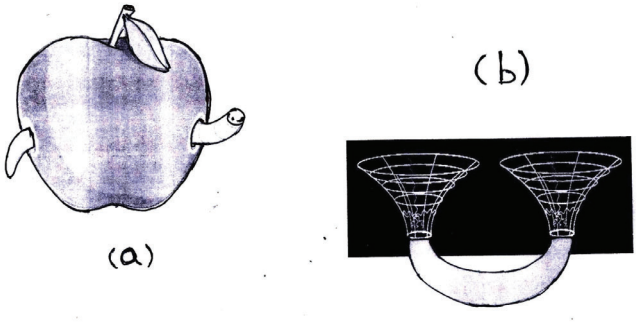


Figure 6.24: (a) A worm eating through an apple, producing a wormhole. (b) The apple has been topologically cut in half, so that the two ends of the wormhole lie on the same surface and the wormhole stretches out from one part of the surface to the other. Now see that as an embedding diagram with the maximal extension being part of the same Universe.

a Physics background would be familiar with it, but those from a Mathematics background may be lacking in it, it is necessary for me to provide something of a background. I cannot, here, do justice to the subject, which includes classical thermodynamics, the kinetic theory of gases, statistical mechanics and quantum statistics. The subject is covered in numerous very good books (see for example [44]). Without attempting to provide a proper grounding of thermal physics, or even a reasonably complete review of the subject, I will provide a very brief review of those parts of it that are essential for my purposes and leave the reader to fill in the rest of the background for themselves.

Those from a mathematical background normally think of the law of conservation of energy as being peculiar to Mechanics, because they have studied the inter-conversion of kinetic and potential energy there. However, it was formulated in 1850 for Thermodynamics, which deals with the inter-conversion of *all* forms of energy. Thermodynamics makes the fundamental idealization that one can totally insulate systems from their surroundings (called "isolated systems"), which can then be allowed a controlled interaction with the environs. It started with the innocuous looking statement that for an isolated system, the work done on the system would equal the heat gained by it. Clearly this statement was built for understanding the working of heat engines. Thermodynamics later evolved the concept of a total "internal energy" and the law was modified to the statement that the increase of internal energy of an isolated system equals the difference between the heat provided to it and the work done by it.

The second law of thermodynamics says that in any inter-conversion of energy, the overall change in heat energy of an isolated system is non-negative. If we use a temperature difference to work a heat engine, the amount of work done by it will not be greater than the energy used, but it can be less. This is often paraphrased as "There is no such thing as a free lunch", i.e. you have to pay in other ways for what is given as "free". Another way to see it is by analogy with changing currencies. Go to a money-changer and you have to pay a fee. Privately exchange with a friend and it will be break-even (or you will not remain friends). Processes in which it is equal are called *reversible* and the

others (wait for it) *irreversible*. In statistical terms the extent of *disorder* of the system, measured by an unmeasurable physical quantity called *entropy* (and generally denoted by S), must be non-decreasing in any process. This is one of the most fundamental and profound laws of nature. It is one of the very few laws of physics that is *not* invariant under time reversal. As such, it has been conjectured that it is at the root of *all* laws that are not time reversal invariant. In particular, it may be that it is at the root of the asymmetry in time that is one of the key features of the Universe as we see it. This asymmetry, or "directedness", is often called the "arrow of time".

In terms of Statistical Mechanics, entropy is the degree of disorder of a system. Since order carries information; the more order there is the more information is carried by the order of the system. How does one define the "order" or "disorder" of a system? To give consistent results it is taken to be proportional to the natural logarithm of the number of possible states in the system. What is meant by "the number of states"? Imagine a Rubik's cube, which has six faces of different colours, each face consisting of one central piece, four edge pieces and four corner pieces. The six central pieces are "fixed points", in that all rotations of the faces leave those unchanged. Each edge has two colours on the two sides of the edge; and each corner has three. There being 12 edge pieces there are 12! possible permutations of them. Similarly for the 12 corner pieces. Further, the edges have 2 orientations while the corners have 3. (One might have said 3!, but that would ignore the fact that we can only get cyclic permutations and not anti-cyclic permutations.) Thus all permutations would give 1.38×10^{18}. However there the face chosen down can be one of 6 and the choice of which face is directed in front is one in 4. Thus one must divide by 24. The next number of possible arrangements is 5.74×10^{16}. For our purpose, this is the number of states and so the entropy is proportional to 38.59. For physical purposes, "states" means elements of phase space, i.e. blocks of momentum space times blocks of physical space, $\Delta V_{\mathbf{p}} \Delta V_{\mathbf{x}}$. Further, the entropy times the temperature give the thermal energy of the system. Thus the entropy must have units of the inverse of temperature. Hence there must be a dimensional constant involved in the definition of entropy with units of energy per unit temperature. This is Boltzmann's constant. Hence we get the statistical definition of entropy as

$$S = -k \ln N \ , \tag{6.87}$$

where N is the number of states and k is Boltzmann's constant. Thus, in our example, $S = 38.59k$.

The third law of thermodynamics says that the absolute zero of temperature can never be achieved by a finite number of isothermal (constant temperature) and adiabatic (constant heat) cooling steps. It will be remembered that in classical thermodynamics absolute zero is the temperature at which an ideal gas would have zero volume. From the kinetic theory of gases and Boltzmann's statistical mechanics, the absolute temperature, T, is related to the energy of random motion of the particles in the system by the formula

$$E = \frac{3}{2} nkT \ , \tag{6.88}$$

where n is the number of particles in the system, which has a value 1.38×10^{-16}ergs/°K (i.e. the temperature being measured in degrees Kelvin, in which

the freezing point of the water is 273.16° and the boiling point 373.16°). The "random motion" is defined in the frame in which the average of velocity of all the particles is zero.

After these laws were stated it was considered necessary to state a prior law which, in effect, defined temperature. It was called the *zeroth law* so as to be *prior* to the other three without disturbing their numbering (much as is done in Relativity). It states that two objects in thermal equilibrium with a third object are in thermal equilibrium with each other. Thus we can define a parameter, temperature, such that heat flows from a higher temperature and so there is no heat flow when two systems are at the same temperature. The set of temperatures is then ordered. The *scale* for the definition of temperature comes from the mode of measurement. One often assumes that all that is needed is the zero of time and one other temperature (like the freezing or boiling point of water). If one thinks about it, that is not really true. It depends on the procedure for measurement of temperature. If, on conversion to another method, the two do not match, one would define one of them to be temperature-dependent.

Another law, called *Stefan's law*, deals with radiating rather than isolated systems. The darker a body the more efficient it is at absorbing or emitting radiation. The idealized limit of a perfect radiator is called a *black body*. Stefan's law states that the total energy density radiated by a black body is proportional to the fourth power of the temperature,

$$\rho = \sigma T^4 \ , \tag{6.89}$$

where $\sigma \approx 7.6 \times 10^{-15}\mathrm{erg/cm^3/^\circ K^4}$, is the Stefan-Boltzmann constant. This was stated as a purely empirical, and somewhat crude, result. There was no attempt at analysing the amount of energy involved for different frequencies. After all, if the amount depend on the frequency, what was the guarantee that the above formula would hold?

An attempt at spectral analysis of this energy radiation led to confusing results. For short wave-lengths (high frequencies) Wien found that the radiant energy, $E \propto \lambda^{-2} \propto \nu^2$, where λ is the wave-length and ν the frequency of the black body radiation being observed. However, for low frequencies Rayleigh and Jeans found that the radiant energy density, $\rho \propto \nu^3 e^{-\nu/\tau}$, where τ is proportional to its absolute temperature. In 1900 Max Planck deduced a formula for the energy density of radiation if it was emitted and absorbed by matter in discrete quanta of $h\nu$, where $h \approx 1.9\times10^{-26}\mathrm{gcm^2/sec}$ is called Planck's constant. This energy density distribution, which Planck verified experimentally, known as the *Planck spectrum*, is

$$\rho(\nu, T) = \frac{8\pi c^3 h\nu^3}{e^{h\nu/kT} - 1} \ . \tag{6.90}$$

On integration over all ν this gives Stefan's law. In the long wavelength limit it gives Wien's law and in the high frequency limit the Rayleigh-Jeans law. It clearly provided a more fundamental understanding of radiation from black bodies and is generally regarded as the birth of the quantum theory. It is ironic, however, that Planck regarded the duality relation, $E = h\nu$, as only referring to the emission and absorption of energy by matter, and not as a property of the radiation itself. In fact, when he nominated Einstein for the Nobel Prize, he

mentioned Einstein's explanation of the photoelectric effect as "his only aberration". The idea of discrete quanta of electromagnetic energy (nowadays called *photons*) comes solely from Einstein, and later Louis De Broglie.

The connection between black holes and thermodynamics started with an argument by Penrose about violating the second law by using black holes, that I heard in 1968 [45]. (By then he had already talked about it during his visit to America.) He started by considering non-usable objects, to extract their rest energy in usable form. The object could be lowered from a safe distance into the black hole. While going down it would wind up a spring. At the surface of the hole it would be released and the freed spring could then do useful work. The natural next step, as he soon realized, was to fill up a box with non-usable thermal radiation and lower the box to the black hole, as before, then open it near the surface, to let the thermal energy fall into the black hole. Due to the equivalence of mass and energy this would allow the relaxed spring to do useful work. We would, in fact, not only have "got something for nothing" but would have got rid of the thermal pollution far from the hole! We would not only have had a free lunch", but have got paid for eating it. We get extra energy *and* our heat engines work more efficiently. This is the connection between black holes and thermodynamics.

The only way to save the second law was to associate an entropy with the black hole. If its entropy increases more than the entropy of its surroundings decreases, the second law is saved. Here Penrose ran into a problem. What would be the measure of the entropy of the black hole? It would have to be given by an invariant expression and not be coordinate dependent. The Ricci scalar could not be used, as it is zero for the black hole. Of course, the gravitational field is actually described by the *Weyl* tensor. Since that tensor is traceless by definition, one has an impasse — unless one uses a quadratic expression instead of a linear one. This was as far as he got.

John Wheeler, having heard Penrose's talk, gave the problem of working out the entropy of the black hole to his student, Jacob Bekenstein, who took the identification of entropy with information to calculate the change of the entropy of a black hole when one unit of information is put into it. The problem of what constitutes one unit of information was beautifully resolved by Bekenstein. He took the simplest possible physical object, a single photon, as the unit of information. Now he took the requirement that the wavelength of the photon be less than the diameter of the black hole, so that it "fits into" the black hole. Thus $\lambda \leq r = 2Gm/c^2$ and, by the de Broglie duality relation, $\lambda = h/p = h/\delta mc$, where δm is the effective mass of the photon. Thus, $\delta m \geq ch/2Gm$. Clearly the ratio of the total entropy to a unit of entropy, k, must be proportional to the ratio of the total mass to the increase of mass, δm. Replacing the inequality by an equality, we get the entropy, $S = 2kGm^2/ch$. You might have noted that the argument is a little bit fuzzy in the precise numerical factors involved. Should we have taken the radius rather than the diameter? Should it really be h and not $\hbar = h/2\pi$? Thus we can really only say that $S \sim km^2/ch \sim kA/ch$, where A is the area of the black hole. It is only with the work of Hawking that the numerical factor crystallizes.

Penrose , with his student Roger Floyd, went on to construct a mechanism to extract energy from rotating black holes. The essence of the Floyd-Penrose process [46] is to break a compound object entering the ergosphere of a Kerr black hole with an angular momentum in the same direction as the rotating

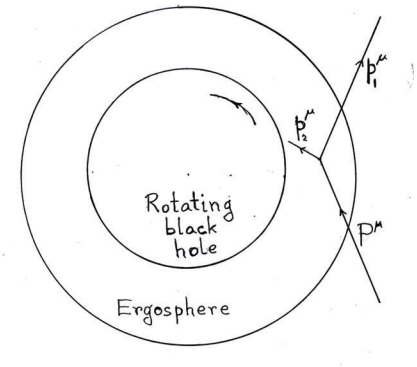


Figure 6.25: The Floyd-Penrose process. An object enters the ergosphere of a Kerr black hole, with momentum 4-vector P^μ and breaks into two parts, one of which goes out to infinity with a momentum p_1^μ and the other falls back into the black hole with a momentum 4-vector, p_2^μ. While both are timelike vectors, and so locally have positive energy, as seen from infinity the second one seems to have a negative zero-component, i.e. $p_2^0 < 0$ as seen from infinity. Hence $p_1^0 > P^0$ as seen from infinity, and we have extracted energy from the black hole.

black hole, in such a way that one part falls into the hole and the other goes away to infinity, see Fig. 25. Since the locally-rest particles appear to travel faster than light as viewed from infinity, the mass-energy would appear to be negative to that observer. Applying energy conservation, as seen from infinity the outgoing particle appears to have more energy than the original object did — even counting the rest energy in. *As such, the rotating black hole has lost energy in all!.* At the time when Penrose presented this argument it seemed unbelievable that energy could ever be extracted from a black hole. However, Remo Ruffini and his student, Demetrios Christodoulou, demonstrated that the charged Kerr black hole has an irreducible mass, and extra energy of a rotational or electromagnetic nature can be stored in its ergosphere — which lies *outside* the black hole proper [47]. As such, it only loses the mass stored as energy *outside* the black hole.

Hawking noticed that the black hole area in any process involving collision of black holes (or matter being absorbed into black holes) always increases [49]. The essence of Hawking's argument is explained in Fig. 6.26. Clearly, it implies that $(m_1 + m_2)^2 \geq M^2 \geq m_a^2 + m_2^2$. For a charged Kerr black hole the area is given by

$$\begin{aligned} A &= 4\pi[r_+^2 + a^2/c^2] \\ &= 4\pi[(Gm/c^2 + \sqrt{G^2m^2/c^4 - GQ^2/c^4 - a^2/c^2})^2 + a^2/c^2] \ . \end{aligned} \quad (6.91)$$

The *irreducible mass* is related to the area by

$$m_{ir} = (A/16\pi)^{1/2} \ , \quad (6.92)$$

and to the total mass by

$$m^2 = (m_{ir} + Q^2/Gm_{ir})^2 + L^2c^2/4G^2m_{ir}^2 \ , \quad (6.93)$$

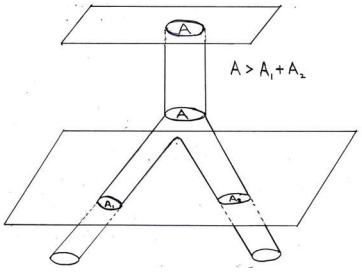


Figure 6.26: The Hawking area theorem, according to which the the area of two merged black holes, A must be greater than the sum of the areas of the two black holes, A_1 and A_2, that merged. Clearly this implies that $M^2 \geq m_1^2 + m_2^2$. Obviously, $m_1 + m_2 \geq M$, as some of the energy would be radiated in the process of merging. Hence $(m_1 + m_2)^2 \geq M^2 \geq m_a^2 + m_2^2$.

where L is the total angular momentum of the black hole, i.e. ma. Even in the Floyd-Penrose process the irreducible mass and surface area of the black hole increase.

The first law of black hole thermodynamics, like the first law of ordinary thermodynamics, is a statement of energy conservation, that the total energy (including rest energy) going into a black hole equals the increase in the rest energy of the black hole plus the energy radiated away from the in-falling object. Correctly speaking energy conservation holds only if there is time translational invariance. If this holds the spacetime is unchanging and nothing can happen in it. The difference from usual thermodynamics arises because classically we can conceive of an "isolated system" as one with adequate thermal insulation. *There can never be gravitational insulation* and hence we can not really think of the equivalent of an isolated system in the context of black holes. Nevertheless, we *can* treat a black hole as an approximately isolated system.

The second law of black hole thermodynamics is essentially contained in Hawking's area theorem along with Bekenstein's identification of the entropy of a black hole with its area, as the sum of the areas is a non-decreasing function of time, in very close analogy with the second law of ordinary thermodynamics, that entropy is a non-decreasing function of time.

The area of the black hole is an *extensive* variable. The natural choice of the corresponding *intensive* variable is the surface gravity of the black hole. Notice that the surface gravity of an extreme Reissner-Nordström black hole can even be zero but cannot be negative. As such one could identify the surface gravity with a "temperature" of the black hole. This was proposed by Bekenstein for consistency of black hole thermodynamics internally and its correspondence with ordinary thermodynamics, but with no further justification for the identification.

Fulling tried quantizing a scalar field in an accelerated frame in Minkowski space [50]. Defining the creation and annihilation operators in this frame and constructing the number operator from them, he calculated its expectation value in a Minkowski vacuum. The result was that the vacuum was *not empty* — the expectation value was non-zero! Worse still, it came out fractional in general. He noted this as an "ambiguity in quantizing in a non-flat space". (Correctly speaking the *spacetime* is flat but the *space* does not look it from an accelerated

frame.) The problem with the procedure is that most of the Minkowski space is inaccessible in the accelerated frame and one can only see a part (as explained in the previous chapter), see Fig. 5.4.

Hawking tried the same calculation in the Schwarzschild spacetime, using the frame of the "observer at infinity" [51]. He again found a non-zero vacuum expectation value. The calculation was done for a massless scalar field with a given frequency only. When Hawking tried looking at all modes he found that the vacuum expectation value of the energy at infinity came out with a thermal (Planck) spectrum! The temperature of the black hole turns out to be

$$T = \frac{\hbar c^3}{8\pi k G m} = 6.169 \times 10^{-8} \frac{M_\odot}{m} \ ^o\mathrm{K} \ , \tag{6.94}$$

where $M_\odot$ is the solar-mass (Sun's mass) $\sim 2 \times 10^{30}$ kg. More generally, the temperature turns out to be the surface gravity of the black hole, exactly as required. (It has since been found that for linearly accelerated observers there is also a Planck spectrum.) Hawking radiation can be understood as the quantum analogue of the Floyd-Penrose process, whereby the vacuum is polarized by the black hole similar to the electromagnetic vacuum polarization by parallel conducting plates giving the Casimir effect and quantum tunnelling through a potential barrier. As I am not assuming a background of Quantum Mechanics proper, leave alone Quantum Field Theory, I will not discuss these matters in any detail.

Since the black hole would be radiating away energy it must be losing mass. Consequently, after some time it would have a lower mass and a correspondingly smaller radius. Thus the surface gravity, and hence temperature, of the black hole would rise. The faster it would rise the more it would radiate. Hence we expect the black hole to evaporate away totally! For a "solar-mass mass black hole" (taken to be $\sim 10 M_\odot$ as black holes formed by gravitational collapse must have a mass $\geq 3.2 M_\odot$) the surface temperature would be $T \sim 10^{-8}$ ^{o}K, much less than the pervasive cosmic microwave background radiation (CMBR) temperature of the Universe. (The CMBR will be discussed in the next chapter and actual black holes in Chapter 8). As such it would go on *gaining* mass instead of losing it. One might say its "cosmic background radiation diet" would be too rich for it to lose mass by the slow "exercise" of Hawking radiation. Sufficiently low mass black holes, $\sim 4.5 \times 10^{22}$kg, that may have been produced in the early Universe, would be currently in equilibrium with the cosmic background. As the Universe cools such a black hole would start radiating and finally evaporate in a γ-ray burst. Black holes of less mass than this would be evaporating now. These postulated black holes are called "mini-black holes", or "primordial black holes".

For charged Kerr black holes the surface temperature of the black hole can *decrease* by radiating away the opposite charge to the black hole. Thus the hole would get charged up and its surface gravity go down. This could be achieved by putting the black hole inside an electric field which has the same charge as the black hole (as a thought experiment only, of course). The third law of black hole thermodynamics, like the third law of ordinary thermodynamics, says that zero temperature can not be obtained in a finite time.

In 1976 Hawking published another paper that led to intense debate at the foundations of Relativity *and* Quantum Mechanics [52]. He argued that if a pure quantum state absorbed by the black hole, it would be radiated as a mixed state

and the information contained in the pure state would be lost, thus violating Liouville's theorem. Since time evolution in Quantum Mechanics is given by unitary operators, except when a measurement is made, no information could be lost according to it. This paradoxical result became known as the "Information Loss Paradox". In effect this was a revival of Einstein's attempts to show that "the Quantum Theory is incomplete". There had been long-standing debates between Neils Bohr and Einstein on this matter, culminating in a paper by Einstein, Boris Podolsky and Nathan Rosen (EPR), attempting to provide a paradox if one accepts the completeness of Quantum Theory. The EPR argument was countered by Bohr and others. Finally, a test was made and it gave the remarkable result that *both Bohr and Einstein were wrong*! Now others took on the role of Bohr and Hawking placed himself in the position of Einstein. There was a bet placed in 1997 by Hawking (joined by Kip Thorne) and John Preskill, that Hawking's argument would be proved valid. In 2004 he admitted that he had lost the bet and this became very big news. The matter remains a subject of debate. An excellent exposition is given in [53]. (I have argued that the foliation of the spacetime may have a direct bearing on this matter [45].)

6.12 Exercises

1. For a Reissner-Nordström black hole of $10M_\odot$, what would be the charge (in Coulombs) for r_- to be 10% of r_+? What electromagnetic effects would one expect to see with such a black hole?

2. For the metric $ds^2 = f^2(u)dt^2 - f^{-2}(u)du^2 - u^2(dv^2 + \sin^2 v\ dw^2)$, where $f^2(u) = (1 - u/a)(1 - u/b)(1 - u/c)$ ($a < b < c$) and the speed of light is absorbed into the definition of time (or equivalently, units are taken with it equal to one), determine the singularity structure of the metric, i.e. where there are singularities and the nature of the singularities. Further, determine the physical features of the black hole, such as the red-shift, the time-delay, null or trapped surfaces. Hence determine if it is a black hole spacetime.

3. Take $a < b = c$ above and discuss the singularity structure. Similarly, take $a = b < c$ and repeat. Finally take the extreme case $a = b = c$.

4. Is this a vacuum solution of Einstein's field equations? If not, is it a solution of the equations with cosmological constant? If neither, can it be made a solution of either by some special choice of a, b, c? [Note that we can take a point mass at $u = 0$ without putting in any new field.]

5. If "none of the above", one has to take it to be the solution of the Einstein field equations coupled with some other field. Try to construct a field (or fields) that could provide the metric.

6. Construct the Kruskal, Kruskal-Szekres and compactified Kruskal-Szekres coordinates for these spacetimes. Hence construct the Carter-Penrose diagram for them.

7. For a point gravitational source of mass m, solve the Einstein field equations with cosmological constant if $T^{\mu\nu} = 0$. This is the Einstein-de Sitter metric. Again analyse the singularity structure and construct the various coordinate systems to make the Carter-Penrose diagram.

8. Repeat the above with a point mass, m, if it has a charge Q. How is the singularity structure modified?

Chapter 7

Relativistic Cosmology

When GR was first formulated it was so far in advance of its times that there was no question of its being forced on us by experimental or observational inconsistencies. True, it *had* explained the observed perihelion shift of Mercury's orbit. However, all other predictions seemed beyond the reach of contemporary technology. A theory that only explained one new phenomenon, or observation, could be regarded as tailor-made for just that purpose but having no further claim to validity. This applies particularly to a theory that diverged so radically from previous world-views and introduced such a plethora of new mathematical techniques. It is, perhaps, partly for this reason that when Einstein was awarded the Nobel Prize, it was for the photoelectric effect and Brownian motion and the Nobel Prize Committee hastened to add that it was *not* for speculative theories like General Relativity. (The other part would have been that Einstein was a *Jew* and the Germans had declared that the theory bore the hall-mark of Jews and so was not sound. Such political considerations have, unfortunately, often influenced decisions about awarding the Prize. It may be remarked that extremely important work of women has often been ignored by the Nobel Prize Committee. One can almost hear some of the older generation saying "We know that their minds are not made for such subjects."

Clearly, there was urgent need for fresh fields of applicability that could provide new tests. (This search for new fields of applicability of fresh ideas has, since then, become very popular in fundamental Physics and has been instrumental in driving the development and testing of new theories at an ever increasing pace.) As could have been expected of Einstein, he chose a most unexpected field for this purpose. Einstein chose Cosmology, from the Greek word cosmos ($\kappa\omega\sigma\mu\omega\sigma$) for "the Universe", and logos ($\lambda\omega\gamma\omega\sigma$) for "the study of". At the time this discipline would, more properly, have fallen into the realms of Theology than of Science. Theologists and Philosophers had for ages been debating the various aspects of the Universe in the "revealed religious literature".

Einstein's reason for choosing this field was *not* a desire to startle other scientists or to drag a field of Science out of Theology or Philosophy. Looked at in his typical way it was a necessity. As formulated at the time, his theory could be regarded as a non-linear field theory of gravity. There being no possible shielding from this force, there can be no "isolated systems" as are used in Thermodynamics or Electromagnetism for example. The non-linearity of the theory would make it impossible to neglect distant sources of gravity. More

correctly speaking, the purpose being to see deviations from Newtonian gravity, the only way to proceed was to try to construct a complete (internally consistent) model of the whole Universe and then adjust it so that it would fit more closely with the observed Universe.

Before going on to Einstein's attempt and the subsequent development of his programme, I would like to give a flavour of the subject before Einstein — specially the scientific part of the debates that had raged in this controversial subject. Let me mention also, for the sake of completeness, that *after* Einstein's development of *relativistic* cosmologies, *classical* cosmologies were also developed—essentially for pedagogic purposes. Since I do not feel that they make the subject any easier, or provide new insights, I will not dwell on them but will rather try to introduce modern Cosmology.

7.1 The Cosmological Principle

The fundamental assumption made in a physical science is that an experiment repeated at different places and times will give the same result, anywhere in the world and at any time. Till the middle of the twentieth century this assumption had been tested and found true for 'the world' regarded as the Earth and 'any time' over the past $2,000$ years or so. More recently spacecraft have enabled us to extend 'the world' to the entire solar system (at least up to the family of planets). However, to deal with the whole Universe and literally *any time*, requires an extrapolation of a totally different scale. The earliest observations and experiments spanned at most about 10^3 km and 10^9 sec. The Arabs took the scale to $\sim 10^4$ km and $\sim 10^{10}$ sec. By the middle of the twentieth century the experiments were vastly improved and the whole Earth, $\sim 10^5$ km, could be used for the test. By the end of the twentieth century we extended the spatial reach to $\sim 10^{10}$ km. The scales for the observed Universe are many orders of magnitude greater.

About $2,000$ years ago Aristotle and his followers had estimated the scale of the Universe as $\sim 10^5$ to 10^6 km while Eratosthenes and his followers estimated it as $\sim 10^8$ and 10^9 km. The age was given by most mythologies and religions as $\sim 10^{11}$ to 10^{12} sec (though some mythologies had many cycles of this duration). A millennium and more later, the Muslim civilization had improved on Eratosthenes estimate and increased it by an order of magnitude, $\sim 10^{10}$ km. With the advent of the telescope the entire world view changed.

According to Aristotelian science the laws governing the Earth and the Heavens were very different. The Earth was composed of the *mundane elements* (in order of increasing perfection): earth; water; air and fire. The Heavens were composed of the perfect (and eternal) element: aether. The Earth, being imperfect, was subject to change while the Heavens, being perfect, were unchanging. The mundane law of motion was that "every terrestrial object goes to its natural state of rest" (and hence "dust" goes "unto dust"). The celestial law of motion was that "all heavenly objects move in perfect circles except in so far as mundane contamination causes them to develop epicycles". An epicycle is the path described by an object moving about a centre in a perfect circle while that centre moves in a perfect circle about another centre. Similarly one could have more epicycles. The closer an object was to the Earth the more it would be contaminated by mundane elements causing it to become more changeable

and develop more epicycles. Thus the closest was the Moon, whose changeability and orbital changes were most apparent. Other planets were less and less changeable and had fewer epicycles. (The Sun and Moon were regarded as planets in those days.) These laws were, according to Aristotle, "self-evident truths" or obvious consequences of such self-evident truths as that "the Heavens are perfect", "the circle is a perfect figure", "the sphere is the perfect shape" etc.

Using his telescope Galileo was able to show that the Moon was a body much like the Earth and the natural laws that applied here also applied there! He measured the heights of the the lunar mountains by measuring the shadows they cast and mapped the lunar surface (incidentally mistaking seas of dust for water). His observations of the Moons of Jupiter and estimates of the sizes and distances of heavenly bodies supported the heliocentric model of the Universe, as opposed to Aristotle's Earth-centred view. Galileo's destruction of the Aristotelian world-view was followed by Newton's construction of a new physical world-view in which laws were to be "universal", applying to *all* objects — i.e. celestial and terrestrial. The scale of the Universe was now set as greater than 10^{10} km.

The discovery of atomic spectra and the use of spectroscopy in Astronomy established that the Sun and stars are composed of the same elements that we find on Earth. In fact, the element Helium was discovered spectroscopically in the Sun, first (hence its name) and later found on Earth. Flaring and exploding stars had been seen around the World since ages past, but were not believed in the Christian World because of Aristotelian dogma and the Church's elevation of that dogma to a revealed faith. Galileo had studied the flaring stars and named them novae (new stars) and seen and sketched the Crab nebula, which is the remnant of an exploded star (nowadays called a "supernova"). Further, proper motions of stars had been seen by comparing the positions of stars in those times and comparing thm with the star charts of the Greeks. Thus the heavens could, no longer, be regarded as "perfect and unchanging". The estimate of the distances of even the closer stars was $\sim 10^{14}$ km. For the Universe the estimate rapidly went up to 10^{17} km by the end of the nineteenth century.

By this stage the time scale had also gone up. In the early 1800s Pierre Simon Laplace that the energy of the Sun came from its gravitational collapse. This seemed to suggest an age of $\sim 10^{14}$ sec, while geological and palaeontological facts seemed to argue for an age $\gtrsim 10^{17}$ sec. Of course both estimates were based on theories that were hotly debated. Astronomical observations guided by our current understanding of Physics give a value of 4.4×10^{17} sec, to an accuracy of about 0.15%.

Naturally, the question arose whether the Universe may not be eternal and infinite in spatial extent. After all, religious beliefs always involved a Creator and hence a time of creation, so that the Universe was taken to have a start and it must have been well nigh impossible for people to conceive of a totally unlimited Universe. However, now it was a scientific question. In the early nineteenth century a German physician, Heinrich Wilhelm Matthias Olbers, tried to prove that the Universe must be finite. He assumed that the Universe has n stars per unit volume on average, with an average intrinsic luminosity of L_*. Then the apparent luminosity due to all the stars in a spherical shell of radius r about us and thickness dr, would be

$$dL = (nL_*/r^2)4\pi r^2 dr = 4\pi n L_* dr \ . \tag{7.1}$$

Thus, for an infinite Universe the total luminosity would be

$$L_\infty = 4\pi n L_* r|_{r=\infty} \ , \tag{7.2}$$

thus implying an infinitely bright sky. Even allowing $n \to 0$ we should expect a uniformly bright sky. Hence, Olbers argued, the Universe must be finite. A counter-argument was presented to prove that the Universe could not be finite. For if it were, there would be a geometrical centre which would act as a centre of gravity and cause the Universe to undergo gravitational collapse. Both arguments rely on classical Physics and would not hold in a relativistic analysis.

It was in this context that Einstein tried to apply his GR to Cosmology. In typical Einstein fashion he first elevated the assumption of the applicability of physical laws everywhere and "everywhen", to the status of a principle. He stated the *cosmological principle* as: "The Universe looks the same to all observers at all times". This is an *extremely* strong statement and restricts the possible cosmological models very severely. To be able to proceed with a physical theory of cosmology the *weakest* statement of this principle would be that: "Physical laws are invariant under spacetime translations (which remain within the Universe)". The former statement turns out to be too strong, in that it is inconsistent with observations and our current understanding of Physics. The latter is too weak to be able to get much out of it without additional assumptions. The form generally used nowadays is that: "The Universe looks the same to observers everywhere". The problem with this statement is that it is not Lorentz co-variant. In other words, it implies the existence of a preferred inertial frame — the rest-frame of the Universe. As we shall see this seems consistent with observation. However, it is not as yet entirely clear that the above statement of the cosmological principle is consistent with all observations.

Before proceeding to the specific models I would like to give some further observations concerned with the distance scale. Messier had identified some nebulae which interfered with stellar observations. Some of these turned out to be relatively close and were identified as the remains or the breeding grounds of stars. Others were found to be much further away and were identified as enormous collections of stars. These are now known as *galaxies* and generally consist of $\sim 10^{11}$ stars, most of which lie within a radius of $\sim 10^{17}$ km with a few percent reaching to $\sim 10^{18}$ km and an average distance of $\sim 10^{19}$ km between collections. These are not uniformly distributed but generally appear to be in clusters as large as $\sim 10^4$ galaxies. Our own galaxy is in a local cluster ~ 20 galaxies. Even these are not uniformly distributed but seem to lie on filaments with enormous voids $\sim 10^{20}$ km across. There is some reason to believe that the Universe is homogeneous at a scale of $\sim 10^{21}$ km but we can not be quite sure of that at present.

7.2 The Strong Cosmological Principle Models

For the Universe to look the same to all observers at all times, the physical parameters cannot depend on the space or time coordinates. Thus the Universe must be static, homogeneous, isotropic and spherically symmetric. The stress-energy tensor for a static, homogeneous, isotropic fluid is given by Eq.(5.26) with constants ρ and p. Thus, the principle of energy conservation applied in

the rest-frame gives

$$T^{\mu\nu}_{\ ;\nu} = 0 = [(\rho + p/c^2)c^2 g^{00}\delta^\mu_0\delta^\nu_0]_{;\nu} - [pg^{\mu\nu}]_{;\nu} \ . \tag{7.3}$$

Since the covariant derivative of the metric tensor vanishes and p is constant the second term automatically disappears. Using spherical polar type coordinates, Eq.(7.3) becomes

$$(\rho c^2 + p)[(e^{-\nu(r)}\delta^\mu_0\delta^\nu_0)_{,\nu} + e^{-\nu(r)}(\{_\nu\ ^\mu\ _0\}\,\delta^\nu_0 + \{_\nu\ ^\nu\ _0\}\,\delta^\mu_0)] = 0 \ , \tag{7.4}$$

where $\nu(r)$ refers to the metric coefficient and ν to the tensor index. Since there is no time dependence the first term in the square bracket and the last term are zero. Thus

$$(\rho c^2 + p)e^{-\nu(r)}\,\{_0\ ^\mu\ _0\} = 0 \ . \tag{7.5}$$

Using the tables of Christoffel symbols given in Eqs.(4.65) and (4.66) we see that Eq.(7.5) is automatically satisfied for $\mu \neq 1$. For $\mu = 1$

$$\frac{1}{2}\nu'(r)e^{-\lambda(r)}(\rho c^2 + p) = 0 \ . \tag{7.6}$$

This equation can be satisfied by taking either:

$$(i) \qquad \nu'(r) = 0 \ ; \tag{7.7}$$

or

$$(ii) \qquad \rho c^2 + p = 0 \ . \tag{7.8}$$

So far we have used the Einstein field equations. From Eqs.(4.57), (4.58), (4.62) and (4.63), without setting the components equal to zero, we see that the static, spherically symmetric Ricci tensor components are

$$R_{00} = \frac{1}{2}[\nu'' + \frac{1}{2}\nu'(\nu' - \lambda') + 2\nu'/r]e^{\nu-\lambda} \ , \tag{7.9}$$

$$R_{11} = -\frac{1}{2}[\nu'' + \frac{1}{2}\nu'(\nu' - \lambda') + 2\lambda'/r] \ , \tag{7.10}$$

$$R_{22} = [\frac{1}{2}r(\nu' - \lambda') - 1]e^{-\lambda} + 1 \ , \tag{7.11}$$

$$R_{33} = R_{22}\sin^2\theta \ , \qquad R_{\mu\nu} = 0 \text{ if } \mu \neq \nu \ . \tag{7.12}$$

The trace of the stress-energy tensor is

$$T = g^{\mu\nu}T_{\mu\nu} = \rho c^2 - 3p \ . \tag{7.13}$$

Thus the Einstein equations, in the form given by Eq.(4.64), become

$$R_{\mu\nu} = \kappa[(\rho c^2 + p)\delta^0_\mu\delta^0_\nu g_{00} + \frac{1}{2}(p - \rho c^2)g_{\mu\nu}] \ . \tag{7.14}$$

For case (i), using Eqs.(7.7) and (7.9) for the zero-zero part of Eq.(7.14), we get

$$\rho c^2 = -3p \ , \tag{7.15}$$

i.e. $T = 0$. Now the one-one and two-two parts of Eq.(7.14) give

$$\lambda'/r = -\frac{1}{2}\kappa(p - \rho c^2)e^\lambda = \kappa p e^\lambda \ , \tag{7.16}$$

$$(\frac{1}{2}r\lambda' - 1)e^{-\lambda} + 1 = -\frac{1}{2}\kappa r^2(p - \rho c^2)e^{-\lambda} = \kappa p r^2 e^{-\lambda} \ . \tag{7.17}$$

Solving Eqs.(7.16) and (7.17) simultaneously, we obtain

$$e^{-\lambda} = \frac{1 + \kappa p r^2/2}{1 + \kappa p r^2} \ . \tag{7.18}$$

Differentiating Eq.(7.18) we get a result incompatible with Eq.(7.16). Hence case (i) is not a viable solution.

For case (ii) the Einstein equations become

$$R_{\mu\nu} = -\kappa \rho c^2 g_{\mu\nu} \ . \tag{7.19}$$

As in section 4.4, the zero-zero and one-one parts of Eq.(7.19) give $\nu' + \lambda' = 0$, so that $\nu(r) = -\lambda(r)$. Using this value in the two-two parts of Eq.(7.19) we get

$$(r\lambda' - 1)e^{-\lambda} + 1 = (-re^{-\lambda})' + 1 = \kappa \rho c^2 r^2 \ . \tag{7.20}$$

Integrating the above equation gives

$$e^{-\lambda} = 1 - \frac{1}{3}\kappa \rho c^2 r^2 + C/r \ , \tag{7.21}$$

where C is a constant of integration. Requiring that $e^{-\lambda}$ be non-singular at $r = 0$, we get $C = 0$ and hence

$$e^{-\lambda(r)} = e^{\nu(r)} = 1 - \frac{1}{3}\kappa \rho c^2 r^2 = 1 - \frac{8\pi G \rho}{3c^2} r^2 \ . \tag{7.22}$$

It can be verified that the other Einstein equations are satisfied by this choice of $\nu(r) = -\lambda(r)$.

7.3 Enter the Cosmological Constant

Einstein had discarded the second case as being non-physical, since the pressure in the Universe is obviously negligible compared with the density, and so Eq.(7.8) cannot be satisfied. In other words it would have to put $\rho = p = 0$ and just give Minkowski space. As such the strong cosmological principle could be regarded as being incompatible with the field equations (unless one puts in a large negative pressure by hand). As mentioned earlier, Einstein introduced a constant of integration, which he called the *cosmological constant*, and modified his equations to the form given by Eq.(4.30). This would, in effect, provide the negative pressure required. In this case, on taking the trace and re-arranging terms, the Einstein equations become

$$R_{\mu\nu} = \kappa[T_{\mu\nu} - \frac{1}{2} T g_{\mu\nu}] - \Lambda g_{\mu\nu} \ . \tag{7.23}$$

He then assumed a negligible pressure and so, in place of Eq.(7.14), obtained

$$R_{\mu\nu} = \kappa[\rho c^2 g_{00} \delta^0_\mu \delta^0_\nu - \frac{1}{2}\rho c^2 g_{\mu\nu}] - \Lambda g_{\mu\nu} \ . \tag{7.24}$$

The zero-zero part of Eq.(7.24) gives

$$R_{00} = (\frac{1}{2}\kappa \rho c^2 - \Lambda) g_{00} \ . \tag{7.25}$$

Since $g_{00} \neq 0$, Eq.(7.7) tells us that

$$\Lambda = \frac{1}{2}\kappa\rho c^2 = 4\pi G\rho/c^2 \ . \tag{7.26}$$

Now, using this value in the one-one part of Eq.(7.24), we get

$$e^{-\lambda(r)} = C - \Lambda r^2 \ . \tag{7.27}$$

Since Special Relativity applies locally, at $r = 0$ the space must be Minkowski, and so $C = 1$. Hence, (for convenience in seeing the units) putting $\Lambda = R^{-2}$, we finally obtain the *Einstein Universe model*

$$ds^2 = c^2dt^2 - \frac{dr^2}{1 - r^2/R^2} - r^2d\Omega^2 \ . \tag{7.28}$$

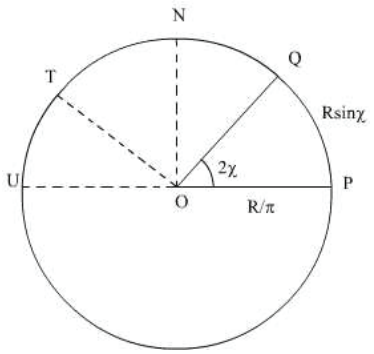


Figure 7.1: The "distance" given by the hyper-spherical angle χ on an S^3. If one passes the point N at $\chi = \pi/2$ the "distance" appears to shrink instead of increasing, so that at U it would again be zero.

The above metric can be cast into various forms which bring out some particular features. Since this is not a very realistic model and is only included to give historical perspective and because it is referred to for its mathematical structure, we will not go into a detailed discussion of this Universe model but will only look at the form that brings out its geometrical structure (and will be relevant for our further discussion). Changing variables to $r = R\sin\chi$, we see that Eq.(7.28) becomes

$$\begin{aligned} ds^2 &= c^2dt^2 - R^2(d\chi^2 + \sin^2\chi d\Omega^2) \\ &= c^2dt^2 - R^2d\Omega_3^2 \ , \end{aligned} \tag{7.29}$$

where $d\Omega_3^2$ is the hyper-spherical (S^3) analogue of the usual solid angle element. Since r is always positive, χ can only vary from 0 to π. Thus, the maximum distance on this 'sphere' is R, at $\chi = \pi/2$. One can think, then, of χ as an angle equal to half the angle on a circle of radius $R/2$. Thus, when χ passes $\pi/2$ the distance can be measured in the reverse direction and will shrink (see Fig. 7.1). Thus the Einstein Universe is a hyper-cylinder in 5-dimensions, with the time direction along the axis and each time slice being an S^3. Thus it is spatially

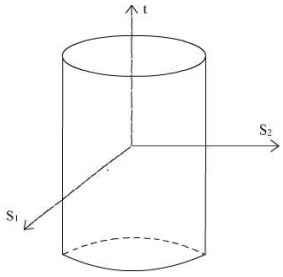


Figure 7.2: The Einstein Universe model with three dimensions reduced, two being suppressed and the third projected down. Time goes up and the projected circle represents a 3-sphere. The cylinder continues in both directions without end from past infinity to future infinity and S_1, S_2 are two of the four spatial directions.

a closed and bounded space. In time it is open and unbounded, being eternal (see Fig. 7.2).

The volume of the Einstein Universe model is given by

$$\begin{aligned} V &= R^3 \int_V d\Omega_3 = R^3 \int_{\chi=0}^{\pi} \int_{\theta=0}^{\pi} \int_{\varphi=0}^{2\pi} \sin^2\chi \sin\theta d\chi d\theta d\varphi \\ &= 2\pi^2 R^3 \ , \end{aligned} \tag{7.30}$$

giving a 'mass' of the entire Universe model,

$$M = \rho V = 2\pi^2 R^3 \rho \ . \tag{7.31}$$

Another form shows how Olbert's argument becomes invalid when the geometry changes. Writing $\bar{r} = R\chi$, the Einstein metric becomes

$$ds^2 = c^2 dt^2 - d\bar{r}^2 - R^2 \sin^2(\bar{r}/R) d\Omega^2 \ . \tag{7.32}$$

Since the luminosity law of reducing as r^2 comes from the geometrical optics of a uniform energy density being spread over a sphere, in this new geometry the factor will become $[R^2 \sin^2(\bar{r}/R)]^{-1}$. Thus there is re-focussing of the light at $\bar{r} = \pi R$. Of course, Olbert's argument is not needed as the model is already finite.

The cosmological constant was also introduced into case (ii) of section 2. The only difference made is that Λ is added to $\kappa\rho c^2$ in Eq.(7.19). Thus putting

$$R^2 = 3(\Lambda + 8\pi G\rho/c^2)^{-1} \ , \tag{7.33}$$

we obtain the *de Sitter Universe*

$$ds^2 = (1 - r^2/R^2)c^2 dt^2 - \frac{dr^2}{1 - r^2/R^2} - r^2 d\Omega^2 \ , \tag{7.34}$$

still with $p = -\rho c^2$ for energy conservation. Here we *can* consider an empty Universe model because of the Λ in Eq.(7.33).

For completeness I would like to also mention the case when $\Lambda < 0$ for Eq.(7.28) which can not be used directly in Eq.(7.33). There we use $-\Lambda$ in

place of Λ. In that case the metric coefficient becomes $(1 + r^2/R^2)$ and one needs to put $r = R\sinh\chi$. The Universe models now become hyper-hyperbolic instead of hyper-spherical. For case (i) we get the non-physical requirement $\rho < 0$ from Eq.(7.26). This is the *anti-Einstein* metric. For case (ii) it is possible to retain a 'physical' interpretation even when we get the change of the factor (if the left side of Eq.(7.33) becomes negative). This is called an *anti-de Sitter Universe model*. However, the anti-de Sitter Universe is not physically interesting but is used in the information loss paradox.

The metric given in Eq.(7.34) is not in a cosmologically convenient form because the time has to be re-scaled according to the distance from the observer. The cosmological time should be the same for all observers. Thus we need to transform to new cosmological coordinates $(\tau, \rho, \theta, \varphi)$ such that it remains diagonal but the zero-zero metric coefficient reduces to c^2. There is a procedure that can be adopted to derive the transformations to such coordinates. Here, however, it suffices to merely give the transformation. In infinitesimal form

$$\left.\begin{aligned} c d\tau &= c dt - \tfrac{r/R}{1-r^2/R^2} dr \\ d\rho &= [\tfrac{dr}{1-r^2/R^2} - \tfrac{r}{R} c dt] \tfrac{e^{-t/R}}{\sqrt{1-r^2/R^2}} \end{aligned}\right\} , \tag{7.35}$$

which can be integrated to give

$$\left.\begin{aligned} \tau &= t + (R/2c)\ln\left|1 - r^2/R^2\right| \\ \rho &= r e^{-ct/R}/\sqrt{1 - r^2/R^2} \end{aligned}\right\} . \tag{7.36}$$

These are the Lemaitre coordinates which give the Lemaitre form of the de Sitter Universe model

$$ds^2 = c^2 d\tau^2 - e^{2\tau c/R}(d\rho^2 + \rho^2 d\Omega^2) \ . \tag{7.37}$$

In the earlier form its geometry is clear as a hyper-hyperboloid of one sheet (see Fig. 7.3), while in the latter form it appears as an exponentially expanding flat 3-space which is infinite (open and unbounded). This latter picture is of great importance in modern discussions of Cosmology.

The spacetime is static (in that it has a timelike Killing vector which is globally orthogonal to a spacelike hypersurface). This staticity is apparent in the metric in the form given by Eq.(7.34), as there is no time dependence in it, even though it is not apparent in the Lemaitre form, which seems to be time dependent. Writing the geodesic equations for the previous form, we get the equation of motion for a test particle initially at rest at some point,

$$\ddot{r} - \frac{r}{R^2}(1 - \frac{r^2}{R^2})c^2\dot{t}^2 = 0 \ . \tag{7.38}$$

Putting $dr = d\theta = d\phi$ in the metric, we see that

$$c^2\dot{t}^2 = (1 - r^2/R^2)^{-1} \ . \tag{7.39}$$

Inserting the test particle mass, μ, Eqs. (7.38) and (7.39) give the force as

$$\mu\ddot{r} = \mu r/R^2 \ . \tag{7.40}$$

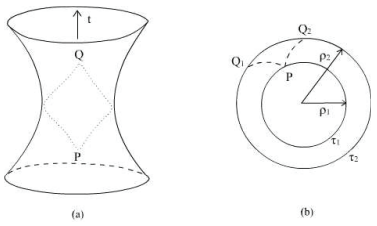


Figure 7.3: (a) The De Sitter model in static form with two dimensions suppressed and time going up. Light rays at P at time t_1 (on the front of the hyperboloid) focus at Q (on the back side) at a later time t_2. (b) The De Sitter model in expanding (Lemaitre) form, with only a spatial representation. The radius expands exponentially so that $\rho_2/\rho_1 = \exp(\tau_2 - \tau_1)/R$. Light starting at P at τ_1 appears to spread out due to the exponential expression but will focus on the other side at a much later τ than τ_2.

This is called the force of *cosmic repulsion* and is clearly linearly dependent on the distance of the test particle from the observer. *This* is what takes on the expansion effect. In the Lemaitre form there is no such force but simply the expansion itself. The $e\psi N$-formalism also gives the above force of cosmic repulsion.

7.4 The Measurement of Cosmological Distances

Since the cosmological models are constructed to explain how cosmological distances are changing with time, we need to take a brief look at the methods for determining them. Since distance measurements within the solar system were discussed in SR (and to some extent in section 1 of this chapter) I shall only mention the parallax method, of them, as that can be used also for the nearer stars. While one moves in one direction closer objects appear to move in the opposite direction while distant objects appear to remain fixed. This motion is due to the change of angular position of the nearer object. For a small change of one's position the change of angle is approximately the ratio of the perpendicular distance moved to the distance to the object, (given in radians). This is called the *parallax shift*. Taking the distance moved to be 1 au (an au is the mean distance of the Earth from the Sun, called an *astronomical unit*, and about 1.5×10^8 km) a parallax shift of 1 arc second would correspond to a distance of 3.09×10^{13} km. This distance is called a *parsec* (pc). Another common unit of distance measurement is the distance light travels in one year, a light year being about 9.46×10^{12} km. Thus 1 pc $\approx$ 3.27 ly. If a star were 1 pc away we would see a 2" parallax shift as the earth crossed its orbit. The nearest star is about 1.5 pc away and shows a 1.4" shift, which was not too difficult to observe. In fact the few near stars do show a 1" shift. However, for stars at $\sim$ 1kpc (1kpc = a kiloparsec = 10^3 pc), the shift becomes difficult to observe even these days. Hence better methods were required in the early attempts to map the known Universe.

The next major development was the use of the interferometer for resolving extremely small angles and hence providing an empirical relation between the sizes and intrinsic luminosities of stars. At a given temperature of the star its

brightness should depend simply on its surface area. The temperature can be estimated if one knows the spectrum of the radiation. This, in its turn, can be "measured" by its colour. Plotting the temperature versus luminosity one finds that there is a direct relationship between them. Conversely, plotting the luminosity against the "colour" measured by wavelength there is an inverse relationship for most stars. This plot is called the Hertzsprung-Russell diagram (see Fig. 7.4). The bulk of the stars fall on the main sequence but there are some "red-giant", stars which have high luminosity despite having a large wavelength (i.e. being relatively cool) and some white dwarfs with low luminosity and extremely short wavelength.

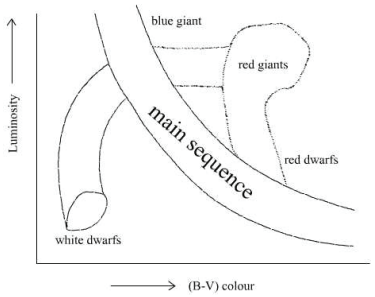


Figure 7.4: The Hertzsprung-Russell diagram plots the luminosity of light coming from the stars against their colour. Most stars lie along the main sequence, where the x-axis goes from blue to red (i.e. increasing wave-length of light), so that the top left corner are blue giants and the bottom right red dwarfs. The top right are red giants and the bottom left white dwarfs. The upper part gives high luminosity stars on account of their large radiating surface area and the lower part corresponds to small stars. The left side corresponds to hot and the right to relatively cool stars.

The understanding achieved about stellar structure led to better distance estimates on a statistical basis, since statistical methods were involved in the initial calibration and in the final estimation there were significant uncertainties. These uncertainties were dramatically reduced by further developments regarding stellar structure.

Though most stars have a constant luminosity over the usual periods of observation, some showed variations. For loose binaries the variation is due to Doppler shifts as they orbit about their common centre of mass. For close binaries there can be an additional variation due to the occlusion of one by the other, periodically. The period of such stars is of the order of days to years. However, some stars were seen to have variable intrinsic luminosity with a period varying from hours to days. The first such star studied was seen in the constellation of Cepheus (the Whale) and so these stars are called Cepheid variables. They provided an extremely precise "standard candle" for measuring distances.

The Cepheid variable was modelled as a star close to the limit of stability of equilibrium, between the gravitational force on the one hand and hydrostatic and radiation pressure on the other. The crux comes when the core pressure becomes adequate to support a nuclear process generating more energy, and hence radiation pressure. This causes the star to expand, thereby reducing the

core pressure and switching off the new nuclear process. This causes the star to collapse and re-ignite the higher nuclear reaction. Thus the star expands and shrinks periodically, causing a corresponding rise and drop in its intrinsic luminosity. The period depends on the mass of the star. The closer it is to the critical mass (called the Chandrasekhar mass $\sim 1.4\ M_\odot \approx 2.8 \times 10^{30}$ kg) the shorter the period. Since the luminosity of such a near critical mass star would be well defined, it is clear that a measurement of the period would give a very accurate estimate of the intrinsic luminosity of the star. This is why it is such a good "standard candle".

Early estimates of stellar distances based on Cepheid variables were off by a substantial factor because one did not know how to calibrate the distance with the luminosity. In other words, we did not know how far away our "standard candle" was. However, the discovery of a group of Cepheid variables whose distance was independently known provided the calibration. Various later developments give a picture entirely consistent with the estimates based on Cepheid variables. As such we now feel that we can rely on them.

By identifying a Cepheid dynamically bound to other stars, the distance to those stars was estimated. In particular, finding a Cepheid variable in a globular cluster gave the distance to that cluster of 10^5 and 10^6 stars. This allowed a detailed picture of the cluster, and the structure of the stars in it, to be built up. One could look at different globular clusters. It rapidly became apparent that they were much the mass and size. This allowed the distances of such clusters to be estimated very reliably. In this way the visible part of the galaxy, which looked like a lens, was mapped. Originally, this lens shaped structure was believed to be the entire Universe. However, while some "nebulae" were seen inside this region it became apparent that others lay outside it. These external "nebulae" turned out to be separate galaxies.

The closest galaxies are the Smaller and Larger Magellanic Clouds, seen in the Southern hemisphere. They are dwarf galaxies $\sim 10^9\ M_\odot$ which are our satellites. The closest genuine galaxy is M31 in the constellation known as Andromeda, and is consequently known as the Andromeda galaxy. It consists $\sim 10^{11}$ stars. It soon became clear that it is extremely similar to our galaxy. The core is about 10 kpc across and it is about 5×10^2 kpc away and so subtends an angle of about 1°. When enlarged one could see that around the core, which is roughly spherical, there is a large region in which matter is spread along some spiral arms. In our galaxy, we are some distance out along a spiral arm and the bulk of the galaxy is occluded by interstellar "dust clouds". We have, since, been able to observe the entire region of our galaxy by using other parts of the electromagnetic spectrum than the visible. There are about 25 galaxies in our local cluster. Using Cepheids in these galaxies gave the distances to them. Direct measurements of the angular diameter gave their sizes. One could, thereby, estimate the number of stars in galaxies, their typical size and range of variation and their types. One knew their intrinsic and apparent luminosity and their constitution (by spectroscopic methods) and even the movements of parts of them (by Doppler shifts).

The local cluster is gravitationally bound. One could see more distant galaxies and superclusters of up to $\sim 10,000$ galaxies. On average, galaxies are about 10 kpc across and many galaxies have smaller or larger spiral arms (while some have no arms). The total diameter may go up to 40 to 60 kpc (or more if one is ready to count individual stars at the extreme ends as constituting the

galaxy). Till now $\sim 10^{12}$ galaxies have been observed. There is evidence that the structure of arrangement of matter seen in the Universe, does not stop at superclusters. All the matter appears to lie along large filaments $\sim 10^4$ kpc long and separated by large voids. It appears that there is homogeneity in the Universe at $\sim 10^5$ kpc scale but one can not yet be too sure of it.

7.5 The Hubble Expansion of the Universe

On the basis of his observations, in 1929 Hubble reported the startling finding that, going beyond the local cluster, there is an average red-shift of galaxies which increases with their distance with a linear relationship. Taking the red-shift to be due to the Doppler effect, the observation indicates that all galaxies are rushing away from us as if we have bad breath. This caused tremendous excitement at the time, as it violates the cosmological principle and put us at the centre of an explosion. The principle can be partially saved by assuming that all observers in the Universe see distances to all points in the Universe as increasing. However, this still violates the strong cosmological principle, as the matter density in a given (large) volume of space decreases with time.

Assuming the weak cosmological principle we are then forced to a metric of the form

$$ds^2 = c^2dt^2 - a^2(t)d\sigma^2 \ , \tag{7.41}$$

where $d\sigma^2$ is the metric for a homogeneous 3-space,

$$d\sigma^2 = d\chi^2 + f_k^2(\chi)d\Omega^2 \ , \tag{7.42}$$

$f_k(\chi) = \sin\chi$ for $k = +1$, χ for $k = 0$ and $\sinh\chi$ for $k = -1$. The Lemaitre form of the de Sitter metric is a special case of this general metric with $k = -1$ and $a(t) = R\ e^{ct/R}$. Could the Universe actually be de Sitter?

To answer this question we need to be more quantitative about how the speeds of galaxies, or more directly their red-shift, changes with distance. Writing the red-shift, $\Delta\lambda/\lambda$, as z, Hubble's empirical estimate was that it is directly proportional to the distance r. Thus

$$z = Hr/c \ , \tag{7.43}$$

where H is the proportionality constant in units of inverse time (so that the corresponding speed is Hd). Much is made of the fact that a quadratic law would have fitted the data better. To me this seems a non-issue. Surely, the greater the number of parameters used, the better the fit that would be obtained. Knowing the enormous uncertainties involved and highly statistical nature of his empirical law, he simply felt that no refinements of the 1-parameter fit were justified by observation. For a 1-parameter fit this law seemed best. Current understanding fits the data superbly. The semi-historical reinterpretation of Hubble's data, thus, seems pointless.

The speed of a galaxy would be simply $v = dr(t)/dt$. In Eq. (7.41) it is clear that $dr(t) = a(t)d\chi$. Thus we see that we should have

$$v = \frac{da(t)}{dt}\chi = \frac{\dot{a}(t)}{a(t)}a(t)\chi = \frac{\dot{a}(t)}{a(t)}r \ . \tag{7.44}$$

Using the de Sitter model we get $H = c/R$ (correcting for using non-gravitational units). The observed value of H is (73.8 ± 2.4) km/sec/Mpc. This gives a radius

of $(4.43 \pm 0.14) \times 10^3$ Mpc. Note that here we have H as a genuine constant, but this would not generally be true. In fact, if we require that H is a constant we are forced to take $a(t) = e^{ct/R}$. How can we check if the Universe is *actually* de Sitter?

The answer comes by considering the general model. In that case we have another quantity that can be measured, called the "braking index", which is dimensionless

$$q(t) = \frac{-a(t)\ddot{a}(t)}{[\dot{a}(t)]^2} \ . \tag{7.45}$$

This gives the rate of slowing-down of the expansion, if any. For the de Sitter model it is clearly -1. However, if $a(t)$ follows a power-law, $a(t) = a_0 t^n$, then $q = (1-n)/n$. Thus, if $n < 1$, q will be positive while for $n > 1$ it is a negative number greater than -1. Notice that now H would *not* be a constant. Defining the Hubble parameter $H(t) = \dot{a}(t)/a(t)$, we see that in general

$$\dot{H}(t) = -H^2(t)\left[1 + q(t)\right] \ . \tag{7.46}$$

Incidentally, Eq.(7.43) was known as Hubble's law and H as the Hubble constant. In view of the later developments Eq.(7.44) is regarded as a statement of Hubble's law now. It is obvious that this will look, empirically, very different from Hubble's law if $\dot{H}(t) \neq 0$ as the "constant of proportionality" changes with time, it will change with distance. This deviation from Hubble's original law has been observed and it corresponds to a value of q close to $1/2$. There is no evidence that $q(t)$ is not constant in the observable part of the Universe.

As a brief historical remark it may be mentioned that errors in the distance scale for the Universe gave a very different value of H in Hubble's time. The changes in the estimates for H have led many people of believe that observational cosmology in unreliable. As I have tried to indicate, this is no longer true. Our estimates of the very large are no less reliable than of the very small. This does not mean that they are as accurate, necessarily. There remains an uncertainty of a factor of 2 in the measurement of the Hubble parameter H, but we *do* know that it lies within that broad range of values *and nowhere else*.

7.6 The Einstein-Friedmann Models

Another model for the Universe had been derived by Friedmann in 1922. In fact this solution to the Einstein equations had been obtained much earlier by Einstein himself, but was discarded by him as non-physical, on account of his faith in the strong cosmological principle. Most people refer to this solution as Friedmann's. However, since Einstein *had* obtained it earlier and *did* believe in its validity later, I include Einstein's name as well.

Using the metric given by Eqs.(7.41) and (7.42), Since the metric coefficients are now time dependent we must re-derive the equations for the Ricci tensor components. Here

$$\left.\begin{array}{ll} g_{00} = 1 \ , & g_{11} = -a^2(t) \ , \\ g_{22} = -a^2(t) f_k^2(\chi) \ , & g_{33} = -a^2(t) f_k^2(\chi) \sin^2\theta \ , \\ g^{\mu\mu} = \left(g_{\mu\mu}\right)^{-1} \ . & \end{array}\right\} \tag{7.47}$$

Writing f, instead of $f_k(\chi)$, the χ derivative by f', χ being the 1-coordinate. The non-zero Christoffel symbols are, (using gravitational units for the present)

$$\left.\begin{array}{l} \{_1\ ^0\ _1\} = a\dot{a}\ , \qquad \{_2\ ^0\ _2\} = a\dot{a}f^2\ , \qquad \{_3\ ^0\ _3\} = a\dot{a}f^2\sin^2\theta\ , \\ \{_2\ ^1\ _2\} = -ff'\ , \quad \{_3\ ^1\ _3\} = -ff'\sin^2\theta\ , \\ \{_0\ ^1\ _1\} = \{_0\ ^2\ _2\} = \{_0\ ^3\ _3\} = \dot{a}/a\ , \\ \{_1\ ^2\ _2\} = \{_1\ ^3\ _3\} = f'/f\ , \quad \{_2\ ^3\ _3\} = \cot\theta\ , \quad \{_3\ ^2\ _3\} = -\sin\theta\cos\theta\ , \end{array}\right\} \tag{7.48}$$

and those obtainable by the symmetry property. It again turns out that the off-diagonal components are zero. For example, consider

$$\left.\begin{aligned} R_{01} &= \{_0\ ^\mu\ _1\}_{,\mu} - \left(\ln\sqrt{|g|}\right)_{,01} + \left(\ln\sqrt{|g|}\right)_{,\mu}\{_0\ ^\mu\ _1\} - \{_\nu\ ^\mu\ _0\}\{_\mu\ ^\nu\ _1\} \\ &= \{_0\ ^1\ _1\}_{,1} - \left[\ln\left(a^3f^2\sin\theta\right)\right]_{,01} + \left[\ln\left(a^3f^2\sin\theta\right)\right]_{,1}\{_0\ ^1\ _1\} \\ &\quad -\{_2\ ^2\ _0\}\{_2\ ^2\ _1\} - \{_3\ ^3\ _0\}\{_3\ ^3\ _1\} \\ &= \left(\frac{\dot{a}}{a}\right)' - \left(\frac{3\dot{a}}{a}\right)' + \left(\frac{2f'}{f}\right)\left(\frac{\dot{a}}{a}\right) - \left(\frac{\dot{a}}{a}\right)(f'/f) \\ &\quad - \left(\frac{\dot{a}}{a}\right)(f'/f) = 0\ . \end{aligned}\right\} \tag{7.49}$$

For the diagonal components we have

$$\begin{aligned} R_{00} &= \{_0\ ^\mu\ _0\}_{,\mu} - \left(\ln\sqrt{|g|}\right)_{,00} + \left(\ln\sqrt{|g|}\right)_{,\mu}\{_0\ ^\mu\ _0\} - \{_\nu\ ^\mu\ _0\}\{_\mu\ ^\nu\ _0\} \\ &= -3\left(\dot{a}/a\right)^{\cdot} - \{_1\ ^1\ _0\}^2 - \{_2\ ^2\ _0\}^2 - \{_3\ ^3\ _0\}^2 \\ &= -3\ddot{a}/a + 3\dot{a}^2/a^2 - 3\dot{a}^2/a^2 \\ &= -3\ddot{a}/a\ , \end{aligned} \tag{7.50}$$

$$\begin{aligned} R_{11} &= \{_1\ ^\mu\ _1\}_{,\mu} - \left(\ln\sqrt{|g|}\right)_{,11} + \left(\ln\sqrt{|g|}\right)_{,\mu}\{_1\ ^\mu\ _1\} - \{_\nu\ ^\mu\ _1\}\{_\mu\ ^\nu\ _1\} \\ &= \{_1\ ^0\ _1\}_{,0} - 2\left(f'/f\right)' + 3\left(\dot{a}/a\right)\{_1\ ^0\ _1\} - 2\{_1\ ^0\ _1\}\{_0\ ^1\ _1\} \\ &\quad -\{_2\ ^2\ _1\}^2 - \{_3\ ^3\ _1\}^2 \\ &= \left(a\dot{a}\right)^{\cdot} - 2f''/f + 2f'^2/f^2 + 3\left(\dot{a}/a\right)a\dot{a} - 2a\dot{a}\left(\dot{a}/a\right) \\ &\quad -f'^2/f^2 - f'^2/f^2 \\ &= a\ddot{a} + 2\dot{a}^2 - 2f''/f\ , \end{aligned} \tag{7.51}$$

$$\begin{aligned} R_{22} &= \{_2\ ^\mu\ _2\}_{,\mu} - \left(\ln\sqrt{|g|}\right)_{,22} + \left(\ln\sqrt{|g|}\right)_{,\mu}\{_2\ ^\mu\ _2\} - \{_\nu\ ^\mu\ _2\}\{_\mu\ ^\nu\ _2\} \\ &= \{_2\ ^0\ _2\}_{,0} + \{_2\ ^1\ _2\}_{,1} - \left(\ln\sin\theta\right)_{,22} + 3\left(\dot{a}/a\right)\{_2\ ^0\ _2\} + 2\left(f'/f\right)\{_2\ ^1\ _2\} \\ &\quad -2\{_2\ ^0\ _2\}\{_0\ ^2\ _2\} - 2\{_2\ ^1\ _2\}\{_1\ ^2\ _2\} - \{_3\ ^3\ _2\}^2 \\ &= \left(a\dot{a}f^2\right)^{\cdot} - \left(ff'\right)' - \left(\cot\theta\right)_{,2} + 3\left(\dot{a}/a\right)\left(a\dot{a}f^2\right) \\ &\quad -2\left(f'/f\right)\left(ff'\right) - 2\left(a\dot{a}f^2\right)\left(\dot{a}/a\right) + 2\left(ff'\right)\left(f'/f\right) - \cot^2\theta \\ &= \left(a\ddot{a} + 2\dot{a}^2\right)f^2 - \left(ff'' + f'^2\right) + 1\ , \end{aligned} \tag{7.52}$$

and it turns out that, as usual, $R_{33} = R_{22}\sin^2\theta$.

Notice that if we take $k = +1$: $f = \sin\chi$, $f' = \cos\chi$, $f'' = -\sin\chi$ and Eq.(7.52) gives $R_{22} = R_{11}\sin^2\chi$. For $k = -1$: $f = \sinh\chi$, $f' = \cosh\chi$, $f'' =$

$-\sinh\chi$ and Eq.(7.52) gives $R_{22} = R_{11}\sinh^2\chi$. Thus we see that generally, for all allowed k,

$$R_{22} = R_{11} f_k^2(\chi) \ . \tag{7.53}$$

Also notice that

$$\begin{aligned} -f_k''(\chi)/f_k(\chi) &= +k \\ &= \left[1 - f_k^{'2}(\chi)\right]/f_k^2(\chi) \end{aligned} \tag{7.54}$$

in each case. Now we can simply obtain the Ricci scalar

$$R = g^{\mu\nu}R_{\mu\nu} = -6\left(\ddot{a}/a + \dot{a}^2/a^2 + kc^2/a^2\right) \ . \tag{7.55}$$

Assuming a homogeneous isotropic Universe, with the density and pressure now functions of time, we get the Einstein equations

$$3\left(\frac{\dot{a}^2(t)}{a^2(t)} + \frac{kc^2}{a^2(t)}\right) = 8\pi G\rho(t) \ , \tag{7.56}$$

$$2a(t)\ddot{a}(t) + \dot{a}^2(t) + kc^2 = 8\pi G p(t)/c^2 \ . \tag{7.57}$$

Rather than using the second equation we use the energy conservation equation (7.3) for the metric given by Eqs.(7.41) and (7.42), since $g_{00} = g^{00} = 1$ here and $\{{}^{\mu}_{0\ 0}\} = 0$, we obtain the equation

$$\delta^\mu_0\left\{\dot{\rho}(t) + 3\left[\dot{a}(t)/a(t)\right]\left[\rho(t) + p(t)/c^2\right]\right\} = 0 \ . \tag{7.58}$$

Given an equation of state relating $\rho(t)$ to $p(t)$ we can solve Eqs.(7.56) and (7.58) simultaneously. Since this is simply the Bianchi identity, Eq.(7.57) must follow from Eqs.(7.56) and (7.57). As Eqs.(7.58) and (7.56) are an easier system to solve we use them.

We can consider some special equations of state. To describe the Universe as we see it now, it would appear reasonable to regard it as dust, on a sufficiently large scale, and consequently pressureless. Thus the equation of state would be $p(t) = 0$. In this case Eq.(7.58) immediately yields

$$\rho(t)a^3(t) = \text{constant} = C_d \text{ (say)}. \tag{7.59}$$

On the other hand, visualise a Universe consisting of only electromagnetic radiation. For that we know that the trace of the stress-energy tensor is zero, as is clear from Eq.(5.43). Thus we have the equation of state for radiation

$$\rho(t) - 3p(t)/c^2 = 0 \ . \tag{7.60}$$

Using this value in Eq.(7.57) and integrating yields

$$\rho(t)a^4(t) = \text{constant} = C_r \text{ (say)}. \tag{7.61}$$

The above results are reasonable and could have been guessed directly. Since $a(t)$ gives the re-scaling of distance with time, from mass conservation we would automatically obtain Eq.(7.59). For radiation we have to remember that if the *number* of photons is conserved in a given region the re-scaling leads to Eq.(7.59) for the *number density*, rather than the mass density. Since the photons are also

red-shifted the energy density picks up an extra factor of $a(t)$. Thus the mass density (due to the famous $E = mc^2$) satisfies Eq.(7.61).

We are now in a position to solve Eq.(7.56). We can replace $\rho(t)$ by some function of $a(t)$, depending on the equation of state. Since we have a first order differential equation in $a(t)$ we can solve it, at least numerically. For $k = 0$ the equation becomes

$$\dot{a}(t) = \pm\sqrt{8\pi GC_d/3} \cdot 1/\sqrt[3]{a(t)} \ , \tag{7.62}$$

for pure dust and

$$\dot{a}(t) = \pm\sqrt{8\pi GC_r/3} \cdot 1/\sqrt{a(t)} \tag{7.63}$$

for pure radiation. Thus, for pure dust

$$a(t) = (6\pi GC_d/3)^{2/3} (t - t_0)^{2/3} \ , \tag{7.64}$$

and for pure radiation

$$a(t) = (32\pi GC_r/3)^{1/2} (t - t_0)^{1/2} \ . \tag{7.65}$$

Setting the origin of time when $a(t)$ is zero we choose $t_0 = 0$ and write the proportionality constant as simply A_d or A_r. Then

$$a_d(t) = A_d t^{2/3} \ , \ a_r(t) = A_r t^{1/2} \ . \tag{7.66}$$

Now consider the case $k = +1$. Put $8\pi GC_d/3c^2 = a_d$. Then, for pure dust Eq.(7.56) becomes

$$\dot{a}^2(t) + c^2 = a_d c^2/a(t) \ . \tag{7.67}$$

To solve this equation put $\dot{a}(t) = c\cot\xi$. Then clearly Eq.(7.66) gives

$$a(\xi) = a_d \sin^2\xi \ . \tag{7.68}$$

Taking the differential of Eq.(7.68) we can write

$$\dot{a}(t) = \frac{da(\xi)}{dt(\xi)} = \frac{2a_d \sin\xi\cos\xi d\xi}{dt(\xi)} = c\frac{\cos\xi}{\sin\xi} \ . \tag{7.69}$$

Integrating this equation to obtain $t(\xi)$ we get

$$t(\xi) = (a_d/2c)(2\xi - \sin 2\xi) \ . \tag{7.70}$$

Putting $2\xi = \eta$ we obtain the equation of the cycloid in parametric form

$$a(\eta) = (a_d/2)(1 - \cos\eta) \quad , \quad t(\eta) = (a_d/2c)(\eta - \sin\eta) \ . \tag{7.71}$$

Notice that we could have put $\dot{a}(t) = c\tan^2\xi$. As can be checked easily, this gives a negative time and an $a(\eta)$ that is not zero for positive η. By re-definitions it reduces to this solution with a shifted origin of time.

For pure radiation, putting $a_r^2 = 8\pi GC_r/3c^2$, Eq.(7.56) becomes

$$\dot{a}^2(t) + c^2 = a_r^2 c^2/a^2 \ . \tag{7.72}$$

We can follow the previous procedure or directly solve the equation by changing the dependent variable to $a^2(t)$ instead of $a(t)$. Either way we obtain the equation of a circle

$$a^2(t) + c^2 t^2 = a_r^2 c^2 \ . \tag{7.73}$$

The full circle comes from dealing with squares. Taking the physical requirement that $a(t) > 0$ for all permissible t, we see that we get the semi-circle

$$a(t) = c\sqrt{a_r^2 - t^2}\ , \tag{7.74}$$

whence $-a_r < t < a_r$. We can shift the origin of time so that $a(t) = 0$ at $t = 0$ to get

$$a(t) = c\sqrt{2a_r t - t^2} \qquad (0 \le t \le 2a_r)\,. \tag{7.75}$$

For the case $k = -1$, we must convert to hyperbolic functions. Putting $\dot{a}(t) = \coth\xi$ and following the previous procedure, for the pure dust case, gives

$$a(\eta) = (a_d/2)\,(\cosh\eta - 1) \quad , \quad t(\eta) = (a_d/2c)\,(\sinh\eta - \eta)\ . \tag{7.76}$$

For pure radiation the equation becomes of a hyperbola

$$c^2t^2 - a^2(t) = a_r^2c^2\ . \tag{7.77}$$

Again, re-setting the time

$$a(t) = c\sqrt{2a_r t + t^2} \qquad (0 \le t \le \infty)\,. \tag{7.78}$$

This completes the solution of the special choice equation of state cases of the Friedmann evolution equation (7.56). We now need to discuss some of them a bit further.

Let us consider the behaviour of the various models for early time. First take dust with $k = +1$. To lowest order Eq.(7.71) gives

$$a(\eta) \approx a_d\eta^2/4 \quad , \quad t(\eta) \approx a_d\eta^3/12c\ . \tag{7.79}$$

Thus, using the latter equation to express η in terms of t and insert into the former equation, we obtain

$$a(t) \approx \left(9a_dc^2/4\right)^{1/3} t^{2/3}\ . \tag{7.80}$$

For the next order correction we write Eq.(7.77) to one higher order,

$$\left.\begin{aligned} a(\eta) &= \left(a_d\eta^2/4\right)\left(1 - \eta^2/12\right)\ , \\ t(\eta) &= \left(a_d\eta^3/12c\right)\left(1 - \eta^2/20\right)\ . \end{aligned}\right\} \tag{7.81}$$

Now, writing the zero-order expression of η in terms of t as η_0, we can ask for the ϵ which will satisfy the second of Eqs.(7.81) when we put $\eta = \eta_0\,(1 + \epsilon)$ and take lowest order in ϵ. This gives $\epsilon = \eta_0^2/60$. Inserting this η into the expression for $a(\eta)$ and find that

$$\begin{aligned} a(t) &= \left(\frac{9a_dc^2}{4}\right)^{1/3} t^{2/3}\left(1 - \frac{\eta_0^2}{20}\right) \\ &= \left(\frac{9a_0c^2}{4}\right)^{1/3} t^{2/3}\left[1 - \left(\frac{9c^2}{500{a_d}^2}\right)^{1/3} t^{2/3}\right] \\ &\approx \left(9a_dc^2/4\right)^{1/3} t^{2/3}\left[1 - 0.262\,(ct/a_d)^{2/3}\right]\ . \end{aligned} \tag{7.82}$$

Following the same procedure for $k = -1$, we get

$$a(t) \approx \left(9a_d c^2/4\right)^{1/3} t^{2/3} \left[1 + 0.262\left(ct/a_d\right)^{2/3}\right] \ . \tag{7.83}$$

Thus there is no distinction between the $k = +1, 0, -1$ models to lowest order in t. However, at the next order we see that $a_+(t) < a_0(t) < a_-(t)$ for the same time. The comparison of the radiation models directly gives the same result.

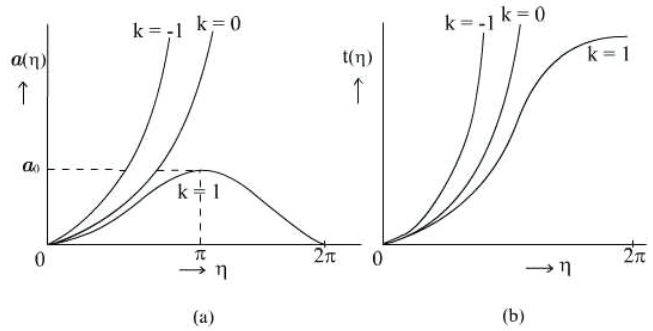


Figure 7.5: 7.5: (a) The scale factor as a function of η for the three Friedmann models: $k = 1, k = 0$ and $k = -1$. Note that the model with $k = 1$ reaches a maximum, a_0, at $\eta = \pi$ and then re-collapses to zero at $\eta = 2\pi$. Though the expansion in this case does not appear symmetric about $\eta = \pi$, the functional form of $a(\eta)$ is clearly symmetric. For $k = 0$ or -1 there is no reason to stop at $\eta = 2\pi$. (b) The time as a function of η for the three Friedmann models: $k = 1, k = 0$ and $k = -1$. Again, only the $k = 1$ model stops at $\eta = 2\pi$.

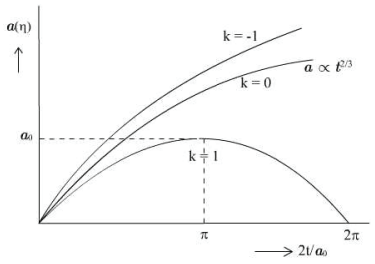


Figure 7.6: The scale factor as a function of time for the three Friedmann models. At $t = a_0\pi/2$ the $k = 1$ model reaches the maximum value a_0 and re-collapses at $t = a_0\pi$. Here $a_0(t)$ is not symmetric about $t = a_0\pi/2$. The other two models do not stop at $t = a_0\pi$. The $k = 0$ model has $a(t) \propto t^{2/3}$ while the $k = -1$ model has $a(t) \sim t^{2/3+\varepsilon}, \varepsilon > 0$.

The $k = +1$ models are specially interesting because $a(t)$ is not a monotonic function. All models start at the same point $a(0) = 0$, but for $k = +1$ we also have an "end" at $a(\eta = 2\pi) = 0$. Thus $a(t = a_d\pi k) = 0$. Further $a(\eta = \pi) = a(t = \frac{\pi a_d}{2c}) = a_d$ in the dust model. Similarly, for the radiation model and $k = +1$ $a(t = 0) = a(t = 2a_r) = 0$ and at $t = a_r$, $a(t) = ca_r$. Though one can take repeated cycles of the dust model, it does not seem to be relevant as $a(t) = 0$ is a singularity. We shall see, shortly, that it is an essential singularity.

Thus the $k = +1$ models have a finite life from beginning to end while the other models are infinitely long-lived but do have a start. The behaviour of $a(t)$ in all models is shown graphically in Figs. 7.5 and 7.6. Of course, the only way to determine the validity of any model is by confronting the theory with observation and not by looking at interesting features.

It is necessary to see how these models fit with the red-shift and braking index observations. Clearly, here $H(t)$ is not a constant. For the $k = 0$ case $H(t) = 2/3t$. Since all three models have a similar behaviour for early times they all showed this feature. This gave the present age of the Universe but the value of the Hubble parameter was not very reliably known, giving an age of 6.5 to 13 billion years. We now know that the upper limit is nearer the correct value. Of course, if $a(t) < a_0(t)$ the correction lowers $H(t)$ and hence raises the age estimate. Contrariwise, if $a(t) > a_0(t)$ the age estimate would be lowered. There have been other relevant observations of great significance that will be talked of later.

We now come to the braking index. To lowest order, the pure dust models give $q \approx 1/2$ while the pure radiation models give $q = 1$. The next order correction leaves the $k = 0$ case unaltered and gives a correction for $k = \pm 1$

$$q(t) \approx \frac{1}{2}\left[1 \mp (3ct/2a_d)^{2/3}\right] \ , \tag{7.84}$$

for pure dust and

$$q(t) \approx 1 \pm 2t/a_r \ , \tag{7.85}$$

for pure radiation. In the 1980s, the estimate came out at $q(t) = 1/2$. The later developments mentioned above render the relevance of the value moot.

Since the spatial geometry of the $k = +1$ model is of a 3-sphere (S^3), of the $k = 0$ model of a 3-cone (which is "flat") and of the $k = -1$ model of a 3-hyperboloid, they are called spherical, flat or hyperbolic models. They are also called closed, flat and open. In all cases the 3-space expands (and may possibly re-collapse for some cases). The sphere would have a finite volume which can be worked out,

$$V(t) = a^3 \int_{\chi=0}^{\pi}\int_{\theta=0}^{\pi}\int_{\varphi=0}^{2\pi} \sin^2\chi \ \sin\theta \ d\varphi d\theta d\chi = 2\pi^2 a^3(t) \ . \tag{7.86}$$

Multiplying this by the density gives the "mass" in the Universe. (Here care needs to be taken in interpretation of this "mass" due to the fact that the gravitational energy has not been accounted for and that would have an effective negative mass. (In fact, it would exactly cancel this mass.) Thus, for a pure dust model

$$M = \rho(t)V(t) = 2\pi^2 C_d \ . \tag{7.87}$$

This relation explains what the constant, C_d, is. In the other cases we have no concept of a "total mass" in the Universe but we could think of the total mass in the *observable* part of the Universe. That would give the significance of C_d more generally. For pure radiation, we get

$$M(t) = 2\pi^2 C_r/a(t) \ . \tag{7.88}$$

Taking $t = a_r$, we see that

$$M(a_r) = 2\pi^2 C_r/ca_r \ , \tag{7.89}$$

is the minimum value of $M(t)$ and it goes to infinity at $t = 0, 2a_r$.

This completes our discussion of the Einstein-Friedmann models.

7.7 Saving the Strong Cosmological Principle

The reason why the strong cosmological principle is violated by Hubble's observations is that they indicate a progressive decrease in matter density. The reason why the de Sitter model can be consistent with both is that it represents an empty Universe. How can a matter filled Universe be ruled by the strong cosmological principle and yet allow Hubble's law? The answer is to violate the law of conservation of mass-energy. Bondi, Hoyle, Narlikar and others examined this alternative. The resulting theories were called the Steady State Theories.

The fundamental idea is, as mentioned above, to hold the energy density constant in an expanding Universe, by having matter-energy created continuously at the rate at which the expansion causes it to decrease. Hence, $T^{\mu\nu}{}_{;\nu} \neq 0$. One introduces a new C-field, $C^{\mu\nu}$, which balances the violation of the matter-energy conservation equation required by the Einstein equations and the Bianchi identities. The matter is supposedly created in the interstices left as galaxies move away from each other. The matter so created forms new galaxies and so the process continues.

These theories suffered from the flaw, as regards scientific methods, that they violate basic principles so as to fit observation without making properly testable predictions. Between energy-matter conservation and the strong cosmological principle, most of us would be ready to sacrifice the latter rather than the former. However, a prediction of sorts does emerge. Younger and older galaxies should be mixed together uniformly according to these eternal Universe theories. However, if the Universe had a beginning we should see young galaxies far away and older galaxies closer by, as the further galaxies are seen as they were a long time ago. Observation supports the finite age theories of the Universe, as the stars in very remote galaxies show the characteristics of young galaxies that have not had time to reprocess the primordial matter as much as the closer galaxies have had. One can find ways to avoid the discrepancies between the expectations based on these theories and observation. However, most of us find those proposals forced and artificial and consequently regard the Steady State Theories as dead. In what follows I shall practically entirely ignore these theories.

7.8 The Hot Big Bang Model Of Gamow

In all the Friedmann models there is an origin of time at which the spatial line element becomes zero and hence the spacetime metric appears to be singular. We need to check whether this is merely a coordinate, or a genuine curvature singularity. Now the curvature scalar given by Eq.(7.55) can be reduced to

$$R = -8\pi G \left[\rho(t) + 3p(t)\right] \ , \tag{7.90}$$

by using Eq.(7.57). It is easy to check that regardless of whether we use the pure dust or pure radiation equation of state $R \sim t^{-2}$ and so the singularity is a genuine one. The only equation of state in which the above behaviour will not

be seen is if $p(t) = -\rho(t)/3$. In this case $R = 0$ for all t and we need to check R_2.

It is easy to verify that the above equation of state gives

$$\rho(t)a^2(t) = \text{ constant } = C_0 \text{ (say)}, \tag{7.91}$$

directly from Eq.(7.58) and then Eq.(7.56) becomes

$$\dot{a}^2 + kc^2 = 8\pi G C_0/3 \ , \tag{7.92}$$

which immediately gives $a(t) = a_0 t$ (say) and $\dot{a}(t) = a_0$, $\ddot{a}(t) = 0$. We now need to evaluate the components of the Riemann tensor.

$$\begin{aligned} R^0_{101} &= \{_1\ ^0\ _1\}_{,0} - \{_0\ ^0\ _1\}_{,1} + \{_\mu\ ^0\ _0\}\{_1\ ^\mu\ _1\} - \{_\mu\ ^0\ _1\}\{_0\ ^\mu\ _1\} \\ &= a\ddot{a} - \dot{a}^2 + \dot{a}^2 = 0 \ . \end{aligned} \tag{7.93}$$

Similarly $R^0_{202} = R^0_{303} = 0$. Now

$$\begin{aligned} R^1_{212} &= \{_2\ ^1\ _2\}_{,1} - \{_1\ ^1\ _2\}_{,2} + \{_\mu\ ^1\ _1\}\{_2\ ^\mu\ _2\} - \{_\mu\ ^1\ _2\}\{_1\ ^\mu\ _2\} \\ &= -ff'' - f'^2 + \dot{a}^2 f^2 + f'^2 = a_0^2 f^2 \ , \end{aligned} \tag{7.94}$$

by virtue of Eq.(7.54). Similarly it is easy to obtain that

$$R^1_{313} = R^1_{212}\sin^2\theta = a_0^2 f^2 \sin^2\theta \ . \tag{7.95}$$

$$\begin{aligned} R^2_{323} &= \{_3\ ^2\ _3\}_{,2} - \{_2\ ^2\ _3\}_{,3} + \{_0\ ^2\ _2\}\{_3\ ^0\ _3\} + \{_1\ ^2\ _2\}\{_3\ ^1\ _3\} \\ &\quad - \{_3\ ^2\ _3\}\{_2\ ^3\ _3\} \\ &= \sin^2\theta - \cos^2\theta + \dot{a}^2 f^2 \sin^2\theta - f'^2\sin^2\theta + \cos^2\theta \\ &= \left(1 + a_0^2 f^2 - f'^2\right)\sin^2\theta \ . \end{aligned} \tag{7.96}$$

The off-diagonal elements all turn out to be zero. Some more easily while some involve many cancellations. For example

$$\begin{aligned} R^1_{323} &= \{_3\ ^1\ _3\}_{,2} - \{_2\ ^1\ _3\}_{,3} + \{_\mu\ ^1\ _2\}\{_3\ ^\mu\ _3\} - \{_\mu\ ^1\ _3\}\{_2\ ^\mu\ _3\} \\ &= -2ff'\sin\theta\cos\theta + \{_2\ ^1\ _2\}\{_3\ ^2\ _3\} - \{_3\ ^1\ _3\}\{_2\ ^3\ _3\} \\ &= -2ff'\sin\theta\cos\theta + ff'\sin\theta\cos\theta + ff'\sin\theta\cos\theta = 0 \ . \end{aligned} \tag{7.97}$$

Using Eq.(7.54), we can replace $1 - f_k'^2$ by kf_k^2 in Eq.(7.96). Now

$$\begin{aligned} R_2 &= 4[\left(R^{01}\ _{01}\right)^2 + \left(R^{02}\ _{02}\right)^2 + \left(R^{03}\ _{03}\right)^2 + \left(R^{12}\ _{12}\right)^2 \\ &\quad + \left(R^{13}\ _{13}\right)^2 + \left(R^{23}\ _{23}\right)^2] \\ &= 4[(g^{22})^2(R^1_{212})^2 + (g^{33})^2\left(R^{13}\ _{13}\right)^2 + (g^{33})^2\left(R^{23}\ _{23}\right)^2] \ . \end{aligned} \tag{7.98}$$

Putting in the values of g^{22} and g^{33} we get

$$\begin{aligned} R_2 &= \left(4/t^4\right)\left[2 + \left(1 + a_0^2/c^2\right)^2\right] \\ &= \frac{4(3 + 2a_0^2c^2 + a_0^4c^4)}{t^4} \ . \end{aligned} \tag{7.99}$$

Hence, as $t \to 0$, $R_2 \to \infty$ in any case, even when $R = 0$. Thus we see that there is a genuine, crushing, curvature singularity at $t = 0$ in all Friedmann models. At this singularity there is an infinite energy density and radiation would go to infinity faster than matter, in a mixed model, as there would be a single $a(t)$ for the whole Universe but $\rho_{rad}(t) \sim a^{-4}(t)$ while $\rho_{matt}(t) \sim a^{-3}(t)$. Thus, the ratios of the densities at times t_1 and t_2 would be

$$\left.\begin{aligned} \rho_{rad}(t_1)/\rho_{rad}(t_2) &= a^4(t_2)/a^4(t_1) \ , \\ \rho_{matt}(t_1)/\rho_{matt}(t_2) &= a^3(t_2)/a^3(t_1) \ . \end{aligned}\right\} \tag{7.100}$$

A pure dust model has to surpass nuclear density at a sufficiently early epoch in its evolution and the whole Universe would be a giant nucleus. As the Universe expanded the nucleus would fission and there would then be a series of successive fissions which would also spew out all sorts of lighter nuclei but would largely give the heavier elements, especially a substantial amount of primordial iron (the stablest element in that it has least tendency to fuse or fission). Observations of the primordial abundance of elements in the Universe (called "cosmogony") show that the Universe consists of hydrogen and helium (and their isotopes) with only trace impurities of other elements. Apart from lithium all the other elements appear to have been "cooked up" in the nuclear furnaces of the stars and released for general consumption when the star explodes. This occurs once every 30 years in a galaxy, on average, currently. Pure hydrogen and helium stars formed would tend to be blue giants, which have a much shorter life. The clouds of this processed matter would themselves form second generation stars with a longer life. The amount of other elements observed in the Universe must be *all* reprocessed. As such, only hydrogen, helium and lithium (and their isotopes) are primordial. The hydrogen is about three times as much as the helium.

The above observations were used by the proponents of the steady state theories to argue against the "big bang" theory embodied in the Friedmann models. (In fact, the name "big bang" was coined by Fred Hoyle to make fun of the theory.) However, George Gamow provided an eminently satisfactory explanation for the observed cosmogony using the developments in Nuclear Physics which he, among others, had been responsible for.

Gamow pointed out that since nuclear reactions would be involved at sufficiently high densities, the Universe could not have behaved as a pure dust Friedmann model and must be regarded as a mixture of dust and radiation. Since the matter and radiation densities would scale according to Eq.(7.100) near the origin of time, the radiation density must begin to dominate over the matter density at a sufficiently early time. Hence at earlier times the radiation would be greater than the binding energy of the nuclei and there would be fewer nuclei being formed than broken up, despite the high density. If the radiant energy were sufficient, by the time that energy considerations allowed nucleosynthesis the nucleons would be too far apart for it to be likely. Contrariwise, if it were sufficiently low there would be too much nucleosynthesis to fit the observed cosmogony. Thus we can tune the ratio of radiation to matter density at the time of nucleosynthesis.

Gamow's explanation has the advantage of making a testable prediction. Knowing the current matter density, $\rho(t_p) \sim 10^{-29}$gm/cc and the matter density at the time of nucleosynthesis, $\rho(t_n) \sim 4 \times 10^{-5}$gm/cc, required to fit the

observed cosmogony, we can use Eq.(7.100) to work out the ratio of the radiant energy density. By Stefan's law for black body radiation

$$\rho_{rad}(t) = aT^4(t) \ , \tag{7.101}$$

where a is Stefan's constant and $T(t)$ the temperature of the Universe. Thus we see that

$$T(t_p)/T(t_n) = a(t_n)/a(t_p) \ . \tag{7.102}$$

Again, using Eq.(7.100) for obtaining the right hand side of the above equation in terms of the ratio of matter densities, we get

$$T(t_p) = [\rho(t_p)/\rho(t_n)]^{1/3} T(t_n) \ . \tag{7.103}$$

Nucleosynthesis starts with the production of deuterium, which has a binding energy of about 0.11 MeV. Now 1.1 eV corresponds to a temperature of 10^4K. Thus $T(t_n) \approx 10^9$ °K. Hence we see that $T(t_p) \approx 3$°K. Due to errors in the distance estimates (and $\rho(t_n)$) Gamow had actually obtained 6°K as the temperature.

7.9 The Microwave Background Radiation

In 1965 two engineers, Penzias and Wilson trying to develop an extremely low noise super-cooled microwave antenna for Bell Laboratories, found that they were unable to eliminate a background noise beyond their expectations, even when they pointed it at the empty sky. Their antenna was shaped like a horn (see Fig. 7.7). First they thought that the instrument itself may be dirty, but no amount of scrubbing removed the background. They noted that the "noise" corresponded to what would be expected of radiation from a body "heated" to 3.5°K seen in the frequency band of their instrument. When they (luckily) mentioned this problem to the experimental relativist, R.Dicke, he identified it as Gamow's 6°K background radiation rather than instrument noise. Armed with the physical basis for their observations they were able to improve their temperature estimate.

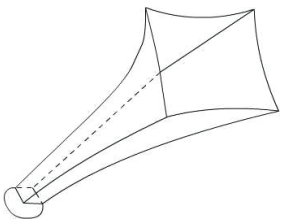


Figure 7.7: Sketch of the Penzias-Wilson "horn" which detected the 2.7°K background microwave radiation depicted roughly. The horn edges have to be flared out substantially to avoid diffraction from them into the horn. The detector has to be super-cooled to be able to detect the very low temperature radiation.

Other researchers used other microwave bands in the windows not absorbed by the atmosphere, or accessed the other bands by sending in balloons or rockets above the atmosphere. All confirmed the Penzias-Wilson observation and the

whole set gave a temperature $T(t_p) \approx (3 \pm 0.3)^\circ$K. By 1980 Woody and Richards had observed the complete Planck spectrum, including the turnover, and found dipole and quadrupole distortions in it. The temperature estimate was, by this time, a lot closer to 2.7°K with somewhat better accuracy than before.

Despite the large uncertainties in the background temperature, it was possible to place much more precise limits on the temperature anisotropy, δT. This was done using two horn antennae pointed in different directions and subtracting the signal received by one horn from the other. This way the systematic errors of such antennae would cancel each other. Reversing the pair of horns (see Fig. 7.8) would cancel out the systematic errors of each horn. Thus with a given angular separation, θ, we could measure the anisotropy in units of the temperature, $\delta T(\theta)/T$. There are technical problems with making θ arbitrarily small. From the early days of anisotropy measurement it was known that at all achievable angles $\delta T/T$ is less than 10^{-3}. In 1982 there was some excitement when Lubin announced a very tentative finding of $\delta T/T \approx 2 \times 10^{-4}$, but this claim was later withdrawn.

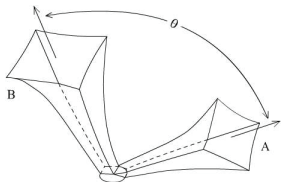


Figure 7.8: A pair of "horns" to measure anisotropy in the microwave background. The signal from one of them (say A) is subtracted directly from B. Thus the precise callibration of the two horns is not required. The pair can be turned round so that A takes the place of B, and B of A, so as to avoid any errors due to the zeroes of the two horn's reading not coinciding. One can vary the angle, θ, between the two horns to get the spectrum of the anisotropy against angular separation. Of course, there are limits to how low θ can be made.

Since the Universe is observed to be inhomogeneous one expects to see the imprint as an anisotropy in the microwave background. Exactly how much this anisotropy should be is not so clear. The reason is that while $\delta T(t)/T(t)$ is a decreasing function of time, $\delta\rho(t)/\rho(t)$ is a very sharply increasing one. Further, since the process of coalescence and "clumping" of matter due to gravitational instabilities is highly non-linear, $\delta\rho(t)/\rho(t)$ can not be computed analytically from some initial $\delta\rho(t_0)/\rho(t_0)$, and any simulation is highly limited and model dependent. However, it becomes very difficult to model the present "clumping" with anisotropy less than about 10^{-5}. As such there was some worry about the lack of anisotropy. The pure Friedmann model was just "too good to be true."

By 1991, observations with the satellite called the COsmic Background Explorer (COBE), which was launched on the 18$^{\text{th}}$ of November 1989, had given a temperature of $T \approx 2.7306\ ^\circ K$. The dipole distortion was explained entirely as being due to our Milky Way galaxy (and indeed the whole local cluster of galaxies) falling towards the Virgo Cluster at a speed of 300 km/sec (600 km/sec). The quadrupole moment claimed earlier turned out to be erroneous. A quadrupole moment *would* be expected on account of very large scale mass density fluctuations, but there is no evidence for this. However, anisotropy at

a somewhat smaller scale which depends on θ *was* seen by COBE at $\sim 10^{-5}$. These observations dramatically confirm Gamow's "Hot Big Bang" model, with the large scale homogeneity depending on the limits on the quadrupole moment in the cosmic background radiation. A far more precise satellite, the Wilkinson Microwave Anisotropy Probe (WMAP), was launched on the 30^{th} of June 2001 and a still more precise one, Planck, on the 14^{th} of May 2009. They not only confirmed the observations of COBE but provided new data that led to a second revolution on the observational aspects of Cosmology, which will be discussed later.

It must be admitted that people like Hoyle and Narlikar still hold to the steady state models. However, most cosmologists class these attempts with those of "saving the aether" and disproving Relativity, or "saving classical physics" and dispensing with Quantum Physics, or even "saving the flat Earth". The consensus is that in essence the Hot Big Bang model is valid. As such it forms the base of what is often referred to as "the standard model", along with the new new revolution mentioned above.

7.10 The Geometry of the Universe

There are many aspects of the evolution of the Universe that are very well understood and tested but there are many more about which precious little is known. Most of the well established parts deal with very early times soon after the Big Bang. Many of the unknown parts relate to relatively recent times. This dichotomy leads laymen to discredit the entire structure of Cosmology. "How can we know about such minute points so long ago" they say, "when we do not know about broad features in relatively recent times?" This lack of credibility is enhanced by the tendency of many practitioners, and particularly those who have entered the arena of Cosmology from High Energy Physics, to present their wild speculations as "Ultimate Truths". I would like to try to clarify this confusion of credibility here.

The first source of confusion has to do with the concepts of space and time. This confusion is enhanced by the misnomer of the standard model — the "Big Bang". This name brings to mind an explosion at some instant and invariably raises the question "What was there *before* the Bang occurred?" This is merely a temporal analogue of a simpler question to deal with "What lies *outside* the Universe?" I will address the simpler question first.

A similar worry bothered the ancients who believed that the Earth was flat. "What lies beyond the edge of the World," they would ask, " and why does the sea not flow off it?" The first problem was resolved by supposing that "the World" is an unbounded surface, either infinite like a hyperboloid of one sheet or a plane, or finite like a sphere or a torus. The requirement that all positions in the World be equally good allows only the hyperboloid, the plane or the sphere. Aristotle, among others, had supposed the Earth to be a sphere as it later proved to be. The second part of the problem was resolved by Al Kindi by postulating a force of terrestrial gravity which pulled objects to the centre of the spherical Earth, to cause them to fall. By providing a totally different concept of "up" and "down" from the universal concepts of Aristotelian Physics, he was able to explain why nothing could "fall *off* the Earth".

Though similar, the above problem is fundamentally different from the corre-

sponding cosmological one, in that we know that there *are* things off the surface of the Earth, the stars above and the ground below. While we directly perceive isotropy along the Earth surface there is no impression of isotropy in all three dimensions. The curved surface of the Earth is embedded, *according to our direct perceptions*, in a flat 3-dimensional Euclidean space. Had we taken the Gaussian view, we would not *look* off the surface of the Earth but keep our gaze firmly fixed on it.

For the resolution of the cosmological problem we take the Gaussian view in 3-dimensions. Again, taking the weak cosmological principle, we must take one of the three analogues of the above "World models" in 3-dimensions. Using an arbitrary $a(t)$ these are the three Einstein-Friedmann models: $k = +1$ for the three-sphere (S^3); $k = 0$ for the flat Euclidean space (E^3); and $k = -1$ for the three-hyperboloid of one sheet (H^3). Though each of these 3-spaces *could* be embedded in a 4-space, in principle, we do not perceive a fourth spacial dimension. (If there were freedom of movement in such an extra dimension we could journey through it and emerge with our left and right sides interchanged. Though such an interchange is never seen in macroscopic phenomena it appears to be seen at the level of elementary particles. This fact has been used to argue that at some, sufficiently small, scale space may have a higher dimensional structure but by the scale of nuclei, $\gtrsim 10^{-13}$ cm, only the 3-space is observable.) From this Gaussian view point *there is no meaning to* "*outside the Universe*". It is only the physically meaningless attempt to regard the 3-space of the Universe as embedded in a 4-space that leads to the misconception of "outside the Universe".

Now let us return to the earlier question of "What was there before the Big Bang?" An unsatisfying, but easy way out is to answer that as with space, so with time, when we reach the origin of time (or the end of time) the concept ceases to make sense and hence the question is meaningless. The vague dissatisfaction with this answer is the feeling that there *is* a difference between space and time. And indeed there is. While there is no preferred location in the Universe we are postulating a preferred time. One may think of the *spatial* Universe as a higher dimensional soccer ball. Taking $a(t)$ to be given by the Einstein-Friedmann models, for $k = +1$ the spacetime Universe resembles more a pointed American Football. This difference does not invalidate the answer given earlier. (The other two model are *infinite* spaces which are unbounded, with the associated difficulty of visualizing something infinite.)

Yet there is more to the layman's uneasiness with the answer. There is a freedom of movement apparent in space which is absent in time. As such we think of a "flow of time". How can that "flow" suddenly stop? The lack of continuity in something that is perceived as necessarily continuous, is what lies at the back of the uneasiness with the answer. Workers in the field get so used to the new ideas that they forget their own earlier uneasiness and sweep the problem under the rug. The present is directly experienced by every observer and hence "exists" in a very obvious way. The past is *recalled* and does not "exist" in the same way. Human memories are reconstructions and not video recordings. The future is *guessed, assuming that nothing unexpected occurs.* It is very difficult to argue that the future "exists" in the same sense as the past, leave alone the present. While others give their own take on it without addressing this issue, I would like to suggest a different resolution of the problem which also addresses the other aspects of the credibility of cosmological models in which

much is claimed about our knowledge of a small portion of the distant part with little understanding of the bulk of the later history of the Universe.

Before Galileo, the concept of time was not very clear as there was no adequately precise measurement of it. The advent of the pendulum clock led to faith in the physically measured time and scepticism regarding the psychologically perceived time. The refinement of the clocks to spring, electrical, electronic and atomic left little room for doubt about the validity of the physically measured time. It is this, proper time, concept that is at the base of Relativity. The tremendous success of this view of time has tended to obscure the element of validity of psychologically perceived time which is needed to clarify the cause of uneasiness mentioned earlier. Let me explain further in my own (round about) way.

Imagine a science built strictly around the concept of time as measured by a pendulum clock. Since the period depends inversely on the square root of the acceleration due to gravity, g, and $g = GM/r^2$, where M is the mass of the Earth and r the distance of the observation point from its centre (so that $r = R + h$, where R is the Earth's radius and h the height or altitude of the observation point), we would conclude that *time slows down with increasing altitude*. Further, at a point where the acceleration due to Earth's gravity is exactly cancelled by the Moon's gravity, say, *time stops*! However, events would continue to occur and would do so in some chronology. The fact that the measurement procedure breaks down cannot mean that "time stops" but only that the procedure is inadequate for the purpose. A re-scaling of time which preserves the chronology and avoids the break-down is required. If the atomic clock ceases to give a good measure of the chronology of time — why, throw it away and acquire a new time-piece that gives a good measure of it.

A good measure would be one which separated off the various events adequately. If many things happened in what we now call a "very short space of time" and little over a relatively long period, we can redefine the measure of time. For example, if a total time interval is T and much happens near $t = 0$ and $t = T$ but little around $t = T/2$, we can define a new time parameter $\bar{t} = \tan\{\pi(\frac{t}{T} - \frac{1}{2})\}$, which ranges from $-\infty$ to $+\infty$. What had appeared as initial and end-points in the t-time are removed to infinity in the $\bar{t}$-time. Similarly, if $t \in [0, \infty)$, we can define $\tau = \log(t/T)$ for some constant measure of time T and have the τ-time go from $-\infty$ to $+\infty$. The point I would like to stress is that there is nothing sacred about our usual time measure. It is convenient for the Physics involved in most phenomena we are interested in for the present epoch. However, if it is inconvenient to describe the chronology of the Universe — why, throw it away. Since the psychological perception is based on how much happens in a given physically measured time interval that concerns (or interests) us, it is a measure of *subjective chronology. This* is its validity and limitation.

In terms of τ (or $\bar{t}$) the Universe has no beginning and no end. It provides a much better description of the chronology of the Universe (as we shall see later) than t does. Though it is not so useful for dealing with usual "macroscopic Physics" (which t does much better) it is well-suited to describe "large scale Physics". Thus the answer to the worry "What happened before the Big Bang?" would be that there never was a "before the Big Bang". In fact the "bang" is a fiction arising from the use of t-time in an inappropriate domain. All it really says is that the usual time measure breaks down for a given value of the usual time parameter. Why can we not use the same argument to say that there is

no meaning to any time before yesterday (or the day I was born, or the year of Hijra or of Christ)? Could we not set $t = 0$ arbitrarily and re-scale the time? The answer is that the new time parameter would *not* provide a good and complete measure of the chronology of the Universe unless the zero of time was set consistent with the Einstein-Friedmann-Gamow model requirements.

7.11 Digression Into High Energy Physics

Starting with Gamow microphysics entered into Cosmology. Many of the recent developments in this field can only be stated in the framework of High Energy Physics. Currently Astrophysics and Cosmology place stringent limits on the various speculations in High Energy Physics and many of those speculations are appealed to by cosmological models. The newly emerged (and very fashionable) subject of Astroparticle Physics is one of the areas of most vigorous research. Regardless of whether much (or even any) of this speculation survives the test of time, it is essential to study it to be able to follow current discussions of Cosmology. As such I will present an extremely condensed review of the relevant aspects of the subject of High Energy Physics.

Matter is composed of combinations of *atoms* ($\sim 10^{-8}$ cm across) which consist of *nuclei* ($\sim 10^{-12}$ cm across) and *electrons* in a diffuse cloud about the nuclei. Nuclei, themselves, are composed of positively charged *protons* and neutral *neutrons* (collectively called *nucleons*). Protons seem perfectly stable but neutrons are only stable in the nucleus. Free neutrons have a half life of $\sim 10^3$ sec decaying by the reaction $n \to p + e^- + \bar{\nu}_e$, to a proton an electron and an *anti-neutrino* (the anti-particle of a *neutrino* which is a neutral particle with very low mass). Nucleons are constituted of 3 *quarks* each. Quarks have never been seen free in nature but only in triplet or quark anti-quark pairs. This fact is called *quark confinement* and any acceptable theory must contain it as a feature. The quarks constituting the nucleons are labelled "up" and "down" (with no physical significance of spatial directions), u and d, with charges $+2/3$ and $-1/3$ respectively. Thus $p = uud$ and $n = udd$. The d quark decays by $d \to u + e^- + \bar{\nu}_e$, giving the observed neutron decay.

In High Energy Physics the basic unit used for energy is that required to move an electron through one volt, called an electron volt or eV. Further units are defined by the thousands: 1 ke$V$=10^3 eV ('k' for 'kilo'); 1 MeV= 10^3 keV ('M' for 'mega'); 1 GeV= 10^3 MeV ('G' for'giga'); and 1 TeV= 10^3 GeV ('T' for 'tera'). These commonly apply: eV for chemical processes; keV for atomic ionization; MeV for nuclear processes; and GeV or TeV for High Energy Physics. Rest masses of particles are generally expressed as energy equivalents. Thus the electron mass is $m_e \approx 7.2$ eV/c^2, a proton $m_p \approx 931$ MeV/c^2, a neutron $m_n \approx 931$ MeV/c^2. The neutrino has a small mass $m_{\nu_e} < 0.09$ eV/c^2. Since quarks do not occur free in nature, their masses are not well-defined. They can be defined in many ways of which the closest to the usual concept of mass of a particle is its *constituent masses* are $m_u, m_d \approx 300$ MeV/c^2. That the sum of three quark masses is *less than* a nucleon mass may seem strange. Components of a nucleus have a *greater* sum of rest-masses, the difference going into the binding energy, in the case of light nuclei. However, since the quarks seem bound far away from each other and free close together (a phenomenon called *asymptotic freedom*) the binding energy behaves differently for them.

You may have noticed that up to now I have steered clear of explaining Quantum Theory to any extent. The reason was that I did not want to lose the mathematicians among my readership, who may not have studied the subject at all, and consequently may shy away from its very mention. Now, however, it has become imperative to introduce the subject, if only minimally. I promise that I will try to make the introduction as "gentle" as possible. The theory is based on Classical Mechanics and replaces the canonical variables, position and momentum, by operators. Thus, for any object the value of the quantity has no meaning till it is determined. If q^i is taken as directly measured, the momentum is given by $p_i = -\iota\hbar\partial/\partial q^i$. Similarly the energy is given by $E = \iota\hbar\partial/\partial t$. Since the operators do not commute, the value will depend on which quantity is measured first. This leads to the Heisenberg uncertainty relations, that the product of uncertainties in the value of conjugate variables is $\gtrsim \hbar$. Thus, $\Delta q_x \Delta p_x \gtrsim \hbar$. Using the angle as the position variable we see that the angular momentum must come in multiples of $\hbar$, as we get back to the same angle on going through 2π. Clearly, the conserved angular momentum is due to an $SO(3)$ symmetry.

Schrödinger wrote the energy conservation equation using the corresponding operators. This is called the Schrödinger equation. The quantity it operates on, representing the quantum object, is called the "wave-function". It turns out to represent the probability amplitude of finding the object at any place (if we use the spacial representation) or with a given momentum (if we use the momentum representation). The wave function must obviously be complex in general. The magnitude squared of the probability amplitude gives the probability. Schrödinger used the classical energy conservation equation to get $[-\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{x})]\psi(t,\mathbf{x}) = \iota\hbar\frac{\partial}{\partial t}\psi(t,\mathbf{x})$. To make it relativistic, we obviously need to take $(\Box - \frac{m^2c^2}{\hbar^2})\psi(t,\mathbf{x}) = 0$, which is the Klein-Gordon equation, where '$\Box$' is the D'Alembertian. This equation runs into problems with the interpretation of the wave function on account of the quadratic time derivative in the $\Box$. To avoid this Dirac tried to convert the equation to first order form. In the process, he had to replace the scalar wave function by a 4-d complex "vector", called a "spinor". (In the next chapter we will deal with two-component spinors.) Further, one needed to introduce what are now called "gamma matrices". These involve the full Lorentz group, and the proper Lorentz invariance of SR leads to a further conservation law. It turns out that this has invariance under 4π rotations, and so the conserved quantity comes in multiples of $\hbar/2$ instead of $\hbar$. This is called "spin angular momentum" and can add to the usual angular momentum. It is a fundamental property of particles.

Particles whose spin is an odd integer multiple of $\frac{1}{2}\hbar$ are called *Fermi-Dirac* particles or *fermions* and those with even integer multiples are called *Bose-Einstein* particles or *bosons* and have very different properties in many ways. For the former the *Pauli exclusion principle* does not allow two of them to belong to the same quantum state (i.e. have all the same quantum numbers in a quantum system) but is silent for the latter. The total angular momentum is conserved in any elementary particle process. Nucleons, electrons, neutrons and quarks are all fermions and have spin $\frac{1}{2}\hbar$. The photon has spin $\hbar$ and is a boson. Of particular interest is the component of the spin along the direction of motion. For the above mentioned fermions it is $+\frac{1}{2}\hbar$ or $-\frac{1}{2}\hbar$. Also, for massless particles it can have only two values, its intrinsic spin and the negative of its intrinsic

spin. Thus, a photon can have helicity $+\hbar$ or $-\hbar$. This provides a "handedness" to particles. Thus we have seen "left" and "right" handed quarks and electrons and photons. However, we have never seen any right-handed neutrinos but only left-handed neutrinos. The significance of this fact will be discussed a bit later. Another quantized property is the behaviour under spatial reflection called *parity*. If the description of the particle is reversed it is called negative parity and if it remains unchanged it is called positive parity. No elementary particles fall in between. The total parity was believed to be conserved in any elementary' process, but we now know that it is not true.

To every particle there is an *anti-particle* with the same mass, spin and parity but opposite charge. The heavier particles formed by quarks are called *hadrons*, the fermions composed of 3 quarks being *baryons* (from the Greek "bary" for heavy) and the bosons composed of a quark anti-quark pair called *mesons* (which were thought to act like a glue to hold the nucleons together despite having so many positively charged particles in it). The lighter particles were called *leptons* (from the Greek "leptos" for light). Earlier the photon was regarded as a lepton but it is now considered differently. There were three spin zero mesons known originally, $\pi^+ = u\bar{d}$, $\pi^- = \bar{u}d$ and $\pi^0 = \frac{1}{\sqrt{2}}(u\bar{u} + d\bar{d})$. Later, three spin 1 mesons analogous to these *pions* were discovered, $\rho^{\pm}$ and ρ^0. There were also two more zero charge spin 1 particles, the ω and φ found which will be explained later. Corresponding to the nucleons there were 4 spin $(3/2)\hbar$ particles: $\Delta^{++} = uuu$, $\Delta^{+} = uud$, $\Delta^{0} = udd$ and $\Delta^{-} = ddd$. The leptons were the electrons and the neutrino. Baryons are given a "baryonic number" $+1$ which other particles do not have. Their anti-particles are given a baryonic number -1. The baryonic number in any reaction is conserved. Since 3 quarks make up a baryon the quarks have baryonic number 1/3. Clearly mesons end up with 0 baryonic number as required. Similarly, leptonic number is also conserved in any reaction.

In the quantum view forces are described by the exchange of "virtual" bosons. (A *virtual* particle is one whose entire existence is as a *quantum state* with no interaction with the *classical* measuring apparatus. As such, by definition, it is never observed.) Thus electromagnatism is described by saying it is propagated by virtual photons. The photon is described by a 4-vector potential which is invariant under gauge transformations (changing the phase of the photon wave function) that generates the Abelian group, $U(1)$. It is called a *gauge boson.* It was found that the weak interactions could be described by the non-Abelian group $SU(2)$, which has 3 $(= 2^2 - 1)$ generators corresponding to 3 spin 1 gauge bosons, $W^{\pm}$ and Z^0. A triumph for this view was achieved when the *virtual* particles were realized in the laboratory. They are very massive — $m_W \approx 80$ GeV/c^2, $m_Z \approx 91$ GeV/c ^{2}.

It may have been noticed that we talked glibly of the proton being composed of 3 fermions, two of which were identical. By the Pauli exclusion principle this should not be possible. This problem was resolved by postulating another quantum number for the quarks, so that the two "u" quarks need not be identical. To ensure that even if all 3 quarks are of the same type (up, down, etc.) they are not identical, the new quantum number must take one of 3 possible values. Like the "up" and "down" labels were provided that were familiar, but without their usual significance. In this case they were called 3 *colours*: "red"; "green"; and "blue" (presumably because those lecturing about these particles used coloured

chalks). Thus the $\Delta^{++} = u_r u_g u_b$ and the $p = (u_r u_g d_b + u_r u_b d_g + u_g u_b d_r)/\sqrt{3}$. The non-Abelian gauge group for the "colour" quantum number is $SU(3)$, which has $3^2 - 1 = 8$ generators corresponding to 8 spin 1 gauge bosons called *gluons*, which bind the quarks together. It is believed that gluons are massless. The theory for strong interactions (of quarks) is called *quantum chromo-dynamics* (QCD) from the Greek word for colour, "chrome", and along the lines of QED (quantum electrodynamics).

There is only one more fundamental force known, namely gravity. So far there is no valid quantum theory of gravity available. One talks glibly of "gravitons" which mediate the gravitational force but there is neither any experimental evidence for it nor a sound theoretical framework in which to place it. (This is one of the most — perhaps *the* most — important problem of fundamental Physics of the day.) There are many views of what the essential problem is, in the attempt to "quantize gravity". (Roger Penrose used to insist on distinguishing the attempt to "quantize gravity" and the attempt to construct a Quantum Theory that is consistent with Relativity. Thus, like him, I would talk of a theory of "Quantum Relativity" and not merely of Quantum Gravity. One point is that the quantum theory is inherently linear while gravity is tied to the geometry of spacetime, via the metric tensor, quantization of the gravitational field entails quantizing the spacetime itself. Thus the basic structure of Physics may need to be modified. Claims have been made for Canonical Quantization, Supergravity, Twistors, Superstrings, Loop Quantum Gravity, Supermembrane (or 'brane') theory, Non-commutative Geometry and Causal Sets (among others) as providing the key to quantum gravity. While each has many minor achievements, which its proponents love to point out, *none* is sufficiently problem-free to provide a viable theory at present. I reserve the attempt of *Quantum Cosmology* for last, not because it shows any better results than the rest, but because it is particularly relevant for this chapter. The point of view taken is that only a *global* approach can take into account the interpretational problems of the Quantum Theory and the non-linearity of GR. Planck had pointed out that if we take units in which $c = G = \hbar = 1$, we cannot then take the classical limit for the Quantum theory or SR or GR. Since there are three "dimensions", L, T, M, we will get dimensional quantities, by putting the three together in different combinations. Most attempts take it for granted that unification will occur at the Planck scale (10^{19} GeV $= \sqrt{\hbar c^3/G}$, the Planck energy or 10^{-33} cm $= \sqrt{\hbar G/c^3}$ the Planck length). All that is clear is that unification should be required *by* the Planck scale, not necessarily *at* it.

Attempts were made to "unify" the fundamental forces of nature. The earliest unification is Newton's *universal* gravitation, combining terrestrial and celestial gravitation. The next is Maxwell's *electromagnetism* (from electricity and magnetism). Einstein tried to unify gravity and electromagnetism without success. Of particular significance for our later purposes, was the attempt of Kaluza and Klein in the same direction (with not much greater success). Glashow, Salam and Weinberg succeeded in unifying electromagnetic and weak nuclear forces into a single *electro-weak* force. The idea may be understood by analogy with electromagnetism. From a static viewpoint the electric and magnetic forces appear distinct. However, for moving charges there are magnetic effects and electric effects of moving magnetic. This "mixing" may be regarded as the hallmark of unification. If we can see electromagnetic effects on weakly interacting

neutral particles and weak interaction effects added into the electromagnetic effect, we would see the evidence of a unified force. The idea was that at sufficiently high energies this signature of unification would be apparent but below the critical energy it would only appear indirectly in virtual processes. The virtual process effect was seen in the 1970s and the proposers were awarded a Nobel prize for this unification in 1979. Direct verification of the theory was achieved later by Carlo Rubia, who observed the $W^{\pm}$ and Z^0. According to the theory, at energies $\gtrsim 100$ GeV the forces are described by the non-Abelian gauge group $SU_W(2) \otimes U_Y(1)$, the former component giving the weak and the latter a force not seen at lower energies. This symmetry is broken at the critical energy and an effective electromagnetic $U_{em}(1)$ remains. The gauge bosons then acquire a mass. The mechanism for this symmetry to break spontaneously, proposed by Peter Higgs, requires the existence of an additional scalar field. The associated Higgs particles have no non-zero quantum numbers and are consequently very difficult to see. It was finally seen in 2011 and the observations confirmed by 2013. Its mass is $m_H = (125.1 \pm 0.3)$ MeV and its mean lifetime is $\sim 10^{-22}$ sec.

It was found that that there are 3 copies of the quarks and leptons with different masses seen so far. They are called 3 *generations*. Thus, as we had (ν_e, e^-, d, u) with charges $(0, -3/3, -1/3, 2/3)$, we also have (ν_μ, μ, s, c) and (ν_τ, τ, b, t), where "s" is the "strange", "c" is the "charm", "b" the "bottom" and "t" the "top" quark. (The names have no special significance and are simply identifying labels.) The μ and τ leptons have associated μ and τ neutrinos. The masses of the new particles are roughly as follows: $m_\mu \approx 106$ MeV/c^2, $m_\tau \approx 1.777$ GeV/c^2, $m_s \approx 486$ MeV/c^2, $m_c \approx 1.55$ GeV/c^2, $m_b \approx 5$ GeV/c^2 and $m_t \approx 177$ GeV/c^2. The six types of quarks are called *flavors*. Quarks of the same, or lighter, generation can decay into each other through strong interactions and involve leptons through weak interactions. The decay of quarks or leptons across generations is severely inhibited. Thus we can have, for example, $t \to b + \tau^+ + \nu_\tau$ or $b \to c + e^- + \bar{\nu}_e$, but the latter decay will be very strongly inhibited as it crosses 3 generations. These are weak interactions; $t \to d + \pi^+ + \pi^0$ is an inhibited strong reaction due to generation crossing. All of high energy physics at present can be understood in terms of the "standard model" consisting of the above "fundamental particles" and the gauge group $SU_C(3) \otimes [SU_W(2) \otimes U_Y(1)]$ which breaks down to $SU_C(3) \otimes U_{em}(1)$ at lower energies.

7.12 Attempts at Further Unification

Despite strenuous efforts there is, as yet, no really convincing evidence of any need to go beyond this standard model. However, the thrill of unification has led to searches for unification of the three forces (embodied in the three groups above) into a single gauge Grand Unified Theory (GUT). The earliest attempt by Pati and Salam tried $SU_C(4) \otimes SU_f(4)$ as a direct product of an $SU(4)$ non-Abelian gauge (colour) group and an $SU(4)$ non-Abelian classification (flavour group). At the time only 4 of the 6 flavours were known, there was no real comprehension of generations and the proposal put leptons as the fourth colour. (With hindsight the idea seems extremely naive.) An immediate implication was that quarks should be able to decay entirely into leptons, thus violating baryon and lepton conservation. Since the proton is composed of quarks, it too should

decay into leptons. The probability of this occurrence would depend inversely on the unification energy. Thus, for a sufficiently high unification energy the decay probability would be so low that the proton would appear to be stable. Early attempts to see how the strengths of the various forces change with energy led, on a fairly wild extrapolation, to a unification energy of $\sim 10^{12} - 10^{13}$ GeV/c^2.

Later attempts at GUTs used a single "simple" group (as apposed to the direct product of Pati and Salam). Thus, while they had two coupling strengths the new attempts had only one. The smallest available such unifying group is $SU(5)$ with $5^2 - 1 = 24$ generators (giving the gauge bosons) and the particles broken up into one set of 10 particles and another set of 5 particles. These were the 3 colours of "left" and "right-handed" up and down quarks, the "left" and "right-handed" electron and the "left-handed" neutrino (which has no observed "right-handed" partner). There is no simple way of choosing the break-up, which is consequently *ad hoc*. A larger group $SO(10)$ has $10 \times 9/2 = 45$ generators. The "spinor" of this group has $2^{(\frac{10}{2}-1)} = 16$ components. These represent the previous 15 fermions and a new, right-handed, neutrino. This is a more elegant attempt at unification but requires many more gauge bosons and predicts an unobserved right-handed neutrino, called a "sterile neutrino". Claims that there is an indication that it is there keep cropping up every once in a while. The theory is that the symmetry broke down at some time when the temperature of the Universe was very high and the right-handed neutrino acquired a mass corresponding to that energy. Thus we should expect to see this neutrino only when we reach those high energies. By now the unification energy had risen to 10^{14} and then 10^{15} GeV/c^2. The proton life-time is inversely related to the unification energy. Thus it started at $\sim 10^{29}$ years and went on rising to $\gtrsim 10^{30}$ years. Many new particles are predicted by GUTs. Thus, when extra matter is needed for our cosmological models, not available in the usual well-known form, the exotic GUT particles are often invoked. This led to a popular slogan of the 1980s, "Cosmology needs GUTs".

Repeated attempts to observe proton decay have failed. The method used is to observe a sufficiently large number of nucleons for a sufficient period to expect many decays and look for the decay particles to emerge. The material used does not matter much as neutrons would also decay to non-baryonic particles through the same decay process, giving correspondingly modified decay products. The present limit of the proton half-life is well over 10^{31} years. The observational limit is at about 10^{34} years on account of contamination by neutrinos produced as a result of cosmic ray showers impinging on the Earth's atmosphere. This provides a background "noise" which would swamp the signal from a genuine proton decay event. To obtain better limits an experiment would have to be mounted on the Moon under some layers of rock. Despite the fact that there is no observational evidence to support GUTs, and in fact the evidence there is seems to refute GUTs, people still believe in it — with a modification coming from another unification attempt.

Wess and Zumino, followed by Salam and Strathdee, tried a totally different type of unification — namely of fermions and bosons, as different aspects of the same fundamental entity related through *supersymmetry* (as it was called). In it, to every particle there corresponds at least one super-partner. If there are n super-partners it is called an *n-extended* supersymmetry. Again there can be many groups and a particularly popular one was the 8-extended $SU(8)$. Here I

do not intend to go into the details of the theory of supersymmetry (abbreviated to $SUSY$) but will just concentrate on the extra particles required by $SUSY$. Thus for a photon there will be the spin 1/2 *photino* and for the neutrino the spin 0 *sneutrino* as super-partners. Similarly, for the graviton (the postulated spin 2 gauge boson for gravity) there will be the spin 3/2 *gravitino*. Even though the original particles (like photons and gravitons) are massless, their super-partners would be massive. Other than the gravitini, the super-partners are expected to have masses in the TeV/c^2 (10^3 GeV/c^2) range.

There is no evidence of any super-partners for observed particles so far. However the proponents of $SUSY$ do point to one piece of observational evidence. It turns out that if we extrapolate the coupling strengths for electromagnetic, strong and weak nuclear forces as functions of the energy they do not meet at one point, but rather pairs of coupling constants coincide at different energies. Nor is the observational error sufficiently large to allow for the possibility of a single grand unification energy. However, the claim is that with $SUSY$ we can get a single unification energy $\gtrsim 10^{17}$ GeV. Though this is an interesting observation, its role as *physical evidence* seems dubious to me. There is no mechanism of spontaneous $SUSY$ symmetry breaking in the offing and the Large Hadron Collider (LHC) in Geneva has passed the energy at which the signals should have been seen. Now, various parameters called "Yukawa couplings" are invoked to maintain the possibility of $SUSY$ being valid. However, appealing to extra parameters violates the spirit of a scientific theory that can make definite, testable, predictions. As someone put it, "At present $SUSY$ is in the ICU on life-support".

The first serious attempt to incorporate gravity into the "Particle Physics" scheme appealed to supersymmetry. Gauge theories may be (only) global or (more strongly) local. The difference can be explained by a simple analogy. An egg shape spun through the centre, but not about the axis of symmetry, returns to an indistinguishable state after a rotation through 2π. Such a symmetry is *global*. However, a round ball rotated about its centre, with any axis, is indistinguishable from its initial configuration for all angles of rotation. This is a *local* gauge symmetry. Thus the conservation of baryonic or leptonic number corresponds to a global gauge theory ($U(1)$ in fact) while the electromagnetic force arises as a local gauge theory (again a $U(1)$). A theory of local supersymmetry is a theory of *supergravity* ($SUGRA$) as it combines internal symmetries with spacetime symmetries, thus bringing gravity into the picture. As mentioned earlier, $SU(8)$, extended to have 8 gravitini corresponding to the one graviton, was the most successful and popular $SUGRA$. Reduced to an ordinary $SUGRA$ it required 10 spacial and 1 time dimension. The suggestion was that as Kaluza and Klein had tried unifying electromagnetism with gravity by introducing a fifth dimension, here we could obtain *super-unification* (of all forces, bosons and fermions, internal and external symmetries) with eleven dimensions. Klein had proposed that the Universe may be "large", or even infinite, in the usual four dimensions but so "small" in the fifth that it could not be observed. This idea, with 8 extra "compact" dimensions was proposed for modern $SUGRA$. The main strength claimed for the theory was that it avoided the unmanageable infinities of attempts to deal with a quantum theory of gravity. However, it was found that it ran into these infinities, in a still worse form, at further steps in the calculations and hence had to be abandoned in its original form.

Another proposal had been floating around to explain the odd behaviour of

quarks, of asymptotic freedom (that they seem not to interact when close together) and confinement (that they attract each other so strongly at a distance that they cannot be pulled apart, and hence are not found free in nature). The idea was that the fundamental entities should not be thought of as particles but instead as quantized excitations of a string. The problem with this theory was that it was inconsistent unless one went to 26 spatial dimensions. Further, the theory was to be used essentially for quarks, which are fermions, but was meaningful only for bosons. It avoided some of the problems for boson excitations but had an unwanted spin 2 massless excitation as well. (There was a super-luminal, or *tachyonic*, non-local field as well.) With the advent of $SUSY$, Green and Scherk tried to develop a supersymmetric string theory. Since the time of $SUGRA$ the extra dimensions no longer worried people and the spin 2 field was actually *desirable* as a graviton. They found a consistent theory with $SO(32)$, $SUSY$, in 10 dimensions and claimed it was unique. For a complete theory one would need 10 dimensions for fermions and 26 for bosons. This was loudly proclaimed as a "Theory of Everything" (TOE). It is now known that the theory is by no means unique and one can build 4 dimensional versions of it. This is the theory of Superstrings and is still largely believed to be able to provide the ultimate theory once we can decide which of the 10^{150} (or so) Calabi-Yau manifolds holds the "correct" Superstring theory.

7.13 The Chronology of the Universe

At the logarithmic time $\tau = -10$ (using seconds to measure t), there was a fundamental change in the Physics governing the Universe. Before this time the temperature of the Universe corresponded to an energy of about 100 GeV and so the forces of electromagnetism and weak nuclear were unified as an electro-weak force. When the laws governing a system change we expect a *phase transition*, such as occurs when water boils or freezes, or when magnetic materials change character. Phase transitions play an important role in the evolution of the Universe. It is not known whether this phase transition played a significant role. However, it is the very first phase transition in the Universe about which we have any really reliable information. Analysis of observational cosmological data led us to believe that there are only three generations of light neutral leptons (mass less than 1 MeV/c^2). As mentioned earlier the sum of all left-handed neutrino masses is much less than 1 eV.

At $\tau = -4$ there was another phase transition from a Universe consisting of quarks and gluons to one consisting of nucleons and electrons, positrons, neutrinos and photons. At $\tau \approx 0$ the neutrinos would decouple from the rest of matter and free-stream through it due to the weakness of their interaction. At $\tau \approx 0.5$ the photons and electron-positron pairs, through the reversible reaction $2\gamma \rightleftarrows e^+e^-$, would break down as the ambient temperature would be about 0.51 MeV and that is the electron rest mass in light units. It is these photons that provide the microwave background observed as being 2.725 $^\circ K$ at present. At $\tau \approx 2$, corresponding to a temperature of about 100 keV nucleosynthesis occurs producing deuterium, He^3, He^4 and some small amounts of Li^7. This epoch also leaves the above trace elements which can be used to verify the model. At $\tau \approx 12$ the Universe ceases to be radiation dominated and becomes matter dominated. This changes the time dependence of the scale factor as discussed

earlier.

At $\tau \approx 13$ the free electrons and nuclei combine to form atoms. Due to the removal of free electrons, by about $\tau \approx 14$, the photons decouple from the other forms of matter in that they also free-stream, like the neutrinos. This is consequently called the surface of last scattering. Any structure in the Universe much prior to this epoch will not leave any imprint in the microwave background, but the image of the structure present at this time would be borne by it. This is the structure seen by the anisotropy in the background radiation.

At $\tau \approx 14$ it is expected that galaxy formation started. Other structures may have started somewhat earlier (according to one scenario) or somewhat later (according to another). The reasons for our uncertainty in this regard are that the processes are highly non-linear and we do not know how much matter there is in the Universe, and are consequently unsure of the nature of the bulk of the matter in the Universe. This latter problem is discussed in more detail in the next section. The usual method adopted to tackle the former problem is by computer simulation. However, it is not currently possible to simulate sufficiently large systems to be able to make reliable estimates for the formation of structure.

The present epoch is somewhere between $\tau = 17.5$ and 18. As can be seen, in terms of the τ-parameter the history of the Universe is not compressed into the beginning, with no major large scale development for most of the time. In terms of the t-parameter it *does* seem strange to claim to be so sure of the early part while knowing virtually nothing about most of the history of the Universe, particularly about the recent part. In terms of the τ-time it does not seem so strange. The non-linearities dominate the last bit, from $\tau \approx 14$ to $\tau \approx 17.5$, but most of the history seems clear, back to about $\tau = -4$. Before that time the fog of uncertainty settles again.

Another implication of τ-time is that we do not think of a "bang" at all. Rather, we think of a continuously expanding 3-space, in the sense that distances between points go on increasing. *There is no explosion!* If the Universe corresponds to a flat or open Friedmann model ($k = 0, -1$) this state of affairs will continue. If it is closed ($k = +1$) the expansion will stop and reverse. In terms of the t-time there will be, what Wheeler called, a "Big Crunch". In terms of τ-time the Universe will merely go on endlessly shrinking, in the sense that distances between points will go on reducing.

7.14 The Composition of the Universe

It had already been mentioned that the matter seen in the Universe is largely hydrogen, with some helium and only traces of other elements. Why, then, should one need to discuss the composition of the Universe again? Well, to start with there are photons as well. In fact, to every nucleon in the Universe there are roughly a billion primordial photons. However, photons have no rest-mass, only energy. As such, while admitting that there are mainly photons in the Universe, one can not account for the *bulk* of it by photons. Of course, as was mentioned in section 11 there are also neutrinos. In fact the number of primordial neutrinos should be comparable with the number of primordial photons and they are not massless. Further, if the right-handed neutrinos expected by some GUTs exist, the left-handed variety *must* also be massive. Even apart from GUTs there are

mechanisms that can be expected to endow neutrinos with mass. As such they could play a decisive role in making the Universe closed. Are there other unseen constituents of the Universe?

It is necessary, here, to focus on the observational evidence that not all that exists (in the Universe) shines. The first point noticed was that the spiral arms of our galaxy (and other spiral galaxies) should not be stable against its rotation if the entire mass of the galaxy was visible. The point is that the centrifugal force on a test particle of mass m at a distance r from the axis and rotating with an angular speed ω is $mr\omega^2$, acting outwards. This is maximum in the equatorial plane. The gravitational force at any point in a spherical distribution of matter, depends on the mass in a sphere interior to that point, $M(r)$. Thus it is $-GM(r)m/r^2$. Denoting the observed angular speed by ω_0, the mass required for stability, $M_s(r)$ would be

$$M_s(r) \geq r^3\omega_0^2/G \ . \tag{7.104}$$

Now, in the spiral arms $r\omega_0 \approx 250$ km/sec and distances go up to about 10^{18} km. Hence M_s $(10^{18}\text{km}) \sim 10^{42}$ kg $\approx 5 \times 10^{11} M_\odot$. Thus it seems that most of the galactic matter is dark. It must be admitted that we cannot really be sure about stability all the way out to 10^{18}km.

It is by no means strange that there should be non-luminous matter present in the galaxy. After all the planets, their satellites, the asteroids, the comets and some tenuous interplanetary dust or gas in our own solar system. The question really is "how much should there be?" and "where is it likely to be found?" We know of enormous dust and gas clouds in our galaxy which occlude 90% of our view of our galaxy. There is evidence of inter-galactic matter as well. How can we determine how much there is? This answer may determine whether the Universe is open or closed. As such it is a vital question for Cosmology.

By mapping our galaxy and studying nearby galaxies one could place an upper limit on the amount of matter in large diffuse clouds. It is inadequate to explain the stability of the spiral arms. In order to check the validity of the mass estimated by the apparent stability of the spiral arms an alternative approach was used. To see this argument consider a system of N self-gravitating particles of mass m_a, position r_a and velocity ν_a $(a = 1, \ldots, N)$. The net force on the b^{th} particle, accounting for gravitational and centrifugal forces, is

$$\mathbf{F}_b = -\sum_{a\neq b=1}^{N} \frac{Gm_a m_b(\mathbf{r}_b - \mathbf{r}_a)}{|\mathbf{r}_b - \mathbf{r}_a|^3} + \frac{m_b}{|\mathbf{r}_b|^2}(\mathbf{r}_b \times \mathbf{v}_b) \times \mathbf{v}_b \ . \tag{7.105}$$

On average this force must be zero. Thus, summing up over all the particles, taking all masses equal and using the mean field approximation, we get

$$Nm\bar{v}^2/\bar{r} = N^2Gm^2/\bar{r}^2 \ , \tag{7.106}$$

where $\bar{v}$ and $\bar{r}$ are average speed and distance respectively. Dividing through by $2N/\bar{r}$ we see that

$$K.E = 2P.E \ , \tag{7.107}$$

where $K.E$ is the average kinetic energy and $P.E$ the average potential energy due to gravity. This result is known as the *virial theorem*. (It turns out that this result holds locally in Relativity as well, but that is not relevant for our present purposes.) Observing the random movements of stars away from the

mean movement gives their *velocity dispersions*, $\bar{v}$. The observed velocity dispersions were considerably in excess of expectations based on the gravitational potential due to luminous matter. Further, the component of velocity dispersion perpendicular to the galactic disc are higher than those in the plane of the galactic disc. Thus there must be more matter in the disc than is seen, i.e. *dark matter*.

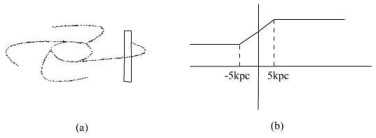


Figure 7.9: Observations of the rotational velocities of galaxies. (a) A sufficiently enlarged picture of a galaxy is scanned by a slit so that only a portion is seen and the light is passed through a spectrometer. This reading gives the red-shift of the light for that portion of the galaxy. (b) The red shifts are plotted against distance of the portion from the galactic centre. Typically, the galaxy would have some red-shift on average. The portions give a relative red or blue shift about the average. It is found that the core corresponds to about 5 kpc for all galaxies.

Another method was developed and extensively used by Vera Rubin (the woman the Nobel Prize forgot) and others [54]. Galaxies were investigated for the red-shift of separate parts. There were the parts moving *away*, due to their rotation, and parts moving *towards* the Earth. Though the galaxies are generally red-shifted on average, one could look at the relative blue and red-shifts of the parts. The procedure adopted was to scan the galaxy through narrow slits and plot the red-shifts against the radial distance from the galactic centre, see Fig. 7.9. It was found that the rotational speed of the stars in the galactic core (about 5 kpc for a surprisingly large variety of galaxies) behaved much as might have been expected on the basis of luminous matter. However, for spiral galaxies the outer regions did *not* rotate as expected. Let us see what "could have been expected".

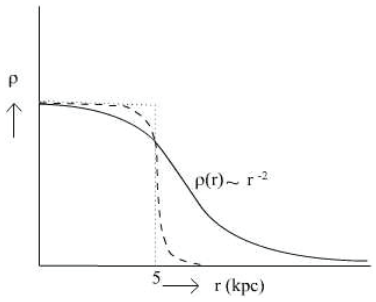


Figure 7.10: Density profiles for the galaxies. The dotted line represents an idealized constant density sphere of radius 5 kpc. The dashed line represents the typical density profile of stars or planets. What is deduced from the rotational velocity curves observed is more like the hard line with a "tail" varying as $1/r^2$.

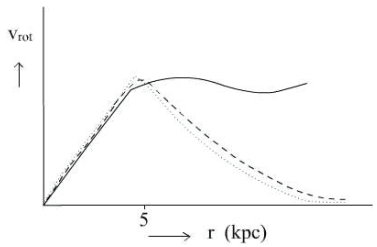


Figure 7.11: The rotational velocity profile for galaxies. The dotted line gives the velocity expected for a sphere of uniform density and 5 kpc radius. The dashed line gives the more realistic estimate with the tail still going as $1/\sqrt{r}$. What is typically seen is as shown by the hard line, with flat rotational velocity curves going out to about 80 kpc.

If we have a pressure-less dust sphere of total mass M, radius a and density $\rho(r)$, Eq.(7.104) gives

$$\omega(r) = \frac{v_{rot}}{r} = \sqrt{\frac{4\pi G \int_0^r \rho(\bar{r})\bar{r}^2 d\bar{r}}{r^3}} \ , \tag{7.108}$$

inside the sphere. If the density is a constant, we get the rotational velocity, v_{rot} given by

$$v_{rot} = \sqrt{4\pi G\rho_0 3}r \tag{7.109}$$

inside the sphere and

$$v_{rot} = \sqrt{4\pi GM}/\sqrt{r} \ , \tag{7.110}$$

outside. Thus the rotational velocity profile expected for a star with the density profile shown by the dotted line in Fig. 7.10 would be as depicted by the dotted line in Fig. 7.11. However, for the more realistic density profile given by the dashed line in Fig. 7.10, the velocity profile should appear close to that displayed by the dashed line in Fig. 7.11, with slight variations. What is seen is like the hard line in Fig. 7.11 with minor modifications from galaxy to galaxy. This means that the core of the galaxy does have a density profile as shown in Fig. 7.10 but the outer regions do not follow that density profile. There the density must be of the form $\rho(r) \approx \rho_0 a^2/r^2$ for $r \geqslant a$, giving the graph displayed as a hard line in Fig. 7.11.

Observational evidence suggests that elliptical galaxies are just like spiral and other galaxies as regards the core. It is not clear whether the density profile of the dark matter is significantly different for them as there is not enough luminous matter far away from the core to trace the matter density there. For the more common spiral galaxies the density seems to decrease as the square of the distance up to about 60 kpc. The ratio of the mass of the core, m_c, to the dark matter *outside* the core, m_d, for a galaxy of core radius a and total radius R, is

$$\frac{m_c}{m_d} = \frac{\int_0^a 4\pi\rho_0 r^2 dr}{\int_a^R 4\pi\rho_0 a^2 dr} = \frac{a}{3(R-a)} \ , \tag{7.111}$$

where we have used the same ρ_0 to ensure the continuity in the density profile apparent in Fig. 7.10. This ratio comes out to be about 3% for the typical

figures mentioned above. Hence most of the matter in the galaxy must be dark and reside in large diffuse "haloes" around the galaxy. There must also be a significant dark component of the galactic disc.

There is little direct evidence telling us of the nature of the dark matter, except that it does not shine. One suggestion is that it is ordinary matter which did not form into sufficiently large clouds to collapse into stars, not having the mass to ignite nuclear fusion, and hence stay dark. These were called *massive compact halo objects* or MACHOs for short. (The significance of this acronym comes from its slang meaning of "he-man" and the fact that another proposal was for *weakly interacting massive particles*, or WIMPS, which is slang for the opposite of manly.) Such objects may collapse to large planet size, Jupiter or a bit larger. In that case they would no longer remain dark but "glow" in the radio and infra-red bands. Wheeler dubbed them "brownies" but the more prosaic "jupiters" has become the established name. The infra-red glow from the haloes of galaxies has been observed! We are unsure of the amount of matter they provide, but they are definitely there.

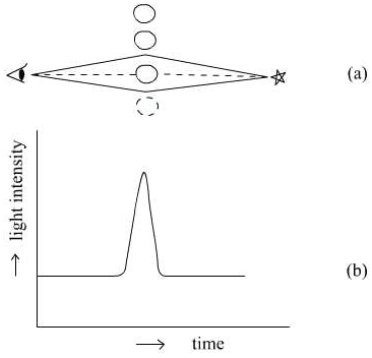


Figure 7.12: Light enhancement from stars due to microlensing by condensed object. (a) The light from a distant star is bent around the dense body providing more paths for the light to reach us and hence more light reaches us. (b) As the object crosses our line of sight the light from the star is enhanced.

If there are MACHOs drifting about they may come into the direct line of sight between us and a star. If the distances involved are appropriate, the light can bend round it due to the gravitational deflection of light and more light reach us than would come directly, see Fig.7.12. If it merely grazes the line of sight the intensity may be only slightly enhanced but if it passes "dead centre" there can be substantial increases in brightness. Instead of occluding the star the MACHO acts as a lens. (The blue shift of the light entering the gravitational field of the MACHO would be exactly cancelled by the red-shift on leaving, so the lens would be *achromatic*.) This *microlensing* has been observed!

There are suggestions for more diffuse MACHOs which never even managed to collapse to planets and various proposals for detecting them. As such there can be much more matter in the form of MACHOs than is seen by the above methods. How much can there be? This is an open question which will be discussed in more detail in this and the next section.

Observing the motion of stars in systems of galaxies — binary, triples, clusters — it is found that the richer the system (i.e. containing more galaxies) the

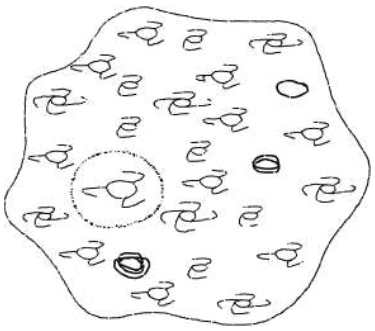


Figure 7.13: (a) Most of the halo matter may not be seen. One can see rotation curves for the main part of the galaxy and even where there are tracer stars available. (b) For binary galaxies much more of the halo matter is seen but much more may not be seen. One can try to deduce the full amount and the nature of its constituents by thermodynamic arguments but the results may not be reliable. (c) For rich clusters of galaxies most of the halo matter would be sees as they would share a common halo.

more dark matterDark matter needs to be postulated. This can be consistently explained with the previous observations by taking R to be much larger than 60 kpc. Since the visible matter is needed to trace the presence of dark matter, one would not "see" the rest of the halo in a single galaxy. However, in a system of galaxies it would be traced by the stars of the other galaxy, see Fig. 7.13. In fact, for a rich cluster of galaxies we could think of one giant halo shared by the entire cluster and then the stars in the outer galaxy could trace nearly all the matter in the halo.

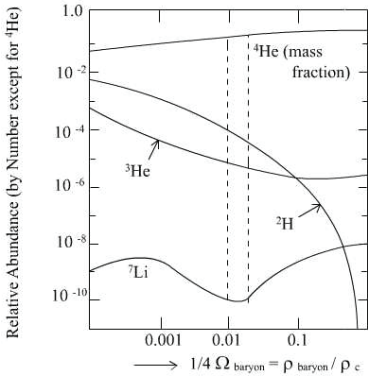


Figure 7.14: The primordial abundance of elements as observed, plotted against the density of baryons (in units where critical density is unity). The total dark matter density required by observations is $\Omega_{dm} \sim 0.24$, while the baryonic dark matter density is only $\Omega_b \sim 0.046$.

We see that there must be a large amount of ordinary dark matter in the Universe. Could there be enough to explain all the dark matter that must reside in the galactic haloes and galactic disk? Could there be enough so that the

Universe is, in fact, closed? There are limits to the amount of "normal" matter due to the constraints of nucleosynthesis. To be precise, if "normal matter" is taken to be composed of atoms, then most of its mass comes from protons — which are baryons. Thus there is a constraint on the *baryonic* dark matter given by the prediction of the amount of each element as shown in Fig. 7.14. The density of baryonic matter turns out to be $\sim 4.6\%$ roughly half of which must be dark.

7.15 Accelerated Expansion of the Universe

Just when we thought we had everything sewn up in Cosmology, Adam Riess, Saul Perlmutter and Brian Schmidt threw the proverbial cat among the pigeons. The Friedmann model gave an expanding Universe, but one that was decelerating. This applied for pure matter and pure radiation and any model in between. In the early days (1931) Georges Lemaître had proposed a model that included the cosmological constant and was otherwise like the Friedmann model. Since the data all seemed to fit with the pure Friedmann model Lemaître's model was forgotten. The three Nobel Laureates, observing high red-shift supernovae, found that that far away one could see that the expansion of the Universe was accelerating. In the beginning it was not clear how much the acceleration seen was. However, with time it crystalized to an exponential expansion. This is exactly what one would expect with a Friedmann-de Sitter, or Lemaître model. As such, one might have thought that this would be the end of the matter — the missing cosmological constant was found and the Universe was Lemaître and. It further appeared that regarding the cosmological constant as a vacuum energy, the total energy of the Universe just fitted a flat model. The current estimate is $\Omega_\Lambda = 72.4\%, \Omega_b = 4.6\%$ and $\Omega_{dm} = 23\%$. What could be better?

However, we physicists are *never* satisfied — especially if it can lead to a flurry of papers. The claim was made that the value of the cosmological constant required by observations was "too small". Assuming that the correct theory for the Universe would be a Quantum Gravity theory, the energy required should be the Planck energy, 10^{19} GeV. This is about 126 orders of magnitude larger than the observed value! (What Hawking called "the worst prediction in the history of Physics".) Such an energy would give would give a life of the Universe $\sim 10^{-40}$ sec. "Since", it was argued, "this is patently absurd the cosmological constant must be zero for some reason." "*Zero!*", you may well ask, "How would that ever help? It should surely make it even worse?" Despite that, a reason was found for not having the cosmological constant. It was claimed that the vacuum would be unstable if there was such a constant. "Hence", they said, "it must be a dynamical field." As they like to call it, "a mysterious form of dark energy". For this purpose they postulate "phantom energy", "quintessence". k-essence, a Chaplygin gas, etc. (You notice the number of papers that can be written for all these candidates?) The papers continue despite the fact that it has been shown that the claims about the instability and the expected value of the cosmological constant are not valid [55].

The current observations fit precisely with the cosmological constant (see Fig. 7.15). Why not simply accept this? As the old joke goes, "If it looks like, and it smells like, and it takes like, then it *is*." Nevertheless, it is insisted that there is a "dynamical field" that exactly mimics the cosmological constant. Is it

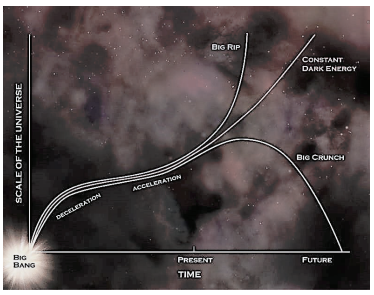


Figure 7.15: The accelerated expansion of the Universe looks *just* as if it is caused by the cosmological constant. Those working on the observational side call it the cosmological constant. and deal with it as vacuum energy. [Picture taken from the web in the open domain.]

supposed to be God's little joke on us — leave it *looking* as if it is a cosmological constant, but *actually it is a dynamical field in disguise?* One can imagine God busting a rib laughing at "how he has confused these little blighters".

7.16 Non-Baryonic Dark Matter

Due to the large uncertainties in the value of the Hubble parameter in the earlier days, the amount of dark matter required to "close the Universe" was not known well. However, all indications were that there was not nearly enough. The dark matter required for the flat rotational velocity curves, including rich clusters of galaxies, seemed under 25% of critical. The rest could be distributed uniformly in the Universe without affecting these curves. This was before the discovery of the vacuum energy. Now the situation has crystalized. All the dark matter needed could be in the galactic halos and disks — and baryonic matter cannot be enough for the purpose on account of nucleosynthesis arguments. The reason for looking for baryonic matter was to stay within the framework of "matter" as we know it. We would not want to propose "the aether" once again. Could there be other forms of energy that we have not included? The simplest is light. However, we know that the current energy density of photons in the CMB is $\Omega_\gamma \sim 5\times10^{-5}$. (It may seem like a non-sequitur to talk of light as "dark matter", but if you think about it, the photons will not be seen in the usual way and are there. the quibble that this is not matter would not hold as we equate matter and energy.) How about neutrinos? It turns out that including all species of neutrinos $10^{-3} < \Omega_\nu < 2\times10{-}2$. Put together they are nowhere *near* adequate for the purpose. So, where do we go from here?

Probably the first suggestion was of "dark stars". However, as the extent of the matter that was needed became clear, Milgrom [56] suggested that it was not that extra matter was present but that Newton's law of gravity failed at the very large. He wrote a series of papers on this idea. The proposal came to be known as "MOdified Newtonian Dynamics" (or MOND). Of course, it runs afoul of GR and so could not be taken so seriously. The earlier hope seems to have been that it could be appropriately tweaked to make it consistent, but that was not the case. Instead one needed to reformulate *ab initio* and obtain something consistent One of the modern approaches is to try to use a modification of

the Einstein-Hilbert Lagrangian, by taking an arbitrary functions of the Ricci scalar in place of the Ricci scalar itself. This is a whole class of non-scientific theories that is, in fact, a mathematical formalism set up to provide proposers with the tools to follow up (hopefully) more scientific proposals. The original "tool-box" [57], called $f(R)$-gravity, comes from well before MOND, but is more cumbersome and is not needed for the specific purpose at hand. The reason I say it is unscientific is that it contains infinitely many parameters and that brings back the epicycles and the aether. It is well to recall Freeman Dyson's aphorism, 'Give me two free-parameters and I can fit an elephant. Give me three and I can make its trunk wiggle". I would add "Give me infinitely many and I can make it jump through the eye of a needle set on fire and held up by demons and fairies!" An extension of GR may be needed but it had better be adequately predictive.

Another suggestion was that there may be primordial black holes ("mini" or otherwise) in sufficient quantities to provide the required dark matter. If black holes formed before the era of nucleosynthesis, they could have been formed of ordinary matter but would still not contribute to the baryonic matter constrained by nucleosynthesis arguments. If Hawking radiation is actual, then the black holes would have to be of a mass more than would be radiating today, as we have not seen exploding black holes. Of course, if it is not (as I have suggested) then this limitation does not apply. Limits on the total mass available in such objects were obtained by the limits on micro-lensing, or image doubling through gravitational lensing, events. However, my own feeling is that though these limits do not exclude the black hole candidature for galactic haloes, the extent of anisotropy required for such a scenario, though it is for very low values of τ, should leave a larger imprint on the microwave background then the observed 2×10^{-5} anisotropy seen. (This is not a quantitatively justified argument but more a statement of a personal hunch.) Better methods to get an answer are bign sought in the large scale structure.

At this stage it is necessary to discuss the spontaneous symmetry breaking mechanism proposed by Higgs. There was a "no-go theorem" that there can be no massive scalar, gauge field as the symmetry would be broken by it. Higgs proposed that the field could be massless while the symmetry is retained and its breaking would provide the field with a mass. Such a field is, thus, called a "Higgs field". As such, the symmetry would suddenly be broken at one specific energy. The idea is that the potential of the field is, say, a quartic function with only one minimum at the field equal to zero. If the coefficients are temperature dependent, at some temperature it could acquire another two extrema: a local maximum; and a local minimum. To start with the second (local) minimum would not be the global minimum, but at some critical temperature it could fall below the previous minimum. At that stage, the theory would lose the symmetry about zero. Massive spinning particles have a helicity (the dot product of the spin with the normalized momentum vector. By changing frame to overtake the particle, the helicity would be seen to flip. However, since we cannot overtake massless particles, they will have a well-defined helicity. One would expect the theory to have both helicities present in general. However, it could be that since the spontaneous symmetry breaking provides chooses one of the two randomly, the particles could acquire either a left-, or a right-handed helicity. This "chirality", or "handedness", is a hallmark of the $SU(2)$ of the standard model and it provides the neutrinos with a left-handed helicity. The neutrinos

would then acquire a very small mass by what is called a "see-saw mechanism" that I will not go into. If there was a GUT that was spontaneously broken at a much higher temperature, it would be in the very early Universe, and the GUT Higgs would provide masses to most of the fermions, with masses of about the GUT-breaking energy. However, there could be some massless particles with their specific symmetry. Thus $SO(10)$ could break to an $SU(4) \times SU_L(2) \times SU_R(2)$, that would than break down to the standard model. This would provide the sterile right-handed neutrino with the "small" mass mentioned (as it is much less than the GUT-breaking energy.)

It is obvious that GUT-breaking provides dark matter candidates. The idea of the very massive sterile neutrinos, $m_{\nu_R} \sim 100$ GeV, led to further speculations of GUT artifacts to provide the required dark matter. Of course, they must be weakly interacting massive particles (WIMPS). (You will notice the juxtaposition of MACHOs and WIMPs.) In particular, one expected the particles to be neutral. Limits could be found for massive neutral leptons produced primordially. It turned out that the sum of the rest-masses of such particles should be less than about 100 eV/c^2 or greater than about 1 GeV/c^2. The tremendous variety of objects within the allowed limits led to it being called a "GUT-breaking zoo". The problem is not one of *finding* a candidate but of *limiting* them. This problem is compounded by a profusion of SUSY artifacts with all sorts of extra exotica from SUSY-GUTs, as GUTs cannot be valid without SUSY. (Just imagine the cornucopia of papers that was opened up by these speculations. I wish we could get back to the old days when physicists sought ultimate truths about the Universe and not exponentially increasing publications. Time was when Biologists used to do experiments with mice running through mazes in the laboratory to determine their intelligence. Now the Scientists seem to have become rats running through the academic corridors to prove their "intelligence" in the rat-race of publications.)

Another reason to consider GUT-breaking for Cosmology came from the question: "Where do all the observed baryons come from?" An obvious answer is that there may be regions of anti-matter so that the net baryonic number of the Universe may be zero. However, there is no evidence to support this view and most available evidence runs counter to this claim. We do not see large chunks of matter and anti-matter annihilating each other. Nor is there apparent any satisfactory way to keep the two adequately far apart that they are not seen to annihilate. In particular, in the rich galactic clusters all matter must be of the same type. The conventional answer was simply "that it was there from the beginning". Since the parallel question of "how the elements came about and happened to be in the observed abundances?" led to the understanding of primordial nucleosynthesis, this answer seems less than satisfactory. Nevertheless, I feel a bit uneasy about the fact that the GUT-breaking answer preceded the question. Also, the GUT explanation of *baryogenesis* (as it is called) relies on an unproved theory and appeals to an undetected instability of protons. Further, there has been no quantitative success of the "explanation" as there was for nucleosynthesis. In fact the attempts have been markedly unsuccessful so far.

It must be admitted that the lack of good candidates of dark matter is one of the most genuinely serious problems faced by modern Cosmology (without reference to the number of papers that may be written. Nothing seems to work and some of us are being driven to desperate speculations like modifying GR,

albeit with limited new parameters introduced.

7.17 Problems of the Standard Cosmological Model

When Particle Theory entered into the cosmological arena, people started looking for problems in the standard model which they could resolve using their latest theories, thus providing a "test" for them, unavailable by using accelerators due to the extremely high energies required. The problem of baryogenesis is one such example. There were already some open problems within the standard model, such as the quantity and nature of dark matter, the mechanism for structure formation in a homogeneous isotropic Universe, etc. What was brought forward now was not this sort of open problem in the model, but problems with the model itself. Many of these problems were artifacts of the Particle theories (mainly GUTs) themselves. However, some problems had already been noted by cosmologists. These problems with the model need to be addressed now.

The GUT induced problems include the "magnetic monopole problem", "the domain wall problem" and other similar "problems". These problems need not worry us unduly. The theories they are based upon are speculative at best. Till they are tested and the details of the mechanisms used understood, it is not clear that there *is* a "problem" in the first place. As mentioned earlier, this procedure of finding "problems" to which we have "answers" seems unsound as a scientific method. Similarly, Superstring attempts at "quantizing gravity" lead to the generation of a cosmological constant of typically Planck scale. This is called the "cosmological constant problem". Again, the theory is totally speculative with no real basis in physicsl observation. There is, in my opinion, no need to explain why the cosmological constant should be zero (or nearly so) instead of Planck scale as we could simply take the fact as evidence against Superstrings.

One of the first serious problems noted was what was called the "causality problem" originally and is currently known as the "horizon problem". Given two points in the Universe, P and Q, at any instant, t_0 (in the background frame used for the Friedmann metric), they will enter into causal contact at some later time, t_2. Though the arguments apply equally to the other models, for concreteness think of the closed Friedmann model, see Fig. 7.16. Of course, they may both be observed at an earlier time, t_1, by an observer R between P and Q, but they can not begin to influence each other till t_2. If we take t_0 to be the time when the microwave radiation was produced and t_1 to be the present time with ourselves at R, it is clear that P and Q could never have been in causal contact. How, then, did they manage to thermalize enough to achieve the same temperature (within an accuracy of nearly 10^{-5})? The fact that the Universe was very small in its early stages is not relevant as the expansion of the Universe pushed neighbouring points away at nearly light speed at that stage. One could introduce a Conformal transformation to remove the re-scaling effect, see Fig. 7.17. In that way of looking at it P and Q were very far from each other.

An associated problem is of *structure formation*. The homogeneous model of the Universe can only be an approximation at best. What we see is structure at different scales. In fact, at more or less any scale up to that of galaxies shows 'clumping', with the density of clumped matter to the average at that scale

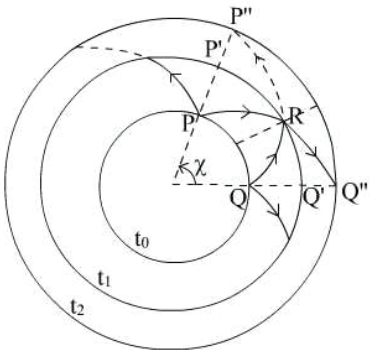


Figure 7.16: Three instants in the history of the expanding Universe: t_0, t_1 and t_2 are represented by three Friedmann spheres. The light cones starting at t_0 at two spatially separated points P and Q "spread out" due to the expansion of the Universe. The "cones" intersect at R on the t_1 slice and P and Q would, thus, be observed at R at time t_1, when the points would be at P' and Q' respectively. An observer at the place Q would see the light sent from P at time t_0, at time t_2 at $Q''$and *vice versa*. Hence causal contact occurs at t_2.

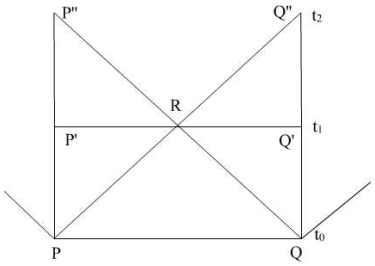


Figure 7.17: The re-scaled conformal picture of any of the Friedmann models. The separation PQ is essentially, now, given by the hyperspherical angle separation, χ, and the radial line have now become vertical lines. Also the light cones now look like usual light cones. P and Q are seen simultaneously for the first time by R at t_1. Causal contact between P and Q occurs at t_2.

$\sim 10^5$. How does this come about while the Universe has CMB anisotropy of $1/10^5$? We tend to assume that "at the largest scale the Universe is homogeneous", but what we see are planets and stars, clusters of stars and galaxies and clusters of galaxies that lie on enormous filaments with great voids between. The smaller scale structure forms due to gravitational nonlinearities, but what provides the *large scale structure*?

The problem of structure formation is compounded by the dark matter problem. The dark matter may *enhance* structure formation or *inhibit* it. This problem was realised in the early days when neutrinos were being considered for structure formation. The fact is that they if they are of an energy to go across a very large collection of matter when they come to it, they may fall into the gravitational potential well before they reach the other side, since the Universe is very small at that stage and so the Universe evolves fast. In any case, they "free-stream" across most of the matter, obliterating structures as

they go, before they fall in — much like water in the tides destroying the structure of a sand castle. If the dark matter particles are more massive they may tend to settle down. However, what we regard as massive depends on the stage of the Universe at which they are forming the structure. At earlier stages the temperature is more and the masses are effectively less.

To be more concrete, we can compare the energy at which the particles are produced with their rest masses. If the energy is much more then the particles will be relativistic. If the energy is barely enough to produce the particles, they will have no surplus energy to move around. Very broadly speaking, "hot dark matter" is likely to destroy structures while "cold dark matter" will create structures. In between temperatures yield "warm dark matter". Notice that mixing hot and cold dark dark matter does not produce warm dark matter. Again, if the particles have large interactions they will thermalise but if they virtually do not interact at all, like neutrinos, the temperature distributions will be "frozen in" at birth and will scale down rather than retaining the Maxwell-Boltzmann distribution. This is likely to be generically the case for WIMPs but not for the other types of objects.

Another point we have to bear in mind is that we can have seed structures from very early on, but the structures must be sufficiently small in size and in the extent of clumping to be contained in the temperature anisotropies of the CMB. The CMB carries information from "the surface of last scattering" about 350,000 years after the Big Bang. Trying to obtain the small and large scale structures seen so that they start at $10^{-5\circ}$K at this time and reach the values seen is a challenging task. This was achieved by the current standard model of Cosmology, the "Λ-CDM Model". (Notice that it is *not* "DECDM", or "Mysterious form of Energy-CDM", etc., just Λ-CDM.)

The standard model depends crucially upon perfect homogeneity, isotropy and spherical symmetry. In a series of papers Khalatnikov and Lifshitz (later joined by Belinski) considered a generic approach to a cosmological singularity [58] (which was later dealt with more rigorously in a series of papers by Penrose and Hawking [59]). They used the homogeneous cosmologies classified by Bianchi (see for example [60]), approximating the metric coefficients by powers of t, the cosmological time from/to the iniitaal/final singularity, using Cartesian coordinates. They found that the metric coefficients for each of the directions diverge with increasing amplitude and frequency as the singularity is approached.

Let us look at it more concretely. The metric tensor is written in terms of a cosmological time, t and a spatial part: $ds^2 = c^2dt^2 - d\sigma(t)^2$. The spatial part is written as: $d\sigma(t)^2 = \gamma_{ij}(t)dx^i dx^j$, where the matrix is taken in diagonal form. Now, writing $\gamma_{ij,0}(t) = y_{ij}$, one can write the Einstein equations is the form $R^\mu_\nu = \kappa T^\mu_\nu - t\delta^\mu_\nu$ and split it into the pure time, space-time and pure space parts. The trace of the spatial part must be a vacuum sufficiently far in the past. This requirement worked through yields $\sqrt{\gamma}y^i_i = 2/T$, whence $\gamma = (t/T)^2$, where $\gamma = det(\gamma^i_j)$. Using x, y, z for the spatial coordinates (*not* Cartesian coordinates), the metric is approximated by

$$ds^2 = c^2dt^2 - (t/T)^{2p_l}dx^2 - (t/T)^{2p_m}dy^2 - (t/T)^{2p_n}dz^2 \ , \tag{7.112}$$

with the condition that

$$p_1 + p_2 + p_3 = 1 \ , \tag{7.113}$$

so that the shape of the Universe is like an ellipsoid at any instant. Requiring that the spatial Ricci tensor also vanish, yields

$$p_1^2 + p_2^2 + p_3^2 = 1 \ , \tag{7.114}$$

which together imply that

$$p_1 p_2 + p_2 p_3 + p_3 p_1 = 0 \ . \tag{7.115}$$

These are called Kasner numbers and they can be written in terms of one parameter, u, as

$$p_1(u) = \frac{-u}{1+u+u^2};\ p_2(u) = \frac{1+u}{1+u+u^2};\ p_3(u) = \frac{u(1+u)}{1+u+u^2} \ . \tag{7.116}$$

The dependence of the three exponents on u is depicted in Fig. 7.18. Writing

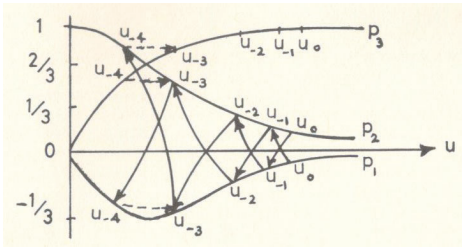


Figure 7.18: The dependence of the three exponents, p_1, p_2, p_3 on u. Because of Eq. (7.114) one of the three must be negative and the other two positive. What started negative goes to zero and on to positive; what started as zero goes to 1; and what started as 1 goes to zero and then negative.

the three coefficients in terms of their logarithms as α, β, γ, we obtain $\alpha+\beta+\gamma = \ln t = \tau$. Each of the previous cycles is one epoch, depicted in Fig. 7.19, and there are infinitely many epochs.

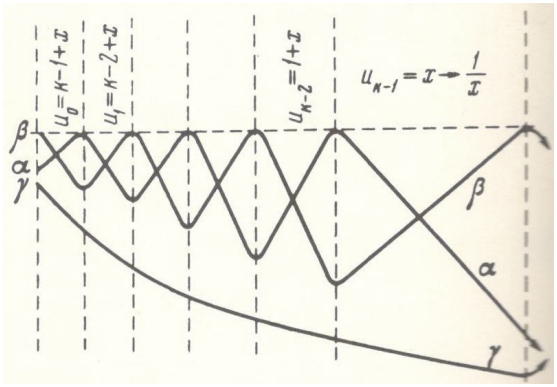


Figure 7.19: The variation of, α, β, γ in one epoch.

The above model was touted as one of the possible solutions of the causality problem, as it would make the Universe homogenous like a liquidiser does. In

fact, Misner [61] called these "mixmaster oscillations" (referring to the brand name of a popular liquidiser at the time). The oscillations, and consequent anisotropies, would rapidly disappear as the Universe expanded. This suggestion is no longer popular because people believe that the oscillations would not have survived long after Planck times since they believe that quantum gravity effects fade past the Planck era. Since the argument has nothing to do with quantum gravity, and the claim that quantum gravity starts at the Planck scale is dubious at best, I do not see that there is sound reason to reject the idea. On the other hand, there is no evidence to support it either. It is interesting to note that the mixmaster oscillations would provide a natural reason to use the logarithmic time as the only relevant time parameter in that era. Atomic clocks would not function when atoms could not exist and even "nuclear clocks" would have no meaning. The only clock that would give a chronology of events would be the "gravitational clock", which would go through an infinite number of ticks.

Another problem is associated with the density of the Universe. For the flat model to be valid it must have, by Eq.(7.56) the critical value

$$\rho_c(t) = \frac{3H^2}{8\pi G} = \frac{3(\dot{a}/a)^2}{8\pi G} = \frac{1}{6\pi G t^2} \ . \tag{7.117}$$

Taking the present age to be about (13.772 ± 0.006) billion years, $\rho_c(t)$ would be about $3.2 \times 10^{-30}\mathrm{g/cm}^3$. Writing the observed density of the Universe as ρ_{obs} and its ratio to the critical density as Ω_{obs},

$$\Omega_{obs} = \rho_{obs}/\rho_c = 1.02 \pm 0.02 \ , \tag{7.118}$$

at present. Some people find the fact that Ω is so close to unity surprising.

In what sense is this "so close"? The point is that for the present value to be within a percent or so of unity, it must have been much closer earlier. Writing the difference between the actual density at time t, $\rho(t)$, and the critical density as $\Delta\rho(t)$, from Eqs.(7.56) and (7.118)

$$\Delta\Omega(t) = \Delta\rho(t)/\rho_c(t) = k/\dot{a}^2(t) \ . \tag{7.119}$$

Clearly this is zero for $k = 0$ and

$$\dot{a}_+(t) = \cot(\eta/2) \quad , \quad \dot{a}_-(t) = \coth(\eta/2) \tag{7.120}$$

yields

$$\Delta\Omega_+(t) = \tan^2(\eta/2) \quad , \quad \Delta\Omega_-(t) = \tanh^2(\eta/2) \quad . \tag{7.121}$$

The current value of η is likely to be about 0.1 (rad). Thus $\Delta\Omega_\pm$ should scale roughly as η^2 as we go to earlier times (since tan and tanh go roughly as η for small enough η). Thus $\Delta\Omega_\pm(t)$ should scale roughly as $t^{2/3}$ (since $t \sim \eta^3$). In other words, when the Universe was about 13.8 years old, $\Delta\Omega_\pm(t) \sim \pm 10^{-6}$. Hence a difference from $\Omega = 1$ of even 10^{-5} at that time would give a current value of $\Omega_+ = 100$ or $\Omega_- = 1/100$. This gives the so-called *flatness problem*: "how is it that the initial value of Ω was so accurately unity?"

We shall discuss the most commonly accepted answer in some detail in the next section. Another popular answer is provided by the *anthropic principle*, that the Universe must be such as to produce intelligent life, for if it did not there would be one around to ask why the Universe should be as it is. Now, if

the Universe had started with an Ω that leads to $\Omega \sim 100$ at present, it would have re-collapsed long since and there would have been no time for intelligent life to evolve. On the other hand if Ω had been $\sim 1/100$ at present we could not have had enough matter to produce nucleosynthesis. To me this teleological argument is unsatisfactory and appears unscientific.

Yet another problem is often mentioned. Since we have an estimate of the baryonic matter density in the Universe and the baryons are practically entirely nucleons, the estimate extends to the nucleon number density. Further, since we know the temperature of the microwave background and its distribution is Planckian, we know the number density of primordial photons. The ratio of the latter to the former, called the *entropy* of the Universe, is $\sim 10^{10\pm1}$. Some people find it strange that the number is "so large". Why not of the order of unity or zero? This is called *the entropy problem.*

This question was constructed, in association with the baryogenesis problem, with an answer in mind (again discussed in the next section). An answer, relevant particularly for any discussion of General Relativity, was proposed by Roger Penrose [62]. As we saw in the previous chapter, the surface area of a black hole corresponds to four times its entropy (in Planck units). He proposed that the gravitational entropy is measured by the Weyl curvature. Thus, when the curvature is mainly due to the Ricci tensor, i.e. due to distributed matter, the gravitational entropy in low. However, on gravitational collapse the curvature is transferred increasingly to the Weyl tensor and the gravitational entropy becomes large. The "photon entropy" is negligible compared with the gravitational entropy so defined. Thus Penrose's answer is that the problem does not exist as the photon entropy is remarkably low — as might be expected near the start of the Universe.

I would like to add some comments on my own view of these "problems". A generic answer to all of them would have been that this is the way things are and so the initial conditions must have been such as to lead to the present situation. Such a glib answer is not satisfactory. A similar answer could have vitiated any attempts to explain the observed cosmogony and left us without an understanding of primordial nucleosynthesis or a prediction of the background radiation. On the other hand, we can carry the questions too far — beyond the context of the theories in terms of which we are trying to formulate answers. Further, there is a tendency to take speculative and tentative estimates as "hard fact". For example, it is not clear that there *is* an entropy problem. I could have asked why the inverse number is so close to zero. This is, in fact, the baryogenesis problem! While I *do* eel that it would be necessary to find out how an excess of baryons arose.

The most serious problem is the causality or horizon problem. One is forced, by present observations, to postulate initial conditions that mimic the effects of statistical "smoothing out" of unevenness in the matter and energy distribution. Let me add that there are various proposals which try to shift the problem to one of providing a mechanism for providing such uniform "initial conditions" after the big bang and restarting the Universe, as it were. In my opinion that is all that the most popular solution, given in the next section, does. It may be better to admit our ignorance and wait till a more satisfactory solution emerges.

7.18 The Inflationary Models

Brout, Englert and Gunzig [63], noted that that the total energy of the Universe, normal matter and negative gravitational energy, adds up to zero. Could it be, they then asked, that the Universe arose as a quantum fluctuation from nothing? What a field day for *all* believers: the atheists who believe that there is no Creator and the theists who believe there is! One could equally argue, like Pierre Simon Laplace, that as there is no need to postulate a Creator one dispenses with Him. As he said when he presented his book on Celestial Dynamics to Napoleon Bonaparte, when the latter commented that there was no mention of God in his book on the Heavens, "Je n'avais pas besoin de cette hypothèse-là" ("I had no need of that hypothesis"). Or, one could say as a theist, *that* is the intervention of the Creator.

Alan Guth [64] presented a novel proposal. Consider a Friedmann Universe starting with a GUT symmetry. Till the potential for the GUT Higgs field acquires a vacuum expectation value (VEV), there is a vacuum energy in the Universe, which is effectively a cosmological constant, causing exponential expansion of the Universe. Once it acquires the VEV, after a "GUT-breaking effort", normal matter is produced as the Higgs field decays and the exponential expansion gets "switched off". Comparing the expansion phase to what was then perceived as skyrocketing prices, he called this "cosmological inflation". Regions of the Friedmann Universe that had been in causal contact would suddenly get pushed out of apparent contact during the exponential expansion thus solving the causality, or horizon problem. At the same time, the Universe would look almost flat as the mass density produced would be just sufficient to stop the exponential expansion. In addition, since there would be a GUT-breaking event there would also be baryogenesis. All the problems would be resolved and "we would all live happily ever after".

Unfortunately, in all fairy tales an evil witch enters, and in every Garden of Eden there is an insidious serpent. The symmetry breaking has to occur over the whole Universe in regions that are out of causal contact, and this would lead to the production of "defects", as occur when crystals are formed over a large area very fast. The problem of defects can be picturesquely explained in terms of an analogy originally used by Salam to explain spontaneous symmetry breaking, called *Salam's banquet*. The purpose was to show how an inherently symmetric system can lose that symmetry due to asymmetric boundary conditions. Consider a banquet at a crowded round table where places are set with plates and one glass per guest. When the guests seat themselves in front of the plates, they will find glasses symmetrically placed to their right and left. When a guest reaches for a glass the symmetry is broken as a glass to the right or left is chosen. This is the spontaneous breakdown of symmetry. The table will then be set "right-handedly" or "left-handedly", depending on the choice of the first guest.

Now conceive of a very large banquet table so that when one guest picks up a glass at one place it does not induce all other guests to pick up the corresponding glasses. Another guest far away does not notice the event and picks up a glass on the other side. Thus many guests would spontaneously break the symmetry and so some would break it left-handedly while others would break it right handedly. As the effect would propagate around those who broke the symmetry there would arise places where a guest was left without a glass, as those on

the left broke the symmetry right handedly while those on the right broke it left handedly. Contrariwise, there would be guests with two glasses where the situation is reversed. These places would be "defects" in the distribution of glasses. It would take energy to move the extra glass to the guest with no glass. Unless that energy is supplied the defect would persist. *The same would apply to the spontaneous breakdown of a GUT in different regions of the Universe!*

Salam's banquet, being along a circle, was a 1-dimensional example and so the defects could only be 0-dimensional, i.e. points. For a field theory in 4-dimensional spacetime we can have *points*, *lines*, 2- and 3-dimensional topological defects. The first type behave like particles endowed with a magnetic charge and are consequently called *magnetic monopoles*. This is only the comparatively far-field behaviour. They have a very rich structure in the near field that is nothing like a classical magnetic monopole. None have been observed and there are stringent limits on the number density around the Earth. Any theory that requires magnetic monopoles would have to provide a mechanism to sweep away enough from our range of observation. It was pointed out that with exponential expansion all inhomogeneities would get smoothed out like removing wrinkles from a fabric by stretching it. However, the net result seems to be that the monopoles remain a problem. The lines defects are called *cosmic strings*. The others are called *domain walls* and *textures* respectively. There is a tendency of a GUT Cosmology to overproduce magnetic monopoles, as their mass is equivalent to the GUT-breaking scale ($\sim 10^{16} - 10^{17}$ GeV) and their number density would produce so much total mass that the Universe would be extremely short lived — much less even than its present age. Similarly, the domain walls create a problem of too much mass and, *still worse*, far too much anisotropy. The textures seem harmless but uninteresting cosmologically. The strings, however, would be able to provide a mechanism for creating the present inhomogeneous structure starting from a homogeneous Universe. As such they could be very useful.

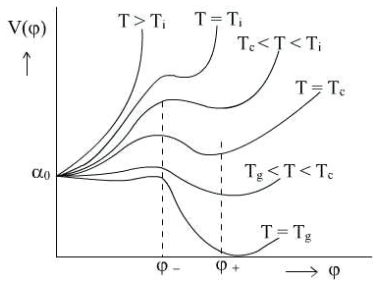


Figure 7.20: A quartic inflationary potential. The shape of the potential changes with temperature as new extremal points appear in the potential leading to phase translations at $T = T_i, T_c$ and T_g. At T_c the "true vacuum" shifts from 0 to φ_+. The barrier at φ_- stops the Universe entering the true vacuum and so it acquires a large vacuum energy causing inflation.

To make the discussion more concrete, consider the quartic potential of the Higgs field ϕ, with positive coefficients, see Fig. 7.20,

$$V(\varphi) = \alpha_0 + \frac{1}{2}\beta(T)\varphi^2 - \frac{2}{3}\gamma(T)\varphi^3 + \frac{1}{4}\delta(T)\varphi^4 \ , \qquad (7.122)$$

where α_0 is constant but β, γ, δ are temperature dependent. Its extrema are obtained by setting its first derivative,

$$V'(\varphi) = \varphi[\beta(T) - 2\gamma(T)\varphi + \delta(T)\varphi^2] \ , \tag{7.123}$$

equal to zero. Assume a Friedmann model Universe to start with and that at early times (high temperatures) $\gamma^2(T) < \beta(T)\delta(T)$. In this regime there is only one extremal, the global minimum at $\varphi = 0$, namely $V(0) = \alpha_0$. Now as temperature decreases, at some T_i, $\gamma^2(T_i) = \beta(T_i)\delta(T_i)$, $V'(\varphi) = 0$ at $\varphi_i = \gamma(T_i)/\delta(T_i)$. This is a point of inflection with

$$V(\varphi_i) = \alpha_0 + \gamma^4(T_i)/12\delta^3(T_i) \ . \tag{7.124}$$

Below this temperature there will be a local maximum and a local minimum at

$$\varphi_\pm = [\gamma(T) \pm \sqrt{\gamma^2(T) - \beta(T)\delta(T)}]/\delta(T) \ , \tag{7.125}$$

apart from the local minimum at $\varphi = 0$, namely $V(0) = \alpha_0$. The local minimum will be at φ_- and the other local minimum, at φ_+, will be

$$V(\varphi_+) = \alpha_0 + \frac{1}{2}\varphi_+[\frac{1}{2}\beta(T) - \frac{1}{3}\gamma(T)\varphi_+] \ . \tag{7.126}$$

The new local minimum will be higher than $\varphi = 0$ till we reach the *critical temperature*, T_c, at which

$$\gamma^2(T_c) = 9\beta(T_c)\delta(T_c)/8 \ . \tag{7.127}$$

At this stage there are two global minima. For $T < T_c$ the global minimum shifts to φ_+ and finally hits the ground state when $V(\varphi_+)$ given by Eq.(7.126) becomes zero, at some temperature $T = T_g$.

The local maximum at $\varphi = \varphi_-$ provides a barrier between the two local minima. Thus, even though the global minimum would be at φ_+ for $T \leq T_c$, the Universe would tend to stay at the previous minimum till it can tunnel through the barrier. Thus it stays in a "false vacuum" and the energy difference between the "true" and "false" vacua remains as a vacuum energy. A vacuum energy is indistinguishable from a cosmological constant and hence gives an essentially De Sitter model (with a slight Friedmann contamination), with its concomitant exponential expansion. This expansion was called inflation. The problem with this model was that the inflation switches off due to a tunneling between different vacua, leading to the formation of defects *ála* Salam's banquet. These defects appear in the form of domain boundaries. One can obtain an inflation factor of e^{65} with this procedure, which would be enough to eliminate the monopoles adequately and provide the proposed solution of the horizon problem. However, the energy density in the domain walls would be unacceptably high.

A new inflationary model was proposed by Andre Lindé [65], which avoided this problem and provided "a graceful exit" from inflation. One regards the Universe as represented by a "ball" on the potential diagram which can "roll" on it exactly as in Classical Mechanics. The idea was to replace the tunneling by a "slow roll" to the true vacuum. This was achieved by taking a very slowly varying potential near the false vacuum, see Fig. 7.21. Further, he pointed out, the field would oscillate down to its ground state. This oscillation would lead to radiation.

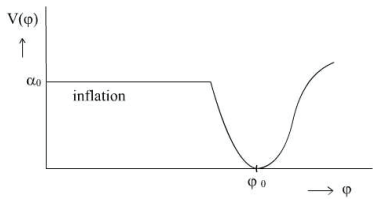


Figure 7.21: The new inflationary potential. The Universe inflates as it "rolls" slowly from α_0, at $\varphi = 0$ towards the minimum, 0, at $\varphi = \varphi_0$. Then it rolls quickly as it enters the deeper part of the potential well, overshoots, and oscillates about φ_0 causing re-heating. [Bear in mind that *if* there is a GUT and a GUT-breaking Higgs there is no basis to suppose the potential to be either quartic or like this potential.]

The original exponential expansion would exponentially cool the Universe but the oscillation down would release all the vacuum energy as radiation which would reheat the Universe, see Fig. 7.22. This radiation would then provide the bulk of the matter-energy content of the current Universe, with the original matter energy content as a trace contamination. Since this process would be responsible for switching off inflation and re-starting Friedmann expansion the density provided must necessarily be critical. Thus $\Omega_{inflation} = 1$. This solves "the flatness problem". Of course, due to the original contamination $\Omega > 1$ according to this model. Again a 65-fold inflation can be obtained with the scale factor varying as depicted in Fig. 7.23. This is the model most popular with the modern Particle theorists.

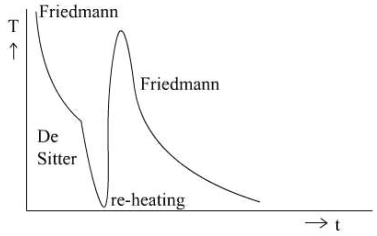


Figure 7.22: The thermal history of the Universe according to inflation. The Universe first cools as a Friedmann radiation dominated model $T \sim 1/\sqrt{t}$. Then, when inflation sets in it cools much faster, dropping exponentially till the re-heating raises the temperature again and it cools as a Friedmann radiation then matter dominated model.

Lindé [66], however, was not satisfied with this model. His point was that it may be that "the cure is worse than the disease". We are solving a "fine-tuning" problem of having to choose one number in 10^{10} or 10^{11} (i.e. a number to the accuracy 10^{-11}) by choosing a very special potential. There being infinitely many more functions than numbers, this amounts to a "finer tuning". He therefore considered a more generic potential, see Fig. 7.24, with the field not at a local minimum (which is taken to be a global minimum). The "roll down" would cause inflation and the oscillations reheating. Some regions of the Universe would acquire too much matter density and consequently re-collapse.

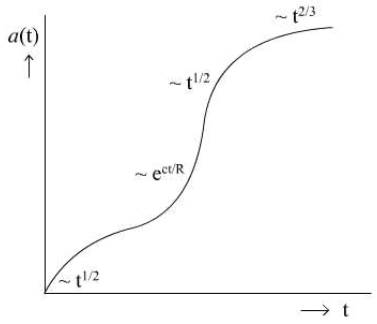


Figure 7.23: The scale factor history of the Universe according to inflation. It starts with $a(t) \sim t^{1/2}$, then inflates as $\exp(ct/R)$, where R comes from the effective cosmological constant arising from a φ_0 (or φ_+) $\neq 0$. It then again goes as $t^{1/2}$ after inflation switches off due to reheating and later as $t^{2/3}$ when it is matter dominated. The initial case could equally well have been $t^{2/3}$ or a $t^{1/2}$ changing to $t^{2/3}$.

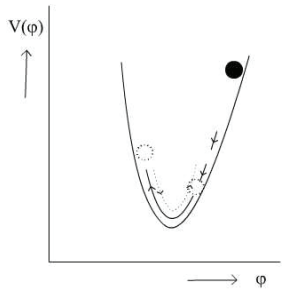


Figure 7.24: Chaotic inflation. The Universe starts at any point on a generic potential well, rolls down and oscillates about the minimum. There are repeated inflations and reheatings in this model, which avoids all the other problems of the inflationary models, but leads to many other regions of the Universe not observed so far, but observable in principle.

Some would expand too fast to be reheated and would be totally empty. However, there would be regions which inflated just right and reheated just right to produce what we observe. Thus every possibility occured/occurs/will occur in this Universe, but we could only be where it happened to produce conditions just right to produce us. Notice the strong resemblance to the anthropic principle, though it avoids the teleological argument involved there.

All the inflationary models provide exotic GUT-breaking relics as dark matter candidates. Those models that solve flatness *need* these relic particles to provide the difference between critical and baryonic matter density. They all appeal to the GUT for solving the baryonic matter density. They all appeal to the GUT for solving the baryogenesis problem but it requires very fine tuning to make the numbers come out right. In my opinion it would be fair to say that they can *accommodate* the baryogenesis and entropy problems rather than *resolving* them.

There is a severe problem about the claimed "resolution of the horizon prob-

lem". All that is really managed is to provide *causal contact* between the two regions, i.e. *light* from one point to the other and *vice versa*. What is required is that *thermalization* occur, i.e. *matter* from both regions undergo repeated scatterings off each other. Of course, we could also manage with light transferring energy between the two regions, but not just by moving freely. It would need to have an effective speed very much less than that of light in vacuum to provide for all the scatterings along the way. However, this is not the severe part of the problem. Remember that the matter-energy we are dealing with was produced at reheating and is responsible for switching off inflation. Thus inflation can not have taken thermalized matter-energy out of causal contact. Rather, the radiation we see was produced at both regions long after they had lost causal contact! How, then, does inflation "resolve the horizon problem". Essentially, it provides a mechanism for taking the effective initial conditions at both places to be the same. Instead of thermalizing, it allows the initial conditions to be such that the temperature is the same, within the limits of anisotropy, at all places. It provides a place for anisotropy but I feel that the claim that the observed anisotropy "proves" the validity of inflation is too strong. The prediction of anisotropy at the observed level and with the observed features is non-unique.

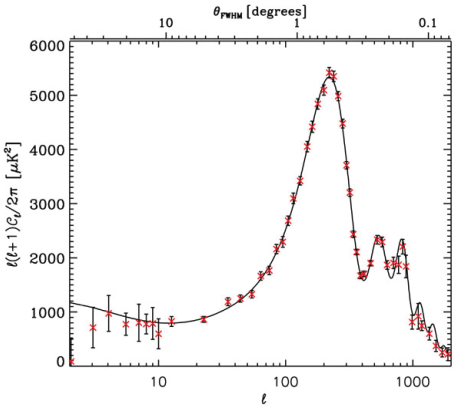


Figure 7.25: The power spectrum of the CMB. The intensity of the anisotropy is plotted against the mode number, or equivalently the angle in degrees. As can be seen, the fit looks good. [The image is taken from the web.]

There *is* "evidence" claimed for the "success" of inflation. Take the Planck spectrum of the CMB and Fourier transform it so that we get a mode decomposition in terms of spherical harmonics. If there was inflation the modes predicted for temperature anisotropies are consistent with the observed values with reference to correlations across angular scales of the order of 8°, see Fig. 7.25. However, there have been no serious attempts to allow for alternate explanations to be considered, as journals tend to say "*we already know that inflation is correct*" and so other possibilities are not checked. Why do I say they are wrong to do so? Because up to now there is no sign of the field causing inflation (the "inflaton" as it is now called). If is to be a serious scientific theory with definite predictions, it cannot be chaotic inflation. In the other case, where many problems are seriously resolved, one needs a GUT-breaking Higgs and there is no evidence for it.

To summarize, the "successes" of inflation must be taken with a pinch of salt. It may be better to admit our ignorance and continue to search for more satisfactory answers.

7.19 Exercises

1. Try to construct a "steady state model" with continuous creation of matter. Now compute the second variation of the Lagrangian and check whether the model is stable or unstable. [Hint: One needs the the second variation be positive for stability and for functions of several variables this requires that the matrix of second derivatives have a positive derivative. Even if it is zero that does not provide stability.] Is an unstable model physically acceptable?

2. Obtain the observational data on the accelerated expansion of the Universe from the net. Now compute the value of Λ required to fit the observations. Write this in Planck units and check the extent to which an expectation of a Planck scale value is off the observed mark. If the value had been of the order of Planck scale, obtain the life of the Universe. Is the expectation of Planck scale Λ reasonable?

3. Construct an "old" inflationary model that would fit the observational data and try to compute the number of defects of various types produced. Assuming the GUT-breaking Higgs to have a mass $m_H \sim 10^{16.5}$ GeV, work out the consequences of each of them in your model.

4. Construct a "slow roll" inflationary potential. Determine the number of defects produced in your "new inflationary" model. [Hint: Instead of using an analytic function, try using patching requiring that the potential remain smooth, i.e. nowhere is the second derivative discontinuous.] Are all the defects adequately removed in your model?

Chapter 8

Some Special Topics

In this chapter I will, on the one hand, put in more Mathematics that is used in the study of GR and, on the other, give further developments of the physical aspects, especially with special reference to recent observational and experimental verification of the predictions of GR. I fear that this makes for a somewhat incoherent potpourri of topics with no central theme running through it. The idea is to provide the tools that are being used in research in various mathematical aspects of GR and to present the latest findings on the observational and experimental side of GR, and not to provide logical cohesion.

8.1 Two-component Spinors

Spinors were introduced by Élie Joseph Cartan [67] and are vectors of a symplectic representation space of the orthogonal group. The symplectic group, $Sp(2n)$, is a group in which the length of all vectors invariant under the action of the group is zero. This can be arranged by having the metric tensor skew symmetric or by taking the vectors to be complex such that with the Euclidean metric tensor the length turns out to be zero. They are related to the orthogonal group and can be shown to be related to the intrinsic angular momentum, or spin. The number of components of the spinor for a $2n$-dimensional space is 2^{n-1}. Thus for 4-d we have 2-component spinors. Two-component spinors were used by Roy Kerr to derive the Kerr metric and are heavily used in discussions of gravitational waves. Their geometrical significance was explained by Penrose , who went on to define twistors from there, which were used in his attempt to develop a theory of Quantum Relativity. They are used by Abhay Ashtekar for his "loop quantum gravity approach". A very thorough discussion of spinors for use in GR is given by Penrose and Rindler [35]. I will give a very brief introduction to the subject.

The spinor basis to write a 4-vector $\mathbf{x}$, with usual spacetime components x^μ, as the spinor equivalent is $\sigma_\mu^{AA'}$, where the $\mu = 0$ component is the 2×2 identity matrix and the other three components are the Pauli spin matrices:

$$\left.\begin{array}{ll} \sigma_0^{AA'} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, & \sigma_1^{AA'} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \\ \sigma_2^{AA'} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, & \sigma_3^{AA'} = \begin{pmatrix} 0 & \iota \\ -\iota & 0 \end{pmatrix}, \end{array}\right\} \tag{8.1}$$

so that

$$x^{AA'} = x^{\mu}\sigma_{\mu}^{AA'} = \begin{pmatrix} ct+x & y+\iota z \\ y-\iota z & ct-x \end{pmatrix} . \tag{8.2}$$

It is obvious that the determinant of the spinor representation of $\mathbf{x}$,

$$|x^{AA'}| = x^{AA'}x^{BB'}\epsilon_{AA'}\epsilon_{BB'} = x^{\mu}x^{\nu}g_{\mu\nu} \tag{8.3}$$

and hence we have the identification

$$g_{\mu\nu} = \sigma_{\mu}^{AA'}\sigma_{\nu}^{BB'}\epsilon_{AA'}\epsilon_{BB'} . \tag{8.4}$$

This is what makes it a symplectic representation of the orthogonal group.

We need to look at the group theoretic aspects of spinors to properly understand their geometry. For this purpose let me briefly define what a Lie group is. It is a manifold which is also a group and it is path-wise connected to the identity. The binary operation for the group is the Lie product of vector fields, which we met as the Lie derivative of one relative to the other. As such, the Lie product is anti-symmetric. It will be recalled the manifold is a connected space. Here we are requiring that all points in the space be connectible by some path to the point that is the identity element of the group. Since the Lie product under multiplication, with the identity being 1 (not 0). Now 0 has no inverse element and so we would have to exclude it from the manifold, $\mathbb{R}^* = \mathbb{R} \setminus \{0\}$. However, $\mathbb{R}^*$ is disconnected and so it is no longer a manifold. Instead, consider $\mathbb{C}$, with the identity being 1 again. Again 0 would have to be excluded, $\mathbb{C}^* = \mathbb{C} \setminus \{0\}$. However, now we can always go *around* 0 and so we have a doubly connected manifold (i.e. having two classes of inequivalent paths).

For a general manifold of dimension n, there will be n linearly independent vector fields at any point, but for the whole manifold there will be a larger group, as the vector field could scale from one point to another in such a way that the space remains invariant. We saw this as the isometries of the space. There will be paths generated by each isometry generator, with a parameter associated. The maximal Lie group for a flat space will be $\mathbb{R}^n \otimes_s SO(n)$, which has $n(n+1)$ parameters, where $\otimes_s$ is the semi-direct product, meaning that the two subgroups involved do not commute. For a space of constant curvature, if it is compact, the group will be $SO(n+1)$ and if it is non-compact, $SO(p,q)$ such that $p+q = n+1$. Thus for Minkowski spacetime we have the Poincaré group $\mathbb{R}^n \otimes_s SO(1,3)$.

For the present purposes, there are 4 independent components of the vector $\mathbf{x}$ in its spacetime and its spinor representations. In the latter the spinor basis is unitary and unimodular. As such, the group under which it is invariant is $SU(2)$. We started off with the connected Lorentz group, $SO(1,3)$ (the translations not being included for the spinor), and have seen that it leads to $SU(2)$. As such, the two groups must be isomorphic: $SU(2) \simeq SO(1,3)$. Of particular interest is the case of the null vector. Since the 4 components are related by one constraint, there are only three independent components involved. In this case we can write $x^{AA'} = \xi^A\overline{\xi}^{A'}$, where $\overline{\xi}^{A'} = \overline{\xi^A}$. To see this, note that ξ^A has 2 complex components, which amounts to 4 real components. However, they satisfy the symplectic constraint $\xi^A\xi^b\epsilon_{AB} = 0$. As such the number of components balance and it is obvious that $\mathbf{x}$ is a null vector, as $x^{AA'}x^{BB'}\epsilon_{AB}\epsilon_{A'B'} = 0$. We call ξ^A a *spinor* and $\xi^A\overline{\xi}^{A'}$ the *conjugate spinor*, having arbitrarily chosen to regard the former as "the original" quantity and the latter as obtained from it.

The spinor gives the arrow on the null cone representing $\mathbf{x}$, but must contain more information because it has 4 real components. To visualise this, Penrose wrote $\xi^A = \xi^0(1 \quad \frac{\xi^1}{\xi^0})$ provided $\xi^0 \neq 0$, in which case it is $(0 \quad \xi^1)$. In the former case, the $(1 \quad \frac{\xi^1}{\xi^0})$ is the projective spinor and represents a direction on the Riemann sphere (the compactified complex plane). In terms of the null cone, it gives the direction of the arrow on the null cone. The question is what does ξ^0 represent? Putting ξ^0 in the polar form $\lambda e^{\iota\theta}$, we see that λ gives the square root of the "magnitude" of the null vector. You might object that the null vector has zero magnitude. However, for the null vector the magnitude gives the "length" of the arrow $\sqrt{2}ct$. Thus the extra, fourth, component of the spinor is a phase, and one is left with interpreting that. To do so, we define another spinor η_A such that $\xi^A\eta_A = 1$. Now define the real vector $\mathbf{w}$ corresponding to the spinor representation $w^{AA'} = \xi^A\overline{\eta}^{A'} + \overline{\xi}^{A'}\eta^A$. It is easy to verify that $x^\mu w_\mu = x^{AA'}w_{AA'} = 0$ and $w^{AA'}w_{AA'} = -2$. As such, $\mathbf{w}$ is a spacelike vector orthogonal to the null vector, $\mathbf{x}$, constructed entirely from the information given in ξ^A. If we multiply ξ^A by a complex phase, $\xi^A \to \tilde{\xi}^A = e^{\iota\phi}\xi^A$, so it is rotated through angle ϕ, we must multiply η_A by $e^{-\iota\phi}$, so as to maintain the condition that their product gives 1. Then $\overline{\eta}^{A'} \to \tilde{\overline{\eta}}^{A'} = e^{\iota\phi}\overline{\eta}^{A'}$, i.e. it must also be rotated through ϕ so that $w^{AA'} \to \tilde{w}^{AA'} = e^{\iota\phi}w^{AA'}$, i.e. it is rotated through 2ϕ. Hence, if we take $\phi = \pi$, $w^{AA'}$ is unchanged, even though xi^A has been converted to $-\xi^A$. Thus $\mathbf{x}$ and $\mathbf{w}$ span a half-plane, rather than a full plane. This gives Penrose's famous pennant representation of the spinor, called the Penrose flag, see Fig. 8.1.

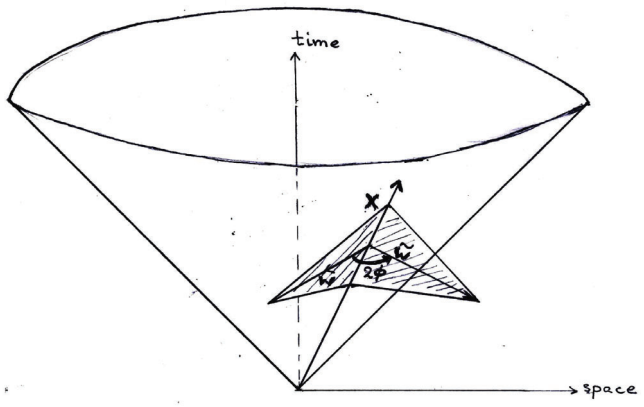


Figure 8.1: The Penrose flag. The flagpole is the null vector $\mathbf{x}$, lying along the null cone, and an American High School pennant is given by the spacelike vector $\mathbf{w}$, which is tangent to the null cone. If the spinor, $\xi^A \to \tilde{\xi}^A = e^{\iota\phi}\xi^A$, i.e. it is rotated by ϕ, the pennant is rotated by 2ϕ, as $w^{AA'} \to \tilde{w}^{AA'} = e^{2\iota\phi}w^{AA'}$.

One can now convert all other quantities to spinor representation by using the spinor basis. Bear in mind that any skew spinor of rank two, X_{AB} must be proportional to ε_{AB}, as it is a 2×2 skew matrix. Thus, for example,

$$R_{\mu\nu} \to R_{ABA'B'} \ , \ R_{\mu\nu\rho\pi} \to R_{ABCDA'B'C'D'} \ . \tag{8.5}$$

For the full Riemann tensor, use the same method of breaking up a tensor

according to its algebraic symmetries, and use Eqs. (3.12 - 3.15). This gives

$$\left.\begin{aligned} R_{ABCDA'B'C'D'} =& \Psi_{ABCD}\varepsilon_{A'B'}\varepsilon_{C'D'} + \bar{\Psi}_{A'B'C'D'}\varepsilon_{AB}\varepsilon_{CD} \\ &+ \Phi_{(AB)(C'D')}\varepsilon_{A'B'}\varepsilon_{CD} + \bar{\Phi}_{(CD)(A'B')}\varepsilon_{AB}\varepsilon_{C'D'} \\ &+ 2\Lambda(\varepsilon_{AC}\varepsilon_{BD}\varepsilon_{A'C'}\varepsilon_{B'D'} - \varepsilon_{AD}\varepsilon_{BC}\varepsilon_{A'D'}\varepsilon_{B'C'}) \ . \end{aligned}\right\} \quad (8.6)$$

The Ricci tensor is obtained from the above equation by contraction, using $\varepsilon^{AC}\varepsilon^{A'C'}$. Obviously, the Ψ terms drop out, as they are symmetric in all indices and are contracted with the skew ε. Thus

$$R_{BDB'D'} = \Phi_{(BD)(B'D')} + \bar{\Phi}_{(BD)(B'D')} + 6\Lambda\varepsilon_{BD}\varepsilon_{B'D'} \ . \quad (8.7)$$

Clearly, this immediately yields $R = 24\Lambda$. Since the Ψ terms are the trace-free part of the Riemann tensor they give the Weyl curvature tensor which represents the pure gravitational field, and hence gives gravitational waves,

$$C_{ABCDA'B'C'D'} = \Psi_{ABCD}\varepsilon_{A'B'}\varepsilon_{C'D'} + \bar{\Psi}_{A'B'C'D'}\varepsilon_{AB}\varepsilon_{CD} \ . \quad (8.8)$$

One can go further by writing the derivative operator, ∇_μ, in spinor form as $\nabla_{AA'}$. While it may seem "more of the same", here we get some further insights and development of techniques to get more information out. Instead of acting on the usual tensor quantities, we can act on a spinor, ξ^B, and get something with no analogue in usual tensor geometry. Even more, we can contract with "half" of the spacetime index to get $\nabla_{AA'}\xi^A = m\eta_{A'}$. This, with its conjugate equation, $\nabla_{AA'}\eta_{A'} = -m\xi^A$, gives the Dirac equation. Thus the equation for a massless spin-half field is $\nabla_{AA'}\xi^A = 0$. This is directly generalized to any spin (linear) zero rest-mass field $\nabla_{AA'}\psi^{(A\ldots N)} = 0$. Thus the equation for linearized gravity is $\nabla_{AA'}\Psi^{ABCD} = 0$, as it is a spin-two field. For the general, massive, Dirac equation there are two spinors, each possessing an $SU(2)$ symmetry and there is a negative sign in the coupling. Hence the symmetry group for the Dirac equation is $SU(2,2)$.

Acting on a basis vector the covariant derivative operator gives the Christoffel symbols (taking the space to be torsion-free). One could do the same for spinors. Of course, once again we need not limit ourselves to the basis vector, but can use the basis *spinor*. A big advantage of selecting a basis spinor is that it has less freedom of choice, and so is "more unique" in a sense that will become clear. For usual four-dimensional spacetime, one has to select three basis vectors (with increasing restrictions in the choice). Then the fourth comes out in an obvious manner. In the same way, choosing one basis spinor, the other will come out in an obvious manner. Thus, instead of choosing nine numbers, one only has to choose four. Here we choose one Penrose flag, $o^A = (1 \ \ 0)$. Then the other one would be ι^A, such that $o^A\iota_A = 1$ and that $2o^{[A}\iota^{B]} = \varepsilon^{AB}$. This is satisfied by $\iota^A = (0 \ \ 1)$. That completes the basis $\zeta_a^A = (o^A \ \ \iota^A)$. This is the sense in which the spinor basis is "more unique".

The spinor basis automatically leads to a null basis for spacetime, called the *null tetrad*. You may well wonder how there can be four independent null vectors, since the null cone is a three dimensional hypersurface. However, with the spinor basis we can define:

$$l^\mu \sim o^A\bar{o}^{A'} \ ; \ m^\mu \sim o^A\bar{\iota}^{A'} \ , \bar{m}^\mu \sim \iota^A\bar{o}^{A'} \ , \ n^\mu \sim \iota^A\bar{\iota}^{A'} \ , \quad (8.9)$$

As you see, there is a conjugate pair of complex vectors. *This* is what makes it possible to get four null vectors in. We can use the spinor basis to write the components of the pure gravitational field "uniquely" (up to the choice of o^A) as $\Psi_0 = \Psi_{0000},\ \Psi_1 = \Psi_{0001},\ \Psi_2 = \Psi_{0011},\ \Psi_3 = \Psi_{0111},\ \Psi_4 = \Psi_{1111}$, which are 5 complex, or 10 real components. Using the null tetrad with the Weyl tensor we easily see:

$$\left.\begin{array}{ll} \Psi_0 = C_{\mu\nu\rho\pi} l^\mu m^\nu l^\rho m^\pi \ , & \Psi_1 = C_{\mu\nu\rho\pi} l^\mu m^\nu l^\rho n^\pi \ , \\ \Psi_2 = C_{\mu\nu\rho\pi} l^\mu m^\nu \bar{m}^\rho n^\pi \ , & \Psi_3 = C_{\mu\nu\rho\pi} l^\mu n^\nu \bar{m}^\rho n^\pi \ , \\ \Psi_4 = C_{\mu\nu\rho\pi} \bar{m}^\mu n^\nu \bar{m}^\rho n^\pi \ . & \end{array}\right\} \tag{8.10}$$

To extend the spinor analysis from algebra to *calculus*, we take the covariant derivative of our spinor basis and write it back in terms of the spinor basis, just as we had done for the Christoffel symbols. Here we get "spinor symbols" $\gamma^B{}_{AA'C} = \zeta^B_a \nabla_{AA'} \zeta^a_C$. They are called *Ricci spin coefficients*. These are 16 complex quantities, which amount to 32 real components — not 40, as in their vector counterpart. Even more than before, *this* is how the spinor basis is "more uniquely" defined. Much of the spurious freedom of choice of basis vectors has been eliminated. Newman and Penrose used this spinor basis and spin coefficients to set up a formalism to investigate gravitational radiation [68]. It has been found extremely useful to study the behaviour of classes of geodesics in these spacetimes.

I am now in a position to justify my claim in Chapter 3 (section 3.9) that the Weyl tensor of valence $\left[\begin{smallmatrix}1\\3\end{smallmatrix}\right]$ is invariant under conformal transformations and with other valences simply scales, see Eq. (3.204). By definition $2\xi^\mu{}_{;[\rho\pi]} = R^\mu{}_{\nu\rho\pi}\xi^\nu$. Putting this into spinor form the skew second derivative becomes a symmetrization over the the non-contracted $\nabla_{X'(A}\nabla^{X'}_{B)}$, along with the ε^{AB}. We now use ξ^A instead of ξ^μ, which will lead to extra terms apart from the Weyl spinor. To get rid of them we symmetrize over all the unprimed indices, to get

$$\nabla_{X'(A}\nabla^{X'}_B \xi_{C)} = \Psi_{ABCD}\xi^D \ . \tag{8.11}$$

The Conformal transformation is $g_{\mu\nu} \to \tilde{g}_{\mu\nu} = \Omega^2 g_{\mu\nu}$. Taking Ω to be real, we can put $\varepsilon_{AB} \to \tilde{\varepsilon}_{AB} = \Omega\varepsilon_{AB}$ and the corresponding conjugate equation. Thus the first basis spinor $o_A \to \tilde{o}_A = \Omega^{1/2} o_A$. However, a vector ξ^μ remains unchanged under Conformal transformations, so ξ^A will also remain unchanged. We need to consider what happens to the covariant derivative operator, $\tilde{\nabla}_{AA'}$. For this purpose take $\tilde{\nabla}_{AA'}\tilde{\varepsilon}_{AB} = 0$ and write the $\tilde{\varepsilon}^{AB} =$ We now need to consider the effect of transforming the affine connection, or covariant derivative operator, as $\nabla_\mu g_{\nu\rho} = 0$ should imply that $\tilde{\nabla}_\mu \tilde{g}_{\nu\rho} = 0$. Putting in the conformal re-scaling, this implies that $\tilde{\nabla}_\mu g_{\nu\rho} = k_\mu g_{\nu\rho}$, where $k_\mu = (\ln\Omega^2)_{,\mu}$. Putting this into the spinor version,

$$\tilde{\nabla}_{AA'}\xi^B = \nabla^{AA'}\xi^B + \varepsilon_A{}^B \Upsilon_{CA'}\xi^C \ , \tag{8.12}$$

where $\Upsilon_{AA'}$ corresponds to k_μ. Raising the unprimed indices by using the spinor form of the metric tensor, the expression would have to be multiplied by Ω^{-2}. Now symmetrizing over the unprimed indices gets rid of the term involving $\Upsilon_{AA'}$. Writing Eq. (8.11) with the unprimed indices raised then yields

$$\tilde{\Psi}^{ABC}{}_D\xi^D = \Omega^{-3}\Psi^{ABC}{}_D\xi^D \ , \tag{8.13}$$

which leads directly to $\tilde{\Psi}_{ABCD} = \Psi_{ABCD}$. Since the Weyl tensor is given in terms of the Weyld spinor with the spinor metric tensor and its conjugate, $\tilde{C}_{\mu\nu\rho\pi} = \Omega^2 C_{\mu\nu\rho\pi}$ or $\tilde{C}^{\mu}{}_{\nu\rho\pi} = C^{\mu}{}_{\nu\rho\pi}$, as claimed.

In general we can write $\Psi_{ABCD} = \alpha_{(A}\beta_B\gamma_C\delta_{D)}$. It may appear that the right side has 8 independent complex components, while the left had only 5. That is not so, since there is only one independent magnitude and phase for all four spinors on the right and the rest are just complex ratios. Thus we again get five complex components. If we contract Ψ_{ABCD} with α^A, since the $\alpha_A\alpha^A = 0$ it is obvious that α^A will appear as an "eigenspinor" for the Ψ_{ABCD}. The corresponding null vector is called a *principal null direction* (*pnd*) of the Weyl tensor. We say that the Weyl field is *algebraically general* if all four *pnd*s are independent, and otherwise we call it *algebraically special*. Petrov [19] provided a classification of the algebraically special cases in terms of the *pnd*s for the Weyl tensor. Since the tensor is of rank 4, the process of determining it is cumbersome. Penrose explained the classification very simply: if all are different call it type [1,1,1,1]; if two are the same, while the other two are different call it type [1,1,2]; if two pairs are the same, but different from each other, call it type [2,2]; if three are the same and the fourth is different, call it type [1,3]; and if all four are the same, call it type [4]. Further, if any one of them is zero, call it of type [-]. Petrov called the first type I; the second type II; the third type D; the fourth type III; the last type N; and the one where the field disappears of type O. In this case the only curvature is due to matter and not due to the gravitational field.

One can generally expect gravitational waves to be produced in a violent event. Thus, close to the source there would, presumably, be no symmetries. However, very far from the source, it seems reasonable to expect high symmetry. Thus, close to the source the wave should be algebraically general and very far away of type [4] (or N). As the wave moves away it should go to type [1,1,2] by two *pnd*s aligning, and then to type [1,3] by one of them aligning with the pair and finally go to [4] by all aligning, giving the *pnd* along the motion of the wave. This is the Sachs peeling property (see [35] vol. 2). The strength of the wave decreases with distance progressively from $1/r^5$ for the algebraically general case to $1/r$ for type [4].

The spinor formalism also provides a classification of the other fields. For example, the Maxwell field is given by $F_{\mu\nu}$, which is a skew tensor. Thus we obtain the corresponding spinor representation

$$F_{ABA'B'} = \phi_{AB}\varepsilon_{A'B'} + \bar{\phi}_{A'B'}\varepsilon_{AB} \ . \tag{8.14}$$

The Maxwell field spinor ϕ_{AB} can be written as $\alpha_{(A}\beta_{B)}$. The field will be algebraically general, [1,1], if the two are distinct and special, [2], if they coincide, in which case it will be type N. If one disappears the Maxwell field will vanish.

8.2 Spacetime Symmetries

Apart from their aesthetic value, symmetries are extremely useful in many ways. We have already seen that spherical symmetry reduces the complexity of the Einstein field equations to manageable proportions. Including time translational invariance provides a unique solution to the vacuum equations with remarkable ease. On account of Noether's theorem, they can provide conserved quantities

for physical processes. Because of this use, local (Lie) symmetry groups lie at the heart of Quantum Field Theory and High Energy Physics. It runs out that symmetries of partial differential equations can reduce the number of independent variables, and of ordinary differential equations reduce the order. They can be used to generate solutions of Einstein's field equations possessing come minimal symmetry. We have just seen how the symmetries associated with the Maxwell tensor and the Weyl tensor help us understand the behaviour of electromagnetic and gravitational waves, by using the Petrov (or equivalently Penrose spinor) classifications. We will start with discussing symmetries of another quantity that we just saw above.

Books like [69, 21] use the null tetrad that was obtained from the spinor basis. Instead one can use Penrose's method for any of the tensors corresponding to the spinors encountered above. The natural one to consider is the *Peblanski tensor* coming from the first term in Eq. (8.7),

$$S_{\mu\nu} \sim \Phi_{(BD)(B'D')} + \bar{\Phi}_{(BD)(B'D')} \ . \tag{8.15}$$

From the spinor representation it is obvious that the trace of this tensor is zero. Since it comes from the Ricci tensor, it must be the trace-free part of the tensor

$$S_{\mu\nu} = R_{\mu\nu} - \frac{1}{4} R g_{\mu\nu} \ . \tag{8.16}$$

Since the electromagnetic stress-energy tensor is also traceless, the most common interpretation for the Peblanski tensor is as a "Maxwell fluid". Peblanski originally constructed the symmetric spinor

$$\Pi_{ABCD} = \Phi_{C'D'(AB)} \Phi^{C'D'}_{(CD)} \ , \tag{8.17}$$

and used the Penrose type of this spinor, just as was done for the Weyl spinor. As in that case, one needs to solve a quartic equation to obtain the eigenvalues and then classify according to the number of distinct eigenvalues, enumerating all possibilities for a given degeneracy.

The Peblanski classification is related to another classification called the Segré classification, by noting that if ξ^A is a principal spinor of Π_{ABCD} then

$$\Phi_{AB(C'D')} \xi^A \xi^B = \eta_{C'} \eta_{D'} \ , \tag{8.18}$$

for some $\eta_{C'}$ and so

$$\Phi_{AB(C'D')} \xi^A \xi^B \eta^{C'} = 0 \ , \tag{8.19}$$

which shows that one can reduce the quadratic fourth rank spinor to the Peblanski tensor itself. Now one classifies this tensor directly. (There is a discrepancy in the notation used by Penrose and by Stephani et al that can lead to confusion unless one is prepared for it.) I will not go further with these classifications.

Let us now proceed to the connection with Noether's theorem for our purpose. The introduction of curvature can destroy the symmetry of a flat space, as is obvious by taking a sheet of paper and crumpling it. As we saw, for n-dimensional space there are $n(n+1)/2$ isometry generators at most. The space can have positive, zero or negative curvature. The same applies to spacetimes. Rather than dealing with other spaces let us stick to spacetime. The flat case is Minkowski space. In this case we know that the group is $SO(1,3) \otimes_s \mathbb{R}^4$ and the translations give energy-momentum conservation, while the rotation subgroup,

$SO(3)$ gives angular momentum conservation. The subgroup $SO(1,3)/SO(3)$ is isomorphic to the rotation group and gives the proper Lorentz transformations, which imply spin angular momentum conservation.

To see what happens to the space if we introduce curvature, first consider what appears to be the simplest extension, the Schwarzschild spacetime. Solving the Killing equations as shown in section 3.8, we find that we are only left with four generators: the rotation group and a time translation, i.e. with angular momentum and energy conservation. Where did the others go? Due to the geometrically provided origin we have lost translational invariance. Going further away from the source makes a difference. Thus linear motion at uniform speed must also be lost. We see that the mathematically computed isometry could have been anticipated by some simple physical reasoning. It is because of this short-cut that John Wheeler used to say "I never start a calculation unless I know the answer". In fact the same reasoning would lead one to anticipate that the same isometry group would hold for the Reissner-Nordström spacetime.

While we took what might have seemed the "simplest" extension, it depends on what we regard as "simple". Let us make it more precise: "Are there any spacetimes or metrics with more symmetry generators lying between the Schwarzschild spacetime and Minkowski space?" (I distinguish between "metric" and "spacetime", as there could be metrics that do not correspond to physically realisable spacetimes.) In that case, what conservation laws would hold for such spacetimes? There are many methods to address this issue and many people have developed and used different ones. A very comprehensive discussion is given in the *Exact Solutions* book by [**?**]. I will not repeat what they have done but will present the method that I developed, which not only gives the symmetry groups lying between the minimal and the maximal, but the spacetimes, or classes of spacetimes, associated with each group. We call this a "complete classification" of the spacetimes possessing the minimal symmetry.

The idea is to solve the system of Killing equations for the components of the Killing vectors *and* the metric coefficients together. Why do so? Because we anyhow wanted to determine the isometries and because this is a system of first order differential equations, which are linear in both the Killing vector components and the metric coefficients, but is (quadratically) nonlinear on account of the products of the two. For comparison, the system of Einstein equations is second order and badly nonlinear. The big problem is that there are only $n(n+1)/2$ equations for $n(n+1)/2$ metric components and n Killing vector components. There are too few equations. The problem is resolved by assuming some minimal isometry group to reduce the number of metric coefficients for which one has to solve. As the reduction of the number of unknown coefficients of the metric tensor is not simply the number of symmetry generators, one can often solve the system without asking for too much symmetry, Since the metric coefficients are obtained here, the Einstein tensor, and hence the stress-energy tensor, can be computed. Of course, there is no guarantee that the stress-energy tensor will represent a meaningful spacetime. Consequently, the meaningful ones have to be picked out. Even then, they may not be of much interest. One simply selects those that are.

I will not try to give all the metrics obtained by the complete classification, but will merely provide one (non-trivial) example, namely of the spherically symmetric, static, metric as the minimal requirement. It has been shown [20]

that this requirement leads to the class of metrics

$$ds^2 = c^2 e^{\nu(r)} dt^2 - e^{\lambda(r)} dr^2 - e^{\mu(r)} a^2 d\Omega^2 \ , \tag{8.20}$$

where a is a constant with units of length. Now, either $\mu(r)$ is a constant function or it is variable. In the former case, without loss of generality, we can take it to be zero, as the constant can be "absorbed" into the arbitrary constant a^2. In that case we get the Bertotti-Robinson class of metrics, which can separately be classified. I will leave these out of the current illustration. Since $\mu(r)$ is a variable function, we can define a new r such that $r^2 = e^{\mu}$. In other words, without loss of generality, we can take $e^{\mu} = r^2/a^2$. Thus the most general spherically symmetric static metric, i.e. having the isometry group, $SO(1,3) \otimes_s \mathbb{R}^4$, is (in light units, so that $c = 1$)

$$ds^2 = e^{\nu(r)} dt^2 - e^{\lambda(r)} dr^2 - r^2 d\Omega^2 \ . \tag{8.21}$$

Writing the Killing equations as done before:

$$(K_{\mu\nu}): \qquad g_{\mu\nu,\rho} k^{\rho} + g_{\mu\rho} k^{\rho}_{,\nu} + g_{\nu\rho} k^{\rho}_{,\mu} = 0 \ , \tag{8.22}$$

which, on using Eq. (8.21) become,

$$\left.\begin{aligned}
(K_{00}): &\quad \nu' k^1 + 2k^0_{,0} = 0 \ , \\
(K_{01}): &\quad e^{\nu} k^0_{,1} - e^{\lambda} k^1_{,0} = 0 \ , \\
(K_{02}): &\quad e^{\nu} k^0_{,2} - r^2 k^2_{,0} = 0 \ , \\
(K_{03}): &\quad e^{\nu} k^0_{,3} - r^2 \sin^2\theta \ k^2_{,0} = 0 \ , \\
(K_{11}): &\quad \lambda' k^1 + 2k^1_{,1} = 0 \ , \\
(K_{12}): &\quad e^{\lambda} k^1_{,2} + r^2 k^2_{,1} = 0 \ , \\
(K_{13}): &\quad e^{\lambda} k^1_{,3} + r^2 \sin^2\theta \ k^3_{,1} = 0 \ , \\
(K_{22}): &\quad 2rk^2 + 2r^2 k^2_{,2} = 0 \ , \\
(K_{23}): &\quad r^2 k^2_{,3} + r^2 \sin^2\theta k^3_{,2} = 0 \ , \\
(K_{33}): &\quad 2r \sin^2\theta k^1 + 2r^2 \sin\theta \cos\theta k^2 + 2r^2 \sin^2\theta k^2_{,3} = 0 \ .
\end{aligned}\right\} \tag{8.23}$$

The easiest equation to integrate directly is obviously (K_{11}). Bear in mind that when integrating for several variables, the "constant of integration" becomes a "function of integration", as it is only constant with respect to the variable that has been differentiated. In this case the variable is r, so it will be a function of the other three variables. Thus

$$k^1 = e^{\lambda/2} A(t, \theta, \phi) \ . \tag{8.24}$$

We can now put this value of k^1 into the other equations. This equation has now become redundant and one is left with only 9 equations.

The equations are easier to solve if a metric coefficient is already known. Since g_{22} is already known and we want to determine the rest of k^1, we can use (K_{12}) and use (K_{22}) to eliminate k^2. Thus we compare $(K_{12})_{,2}$ and $(K_{22})_{,1}$ in a manner reminiscent of the Gauss-Codazzi equations. This procedure yields

$$\frac{A(t,\theta,\phi)_{\theta\theta}}{A(t,\theta,\phi)} = -(1 + \frac{1}{2} r\lambda') e^{-\lambda/2} = -\alpha \ . \tag{8.25}$$

There are now three cases: (1) $\alpha > 0$; (2) $\alpha = 0$; and (3) $\alpha < 0$. To get the same k^1 for $\theta = 2\pi$ as for $\theta = 0$, case (3) is eliminated and in case (2) the linear dependence of A on θ is eliminated. In case (1), putting $\alpha = a^2$, we get

$$A(t, \theta, \phi) = A_1(t, \phi) \cos a\theta + A_2(t, \phi) \sin a\theta \ . \tag{8.26}$$

In case (2), $A(t, \theta, \phi) \equiv A(t, \phi)$.

For case (1), solving for the r-dependence, we get

$$e^{-\lambda(r)} = a^2 - \beta r^2 \ . \tag{8.27}$$

Once again we have three cases: (a) $\beta > 0$; (b) $\beta = 0$; (c) $\beta = 0$. We now proceed with case (1a).

The next step is to take on the time-dependence. Following the procedure used for θ we compare $(K_{00})_{,1}$ and $(K_{01})_{,0}$ and use K_{00} to eliminate the k^0. The result is

$$\frac{A_1(t, \phi)_{00} \cos a\theta + A_2(t, \phi)_{00} \sin a\theta}{A_1(t, \phi) \cos a\theta + A_2(t, \phi) \sin a\theta} = \frac{1}{2}[\nu''(a^2 - \beta r^2) + \beta\nu' r] = \gamma \ , \tag{8.28}$$

where γ is the new separation constant, leading once again to three cases: (i) $\gamma > 0$; (ii) $\gamma = 0$; and $\gamma < 0$. Taking case (1ai) and putting $\gamma = b^2$, the above equation yields the explicit t-dependence

$$\left.\begin{aligned} A_1(t, \phi) =& A_{11}(\phi) \cosh bt + A_{12}(\phi) \sinh bt \ , \\ A_2(t, \phi) =& A_{21}(\phi) \cosh bt + A_{22}(\phi) \sinh bt \ , \end{aligned}\right\} \tag{8.29}$$

and the explicit r-dependence

$$e^{\nu}(r) = (b^2/a^2\beta)(a^2 - \beta r^2) = (b^2/a^2\beta)e^{\lambda} \ . \tag{8.30}$$

We are now only left with the ϕ-dependence to be determined for k^1. Of course, we have to also determine the other Killing vector components. Using the two equations above and (K_{12}) and integrating with respect to θ we obtain

$$\left.\begin{aligned} k^2 =& -\frac{a^2 - \beta r^2}{ar}\Big[\{A_{11}(\phi) \cosh bt + A_{12}(\phi) \sinh bt\} \sin a\theta \\ & - \{A_{21}(\phi) \cosh bt + A_{22}(\phi) \sinh bt\} \cos a\theta\Big] + B_1(t, r, \phi) \ . \end{aligned}\right\} \tag{8.31}$$

Differentiating k^2 relative to r and comparing with (K_{02}) yields

$$\left.\begin{aligned} k^2 =& \left[\frac{\beta r}{b^2(a^2 - \beta r^2)}\right]^{1/2}\Big[\{A_{11}(\phi) \sinh bt + A_{12}(\phi) \cosh bt\} \cos a\theta \\ & + \{A_{21}(\phi) \sinh bt + A_{22}(\phi) \cosh bt\} \sin a\theta\Big] + B_2(t, r, \phi) \ . \end{aligned}\right\} \tag{8.32}$$

Differentiating the above values and comparing with (K_{02}), it is seen that B_1 and B_2 are functions of t and ϕ only. Now repeating the entire process for k^3 we get

$$\left.\begin{aligned} k^3 =& \frac{\sqrt{a^2 - \beta r^2}}{ar \sin^2 \theta}\Big[\{A_{11}(\phi) \cosh bt + A_{12}(\phi) \sinh bt\} \cos a\theta \\ & + \{A_{21}(\phi) \cosh bt + A_{22}(\phi) \sinh bt\} \sin a\theta\Big] + B_3(t, \theta, \phi) \ . \end{aligned}\right\} \tag{8.33}$$

It is clear that if $a \neq 1$ the trigonometric functions will lead to a deficit angle or an excess angle, so that even though the metric is in some sense spherically symmetric, it is not that the functions revert to the original value at the correct places. One can re-scale the θ in those arguments so that one goes from pole to pole by going from $\theta = 0$ to $\theta = \pi$, but in that case the integral of the solid angle will not be 4π. This type of situation arises for circular symmetry about an axis. In that case the deficit angle is visualised as a disc with a slice (as of a pie) excised, and then the edges glued together. This is a cone. These arise in "string solutions" of the Einstein equations, where there is a singularity on the axis. The size of the deficit gives the "strength" of the solution. As such, the spherical metrics talked of above have been called "string clouds", a grossly misleading name for them, as it makes one think of little pieces of strings stuck together in a ball, which will not have any of the geometrical properties of these metrics. We have two cases: $(*)a = 1$; $(\dagger)a \neq 1$. I will not go further with the latter and so we proceed with case (1ai$*$).

What remains is to run all the components and coefficients through the mill of the $(K_{\mu 3})$ equations. Skipping the steps this time, we arrive at an explicit metric with $b = \Lambda = 1/R^2$ the metric is the de Sitter Universe model given in Eq. (7.34). There are 10 corresponding symmetry generators, so this is a space of constant curvature. Obviously, the radius of curvature is $1/R$. I will not waste space giving the explicit expressions but leave it at the fact that the ϕ-dependence involves arbitrary constants times $\cos\phi$ and $\sin\phi$ and the isometry group is $SO(1,4)$. If we had taken $b = -\Lambda$, the term $(1 - r^2/R^2)$ would become $(1 + r^2/R^2)$, and we get the anti-de Sitter Universe model, which obviously corresponds to a negative cosmological constant. The difference made to the symmetries is that the isometry group is now $SO(2,3)$.

The complete classification is obtained by going through all possibilities (including the two signs of b for example). In the first paper on complete classification [70], the full process of eliminating *all* possibilities was *not* carried out and some cases were missed (as pointed out by Stephani and acknowledged by us [71]). The first *correct* use of the procedure was for all spherically symmetric metrics [72]. Others followed.

The classification by isometries does not stop here. There is an obvious generalization arising — extend to incorporate the conformal transformations. (Leave mathematicians with *anything* and they will immediately go and generalise it. As a mathematician I love to do so. As a physicist I ask what use the generalization is. As a student I would probably have said "Can't you leave well enough alone? I already have as much as I can handle.") The physical significance of this generalization is obvious: the conformal curvature tensor gives the gravitational field. Also, for the Friedmann model, for example, the Universe expands. Thus the spatial part of the metric varies with time and so there is no time translational invariance and hence energy is not conserved. However, under the re-scaling supplied by the scale factor, we would expect to be able to get the energy scale invariantly. Requiring Lie transport invariance for the conformally transformed metric, we get the *conformal Killing equation*

$$k_{(\mu;\nu)} = f(x^\rho) g_{\mu\nu} \ . \tag{8.34}$$

where $f(x^\rho)$ is the natural logarithm of the conformal factor. The Friedmann metric has a conformal Killing vector, even though it does not have a Killing vector.

One could now ask for a complete classification by conformal Killing vectors by using the above procedure for this equation. A special case of these vectors is if $f(x^\rho)$ is a constant. Such vectors are called *homotheties*. Classification by homotheties have also been pursued. Since mathematicians *have* got into the game, the classification industry does not stop here. One can ask for invariance under Lie transport of any other tensor. These are called *collineations*. Thus, for the geometric tensors we have curvature collineations for the Riemann tensor, Ricci collineations for the Ricci tensor and Weyl collineations for the Weyl tensor. For all of these one has a conformal version as well. Of course, for the Weyl tensor with one upper and three lower indices, the conformal version has to be identical with the original. However, if we take the fully covariant Weyl tensor, there will be a difference. This brings home the point that these symmetries depend on the valence of the tensor. Complete classifications can be sought for each of these. It does not stop here, one can also ask for matter collineations by requiring that the stress-energy tensor be invariant under Lie transport.

I am not going into the classification problem further here. I would just like to make one point about them. Unlike the metric tensor, which has to have the determinant non-zero, the same would not apply to the other tensors. The easiest to see is the Ricci tensor or the stress-energy tensor. We can have a curved spacetime that is Ricci flat, like the Schwarzschild. In that case the equivalent of the Killing equations would be satisfied for *all* choices of vectors. As such, we have an infinite dimensional Lie algebra for them. In case the Ricci tensor is non-zero but has no time-time or time-space component, and there is no time dependence in the Ricci tensor, the time component of the collineation vector will be an arbitrary function of time. Such degenerate cases arise in collineations but not in isometries.

8.3 More on Gravitational Waves

I left the discussion on gravitational waves, in Chapter 5, very short. I need to take the theoretical developments a bit further and then go on to discuss the observational aspects somewhat. The subject has advanced too much for me to attempt a comprehensive discussion of it and I will still only provide a very basic introduction to it.

In the late 1960's Penrose had come up with a topological construction of an impulsive plane gravitational wave [73]. He took Minkowski space and at a light cone cut it and rejoined the space with a discontinuity. Thus there is a delta function singularity, which represents the gravitational wave. The space before it and after it is flat so it is like a shock wave. Next, he and Khursheed Alam Khan [74] gave a solution of the Einstein field equations that represented colliding impulsive plane gravitational waves, undergoing a head-on collision. (When I say "gave" I mean *gave*. Khan was given the problem of deriving the solution for his PhD. *He never derived it!* He simply arrived at the result which could be verified to be correct; but neither he, nor anyone else, could figure out how it came.) The solution comes in four parts, as it were: the time and place before either gravitational wave arrives; two where one has gone and the other has not arrived; and one where they have collided, see Fig. 8.2. The combined

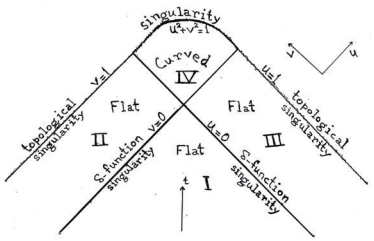


Figure 8.2: Khan-Penrose colliding plane gravitational waves. The spacetime is broken into 4 regions by the δ-function shock waves: I; II; III; and IV; of which the first three are flat and the fourth is curved. A curvature singularity develops at $u^2 + v^2 = 1$, and there are topological singularities at $u = 1$ and $v = 1$. In this toy Universe, doomsday comes without a warning at $u = 1$ and $v = 1$ in regions II and III, and with a warning in region IV.

solution can be written in terms of the Heaviside step (θ)-functions:

$$\left.\begin{aligned} ds^2 =& \frac{2t^3}{rw(pq+rw)^2} dudv - t^2 \left(\frac{r+q}{r-q}\right)\left(\frac{w+p}{w-p}\right) dx^2 \\ & - t^2 \left(\frac{r-q}{r+q}\right)\left(\frac{w-p}{w+p}\right) dy^2 \ , \end{aligned}\right\} \tag{8.35}$$

where u and v are the null coordinates for a wave moving in the positive and negative directions respectively, and

$$\left.\begin{aligned} p &= u\theta(u) \ , & q &= v\theta(v) \ , \\ r &= \sqrt{1-p^2} \ , & w &= \sqrt{1-q^2} \ , \\ t &= \sqrt{1-p^2-q^2} \ . \end{aligned}\right\} \tag{8.36}$$

This is the only new exact gravitational wave solution of Einstein's field equations known. It is interesting to note that in this solution curvature develops not due to the presence of any matter but just due to waves that leave the spacetime flat in themselves. Not only does curvature develop, but it leads to a crushing singularity at $u^2 + v^2 = 1$. What is more, the singularity then spreads into the *past!* Though region II is flat, the presence in the spacetime of the other gravitational wave, *which it will encounter later*, causes a singularity in that region, where the second gravitational wave never entered. In this toy Universe consisting of the two waves, there is a doomsday. The doomsday comes with a warning in the region where they have collided. However, it comes without a warning in regions II and III. The singularity there is not a curvature singularity but a topological one. It is worth also noting the mathematical aspect of this solution that the manifold is C^{-2} in the sense that its second integral would be a continuous function. Thus it is in the sense of generalized functions that the manifold can be dealt with.

Let us now turn to the observational aspects. Since the gravitational force is very weak, only 10^{-38} of the electromagnetic force (as measured by the ratio of the gravitational force acting in the atom, divided by the electromagnetic force, it becomes extremely difficult to observe them. The earliest suggestion and implementation for actually detecting them came from Joe Weber in a series

of papers [75], accompanied by actual experiments. The idea was to provide a massive cylinder that would interact with the waves and be "rung like a bell". The extremely slight vibrations were to be detected by a piezoelectric crystal resonating at 1660 Hz. He used an aluminium bar of 2m and diameter 0.95m suspended by a wire and enclosed in a shielding from acoustic and electromagnetic waves, so that there would be no contamination of the signal by spurious vibrations in the surroundings. To further eliminate random signals in the environment mimicking the gravitational wave, he looked for a coincidence of 3 detectors, placed at the vertices of an equilateral triangle, with a seismometer placed at the centre of the triangle produced by them to look for seismic signals that could set up vibrations in the three bars as the same time.

He claimed to detect gravitational waves coming from the core of the galaxy at a 24 hour period corresponding to the the rotation of the Earth. It was difficult to believe that from all the modes of vibration available in gravitational waves, only *one* would be excited for the bar pointing in the direction of the source. It was pointed out that gravitational waves would not be affected at all by the Earth's bulk, but would pass through it much more easily than neutrinos do. He then claimed that he was, indeed, getting a 12 hour period for the signals. First no one tried to replicate his experiment but later, when they did, no one saw any signals. As such, no one believed him. However, he insisted that he *was* getting the signals. At this unsatisfactory stage it was pointed out that thermal vibrations of the bar set a lower limit on the amplitude of the vibrations from the waves, as several stars would have to be annihilated *per second*. The store would be depleted in a few years.

The natural solution was to try to reduce the temperature of the bar. As such "cryogenic bars" were used. Of course it became much more difficult (and costly) to keep the bars at sufficiently low temperatures. Though this was the mainstream attempt, there *had* been another suggestion that was also being followed up. Bertotti and Carr [76] had proposed using laser ranging of spacecraft and looking for frequency shifts due to the path difference caused by the fluctuation in the metric by the gravitational wave, as such changes could be measured with great precision. Of course, the fluctuations in the time component of the metric could also be used by sufficiently accurate clocks. The possibility of detecting gravitational waves from the Big Bang by time measurement fluctuations was explored [77] but turned out to require much higher precision than could be obtained by atomic clocks. Later it was suggested [78] that one could try to use pulsars to detect a stochastic gravitational wave background. So far, none of these attempts has led to a conclusive detection of gravitational waves. The path difference could also be looked for using lasers in a sufficiently large Michelson-Morley interferometer and this seemed to be a very promising approach.

At a Symposium on Experimental Gravitation held at Nathiagali in Pakistan in 1993 [79], there were talks on the original bar detector, cryogenic bar detectors and laser interferometer detectors. Weber claimed to have developed a theory of Quantum Gravity according to which the signals from any event would be much larger than he had originally calculated. However, he did not share the theory with us. No one was ready to take his claim seriously, but M. Bassan who happened to be present admitted that he had been visiting Weber in 1987 and an event *was* noted in his laboratory a bit *before* news of the Supernova of 1987 came through. He had no explanation for this coincidence but did not believe

that the room temperature bar had seen gravitational waves. The matter was discussed at the conclusion of the Symposium and it was agreed that we needed the laser interferometer to resolve Weber's dispute, as his claimed theory would make no difference to this class of experiments in any case. Finally, it was this method that *did* work, but the way it worked requires that I explain about collapsed stars and black holes.

8.4 Collapsed Stars and Black Holes

One might wonder what holds a star up. I don't mean "up in the sky", but in the sense of maintaining the size of the star. After all gravity should cause it to go on collapsing. If the star were made of a solid, one could say it was the strength of the material of the solid. However, that would easily collapse under its own weight fairly soon. In any case, the star is essentially a gas. What holds *that* up? For liquids it would be the hydrostatic pressure. In fact, the same would apply to the gas at the scale of a star. This would be enough to hold up a planet mass object, even up to a gas giant. But for a mass more than that it should collapse under its own weight.

How far can it go on collapsing? Once the force of gravity exceeds the Coulomb repulsion between nuclei, classically there is nothing stopping an endless collapse. The nonlinearity of GR would only make matters worse, accelerating the collapse. Of course, GR is not the only non-classical theory to be involved. The gas consists of nuclei and *electrons*. When the star is squeezed enough, the electrons would no longer belong to any given molecule but would be shared between all the nuclei. Since they are fermions, they would constitute a Fermi degenerate gas, which would be held up by the Fermi pressure. Essentially, this is Pauli's exclusion principle at work, ensuring that no two Fermions occupy the same quantum state. (Think of it as a moral police that has banned cohabitation between electrons.) For a sufficiently low temperature there is *total* degeneracy, but at higher temperatures there is partial degeneracy.

In 1930 there was an outstanding problem of explaining the spectrum of a faint star near the bright star Sirius. It emitted white light, which should mean that it was extremely hot, but it was very dim. That could be explained by assuming that it was very far away, but it was seen to be gravitationally bound to Sirius, which is one of the closest stars to the Sun. The only explanation was that it was inordinately hot with a mass more than that of the Sun and the size of the Earth. This star was called a "white dwarf". According to the understanding of the time such objects could not exist. Subrahmanyan Chandrasekhar had just obtained his BA in India at the age of 20, having already published a paper at the age of 19 and went to read for his PhD with R.H. Fowler at Cambridge. On the ship journey from India to England, he proposed the theory of partial degeneracy for stars, which provided a solution to the problem of the white dwarf. Essentially, the star would be a giant molecule with the nuclei sharing a sea of electrons. The work was completed for his PhD with Fowler. When he presented it at a meeting of the Royal Society, Eddington tore it to shreds. It gradually came to be accepted despite Eddington and he (along with Fowler) was awarded the Nobel Prize in 1983, more than 50 years after his work! (One wonders how much his race had to do with the delay. Bear in mind that the Jew, Albert Einstein, never got the Nobel Prize for his work on Relativity.) I

will not, here, go into the subject but it is explained in Chandrasekhar's book [80].

Chandrasekhar [81] considered a star of sufficient mass to overcome the Fermi degeneracy pressure of electrons. Then the electrons would get squeezed into the nuclei till the Fermi degeneracy pressure of neutrons stops it from collapsing further. The mass would be about one and a half times that of the Sun and the radius about 10 km. The same result was arrived at independently by Lev Landau [82]. More precisely, there would be equilibrium between protons, electrons and neutrons, each providing its separate partial pressures. At the core there would even be strange particles (Σ, Λ, etc.). They did not go much further because it was thought that these objects could not be seen, even if they existed. Of course, as Wheeler used to say "The Universe is not only stranger than we imagine, it is stranger than we *can* imagine". We have repeatedly seen the invisible.

In 1967 a young PhD student reading for her PhD at Cambridge with Anthony Hewish, Jocelyn Stuart Bell, had been given a routine type of project examining radio signals from stars. She noticed a sort of smudge where there should have none and took it to her supervisor. He told her to ignore it and get on with the work she was supposed to be doing. However, she was intrigued and managed to clear it up a bit and took it back to her supervisor, who now got annoyed and told her not to waste the time she was paid for on this matte. As such, she followed it up on her own time and when she had resolved the smudge into regular pulses of around 30 ms period coming from the direction of the Crab nebula, she took it back to her supervisor. He finally took it up and they published two papers, the second one about another source, (with other authors brought in) in the journal *Nature* [95]. For Bell's discovery her *supervisor*, Hewish, was awarded the Nobel Prize!?! You may well ask "How come?" Notice that the authorship did not put her as first author, despite the fact that she would be first in alphabetic ordering apart from being the person who had actually discovered the object, and pursued it in the teeth of the opposition of the man who would be awarded the Prize. One wonders how much her sex had to do with the Nobel Prize ignoring her and how much is to be ascribed to her supervisor's greed for the Prize. (Bear in mind Vela Rubin — the woman the Nobel Prize forgot). One wonders how the Nobel Prize will sit on Hewish's conscience on his deathbed.

Because of the regularity of the pulses it was first thought that they may be signals from an extraterrestrial civilization. Since the science fiction of the time visualised aliens as "little green men", these objects were first dubbed LGMs. Once properly located, it became clear that these objects were radiating far more energy than any civilization could *produce* (leave alone the fact that if they were civilised they would not waste energy more profligately than the Americans). They were then called *pulsars*. An adequate mechanism for the observed radiation was needed. For astronomical objects regularity is associated with revolution or rotation. The extremely short duration of the pulse excluded revolution. As such, it had to be by rotation. Even then, a white dwarf would have to rotate faster than light to produce such short pulses. These objects had to be $\sim$ 10s of km in radius. Tommy Gold [84] suggested that these could be rotating neutron stars with high magnetic dipole moment spinning obliquely to the magnetic axis. The result would be a beam of pulses with the rotation period as the pulse period, along the line where the planes perpendicular to

the axes lay. This exactly fitted the observations. Hence the observed pulsars are generally referred to by the model for them as *neutron stars*. More than 2600 pulsars have been found by mid 2019, with periods ranging from $\sim$ 1 ms to a few seconds and masses ranging from about 1.4 $M_\odot$ to about 2.3 $M_\odot$ (though such high masses are difficult to square with our understanding of these objects). Their magnetic fields vary from $\sim 10^{12}$ Gauss to a thousand times that! Of course, there could be neutron stars that are not pulsars, because their magnetic axes have got aligned with their rotational axes. Such non-radiating neutron stars would be what Chandrasekhar and Landau had felt could not be seen.

The white dwarf and neutron star are called "collapsed stars" or "degenerate stars" for obvious reasons. The term "collapsed object" is also used to include the third category of extremely small but very massive objects: namely black holes, which I will discuss now. Of course, I already devoted a whole chapter (#6) to them, but only as mathematical objects and not as the type of thing that could be observed. Their biggest problem was that they had a singularity at the core, a region of infinite energy density, and "that was impossible". As such, some new Physics must take over at that stage. The natural candidate for the new Physics would be a theory incorporating Quantum Field Theory and GR, which is the topic of the next section. However, whatever happens at the core, there is a huge step between the outer part of the black hole and the core. There seemed no good reason that they should not form at all. Their next problem was that, "by definition, they can never be seen". (One has to be wary of such negative statements. We have repeatedly seen the invisible.) They have a gravitational field and if such a field is found where there is no visible source it would have to be a collapsed object that has ceased to radiate. There are limits on the masses of collapsed stars. If the source is more massive than those limits, it must be a black hole.

Chandrasekhar had derived an upper limit for the white dwarf mass: 0.7 $M_\odot$ if it is fully degenerate; and $1.4M_\odot$ if it is partially degenerate. This is called the Chandrasekhar limit. It is for this reason that we take neutron stars to be more massive than 1.4 $M_\odot$ — they cannot be white dwarfs. The argument for white dwarf mass limits could be extended to neutron stars by replacing the electron mass by the neutron mass. Here we would need to use an equation of state for nuclear matter to get the partially degenerate limit. Due to uncertainties in the nuclear physics involved, this was not easily obtainable. The rough estimate came out at about 2 $M_\odot$. However, one cannot then say that if it is too massive to be a neutron star, it must be a black hole. After all, who can say that there cannot be some other particles (say quarks) that could provide still larger masses.

Since GR identifies mass and energy, Rhoades and Ruffini argued [85] that the energy required to "hold up" any star (including degeneracy energy) would contribute to the mass causing the star to collapse. This is similar to trying to provide steel reinforcement to hold up an enormously high skyscraper and finding that the building gets dragged down by the weight of the reinforcement itself. Taking into account the nonlinearity of GR, they obtained a mass limit of 3.2 $M_\odot$ for *any* star to become a black hole. This was the limit arrived at independently by Fang Li Zhi in China (referred to in the dedication). With the study of masses of X-ray sources [86], this led of the first identification of a black hole as the X-ray binary source, Cygnus X-1 [87]. The fact of it being a

binary with a normal star was crucial to the identification, as the signals from the black hole were being occluded by the normal star while it was behind the normal star. How did signals emerge from the black hole? Actually they did not. Matter falling into the horizon of the hole was accelerated enough near it, to emit a characteristic spectrum going into the X-ray.

In 1974 an object of interest was seen towards the constellation Sagittarius and was called Sgr A*. It turned out to be at the core of our galaxy, the Milky Way, some 2.564×10^4 light years away. Over time it came to be accepted that this was a black hole — not a little 50 km radius black hole like Cygnus X-1, but a giant of about 2.2×10^7 km radius, and much more massive. With rapidly improving space-borne telescopes one started to see stars in orbit around a gravitational source that could not be seen. One of the stars is a 15 $M_\odot$ star that completes an orbit in 12 years. There are other very bright stars as well. Solving the Kepler problem for the system, it is found that the mass is about 4×10^6 $M_\odot$. It has since been found that there are such *supermassive* black hole holes at the cores of all large galaxies. They vary from some million to some billion $M_\odot$. Black holes like Cygnus X-1 are called *solar mass* black holes, even though they are all from 5 to 50 $M_\odot$, and not 1 $M_\odot$. These are what might be called *astrophysical black holes*. They are *seen*, and so their existence is no more open to debate than that of pulsars or quasars. However, nothing can be seen inside the event horizon, so there is no way to know whether there is a region of infinite energy density inside or not.

We know how solar mass black holes arise. Very large stars go supernova (explode) in such a way that while the bulk of the mass in the outer regions is ejected, they leave behind a core more massive than the black hole mass (lower) limit. Or perhaps white dwarfs or neutron stars accrete so much matter that they literally cannot digest it, and they collapse (implode). The question arises, how do the supermassive black holes arise? Were there supermassive giant stars that exploded? Stars of masses $\sim 10^5$ $M_\odot$ have been postulated as the very first stars to form. They are supposed to have been composed of pure molecular hydrogen and burned furiously for a span brief on the astronomical scale, exploding in some 10^8 years. If they did exist, they are unlikely to have been able to produce any collapsed object, but only provided other elements that went on to be used in the next generation of stars and planetary systems. Where did the supermassive black holes come from? For that matter, why is there such an enormous gap of masses between solar mass and supermassive black holes with no "intermediate mass" black holes?

The suggestion is that there were a lot of solar mass black holes produced in the earlier stages of the Universe and they suffered occasional collisions, in which they often merged. With continuing mergers in the core of the galaxy, the supermassive black hole would have gradually grown. However, it is not so easy to see that there would be enough time to build the supermassive black holes by this method. Anyhow, a supermassive black hole of 8×10^8 $M_\odot$ has been seen, that was there when the Universe was only 6.9×10^8 years old! Where did that monster ever come from? How would there have been time for it to form by coalescence? One suggestion is that black holes may have formed primordially and been coalescing in the very early Universe, when distances were very small anyhow and the number of black holes in "striking range" may have been very large. Then they could coalesce to form the supermassive black holes that we see.

Fresh evidence for the existence of black holes was announced on the 11$^{\text{th}}$ of April, 2019. This is on account of strong bending of light by the black hole. Recall that at $1.5r_s$ light goes into a circular orbit about a Schwarzschild black hole. As such, slightly further out it could bend right around and go back in the direction it came from. Holz and Wheeler [88] called this "retro-lensing" and suggested that we could literally see black holes (one and a half times enlarged) by this method. There would be a succession of rings around the black hole for paths going once around, twice around and so on. They called the resulting haloes around the black hole "glories". If it is a Kerr (spinning) black hole, the ring gets flattened out on the side where the hole spins towards the observer and bulges out on the side rotating away from the observer. Obviously, the dented side is blue-shifted and the bulging side red-shifted. What was seen was a spinning black hole. This, though, is *not* what was seen by the Event Horizon Telescope (EHT). The problem is that the black hole has an accretion disk of matter around it, and that gets lighted up a lot more than the faint halo would be. As such, it is the accretion disc, considerably further out than the event horizon, that is seen. However, it is very much closer than any orbiting stars reach. Further, one does not *see* the black hole, but only deduces its existence by the absence of any visible gravitational source to bend the paths of the stars. With strong bending one sees an outline, albeit considerably further out than the event horizon.

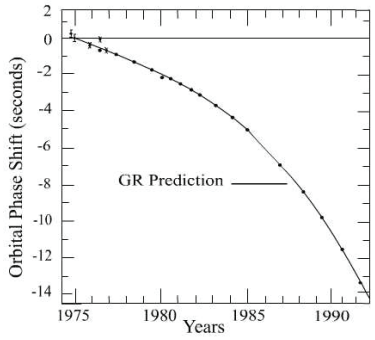


Figure 8.3: The observations of the Hulse-Taylor binary pulsar, by Taylor, have followed exactly the curve predicted by GR. No other theory comes close to such a remarkable fit. Notice that the error bars of observation, visible till 1978, have disappeared as they are too small to be displayed. This is one of the most dramatic and precise tests of GR. (Image taken from the web.)

The first piece of concrete evidence for gravitational waves, albeit indirect, came from the Hulse-Taylor binary pulsar mentioned at the end of Chapter 4. Russell A. Hulse was a PhD student of Joseph H. Taylor Jr. In 1974 they discovered a binary pair of neutron stars [89], now identified by the unedifying label PSR 1913+16. As mentioned before, this provided the best laboratory for testing GR. In particular it was worked out what GR with gravitational waves would expect and what any other theory would predict for the rate of slowing down of the binary period. The result for the period after which the observations were stopped are given in Fig. 8.3. The binary pair follow the choreography of GR, ignoring all dance steps that other theories might try to

make them follow. Incidentally, in contrast to the behaviour of Hewish with Bell, despite the fact that it was Taylor who continued the observations and Hulse dropped out, Taylor ensured that Hulse shared the Nobel Prize of 1993 with him and had him as the first author of their joint paper.

Coalescence of black holes has been seen and provides direct evidence of gravitational waves. When two black holes of masses m_1 and m_2 merge, the mass of the resultant black hole, M, cannot be more than the sum of the two masses. At the same time, as mentioned before, Hawking had proved that the area of a black hole always increases. Thus, if two black holes merge, the area of the resultant black hole will be more than the sum of the areas of the two black hole. Since the area is proportional to the square of the radius, and the radius is proportional to the mass, the mass of the resulting black hole must be constrained by $m_1^2 + m_2^2 \leq M^2 \leq (m_1 + m_2)^2$. As they approach merger they will be highly accelerated and so must radiate. However, since the black holes have no electromagnetic fields associated with them, they cannot radiate electromagnetic energy. The only energy they can radiate is in the form of gravitational waves. The maximum amount available is obviously $\sqrt{2m_1m_2}c^2$. However, some of the energy could go into increasing the mass of the resultant black hole. There is no exact solution for the merging black holes and it is extremely difficult to work out the result of the merger by going to higher order approximations. These are provided by the "post-Newtonian approximation". A major development of the last century was the development of methods, algorithms and a code for "numerical relativity". Using this technology simulations for mergers of solar mass black holes were produced and the typical signal for the radiation from the merger was obtained. There is a slow, low level radiation produced in the earlier phases that could be reproduced by the post-Newtonian approximation and then a dramatic increase resulting in a catastrophic merger with large amounts of mass radiated as gravitational waves. (In the simulation, the "ring-down", as it is called, is quite thrilling to see.)

The best evidence for gravitational waves and black holes came in one packet. The Americans had set up a couple of 4 km $\times$ 4 km laser interferometers, LIGO (Laser Interferometer Gravitational wave Observatory), at Hanford, Washington and Livingston, Louisiana, about 2,100 km apart (provided one could go straight *through* the Earth and not be restricted to its surface). The purpose was to see a coincidence event between the two and, at the same time, to measure a time delay if there is one. The Europeans set up a 3 km $\times$ 3 km interferometer, Virgo. On the 15^{th} of September, 2015, the centenary of GR, LIGO had just come online when a signal was seen [90]. Unfortunately, Virgo, was not ready and so did not see the signal. Three months later [91], on the 25^{th} of December, a second signal was seen by LIGO and Virgo was still not on line. (The coincidences that the first signal was seen just as LIGO was ready, and that the second one came on Christmas day, led to various conspiracy theories.) In both cases the characteristic signal predicted was seen and fitting gave the masses, separation and distances of the two events.
1^{st} **event:** $m_1 = (35 \pm 5)M_\odot, m_2 = (29 \pm 4)M_\odot, M = (62 \pm 4)M_\odot, M_{gw} = 3M_\odot$, $a_{final}/(r_+c) = 0.67 \pm 0,06$, luminosity distance $(410 \pm 170)Mpc$;
2^{nd} **event:** $m_1 = (14.2 \pm 5)M_\odot, m_2 = (7.5 \pm 2.3)M_\odot, M = (22 \pm 4)M_\odot, M_{gw} = 1M_\odot$,
$a_{final}/(r_+c) = 0.74 \pm 0,06$, luminosity distance $(440 \pm 170)Mpc$.
(Actually, the errors for some masses are slightly more and asymmetrical in $\pm$.)

Improvements of LIGO were underway and the observatory was closed down soon after the 2^{nd} event. When it came back on-line Virgo was also functioning, albeit not with such fine resolution (on account of the shorter arms) a signal was seen that corresponded to a neutron star merger [92]. Since the neutron stars were seen by electromagnetic radiation the identification of the source and its distance could be firmly established. The event did not end with such a dramatic ring-down (having an energy of $M_{\odot}c^2/40$, but lasted a lot longer, ~ 100 sec. The further data that came in showed how the trans-iron elements are produced. Earlier it had not been clear that mergers would be the main source of gravitational waves that could be seen, as one had no idea of the frequency of such mergers. Now that a number of them have been seen, one has a much better idea of how many collapsed objects there are in our neighbourhood in the Universe. A space-borne observatory, eLISA (enhanced Laser Interferometer Antenna, is expected to be put up by 2040, which would have much finer resolution and could be used for proper gravitational wave astronomy, to the extent that it could see the inner workings of a supernova event. It should also be possible to use it to probe the early Universe.

8.5 Attempts to Unify Quantum Theory and GR

From the start of Quantum Theory and Relativity, there was tension between the two. The former seemed to require sudden jumps but the latter needed that processes be continuous. A sudden jump implied changes faster than light. Further, the sudden jumps had to be unpredictable for the former and everything needed to be deterministic for the latter. As with a human couple the tensions, came out in little things. P.A.M. Dirac arranged a marriage between the two and initially it seemed that the union was productive. Dirac's (special relativistic) Quantum Electro-Dynamics (QED) led to the Dirac equation with its concomitant electron spin and antiparticles. However, infinities arose in the theory. Further, the founder of both theories, Albert Einstein, and the self-proclaimed "defender of Quantum Theory", Neils Bohr, could not agree on the interpretation of the Quantum formalism. While both agreed that it "gave the right results", they could not agree on what those results mean. Somehow, time did not seem to mean in Quantum Theory what it did in Relativity.

Heisenberg tried to predict the outcomes of sub-atomic experiments by setting up an input-output table, with all possible outcomes included. Given the initial state and the final state, the probability of the outcome could be computed. This way of doing Mechanics was called "Matrix Mechanics". It led Heisenberg to formulate his uncertainty principle, that both position and momentum cannot be known precisely simultaneously, but the product of the uncertainties in the two, $\delta x.\delta p_x \geq \hbar/2$. Einstein was ready to concede that this was true for simple-minded experiments but contended that clever enough experiments could circumvent this limitation. As such, this could not be a fundamental principle. The view was developed by Bohr and Heisenberg that the state of a system during the experiment is unknowable and there is no meaning to ask for a description of it. The only thing that *can* be known is the prob ability of a given final outcome given an initial state. Once the experiment is performed, the probability gets converted to a certainty and "the wave-function collapses". Erwin Schrödinger propounded a thought experiment that led to a

paradoxical result that a cat in a box given cyanide with a probability of half, would be dead and alive till the box is opened and the cat seen to be dead or alive. Thus, as Wheeler put it, "No phenomenon is a phenomenon until it is an observed phenomenon". (This was reminiscent of the theistic belief that things existed because God observed them.) On this Einstein had said that he could not believe that the Moon existed because a mouse looked at it.

The Bohr-Einstein debates led Bohr to formulate various "principles" for Quantum Theory. Einstein had begun the theory with a dual picture of the quantum of electromagnetism. This was extended by Louis de Broglie to a dual picture of particles like electrons. One of Bohr's principles was the "complementarity principle", that in a given experiment a quantum entity behaves *either* as a particle *or* as a wave, but not both. Einstein believed that Heisenberg's probability could not be a fundamental requirement but only arose because of ignorance. With Max Born he had interpreted the wave function for an ensemble of quanta as giving the probability amplitude but, he said, "I, in any case, am convinced He does not play dice with the Universe", and in this context he also said "Subtle is the Lord. Malicious He is not." While Einstein tried to formulate thought experiments that would prove that the Quantum Theory was an incomplete description of Nature, Bohr went on countering the demonstrations and refining his philosophy of the theory. This culminated in a paper in 1935 by Einstein, Boris Podolsky and Nathan Rosen [93], in which they tried to state what would be meant by "physical reality" and demonstrating that it was not fully described by Quantum Theory. Their experiment was to take a composite system breaking into two constituents. By measuring the position of one of them precisely and the momentum of the other, using the standard conservation laws, they said that both quantities could be simultaneously determined. The process of measurement of one of the quantities for one of the parts, could not be communicated to the other part without violating the speed of light limit. Bohr did not provide a sound critique of the idea, but W.H. Furry did in 1936. He argued that the uncertainty principle vitiated the determination, as the momentum is a continuous variable.

From 1936 to 1957 the problem was shelved till David Bohm and Yakir Aharanov [94] revived it with a remarkably simple resolution of Furry's objection. Take the composite system to have zero spin and break it into two parts of opposite spin. *Now* the variable is not continuous. Either the spin would be carried along with each part, as a *local hidden variable* (as it is now called) a'la EPR, or it would be created on measurement a'la Bohr and Co. However, it was not quite so simple to see how to actually get a difference between the two to show up. For this purpose John Stuart Bell [95] developed a set of inequalities in 1961 based on considerations of probabilities of correlations of spin vectors of the broken system. It must be admitted that both Bell and Bohm were quite sure that the EPR argument would be demonstrated. Equally, the majority of working physicists were quite sure that Bohr and Co. would prevail. In the event, when Aspect et al. actually performed the experiment in 1981 [96] the result completely contradicted the EPR expectation. People claimed that this proved Bohr's point. However, Bohr had actually said that the proposed experiment was meaningless as the question asked was meaningless. As such, Aspect's experiment *disproved both*.

Wheeler [97] brought the problem of the concept of time into focus with his variant of Young's double slit experiment. He considered sending a single

quantum entity through a double slit and placing the screen on which it was to be received sufficiently far away that the experimenter could choose how to observe the quantum, as a particle or a wave, *after* it had gone through the double slit. The wave is seen on the screen if there is no recoil on it (which would measure the momentum and hence see a particle). If the quantum is to be seen as a wave, it should have gone through both slits. However, if it is to be seen as a particle, it could only go through one slit. By making a choice *later*, he said, you could change what happened *earlier*. The experiment has been performed with the result anticipated by Wheeler [98]. Clearly, for Quantum Theory, "past", "present" and "future" have different meanings than they do for Relativity. Sending single quanta through a double slit builds up an interference pattern. This would imply interference of the quanta *over time*, which would be absurd in relativistic terms.

Most people talking of the problem of putting Quantum Theory and Relativity together, do not even mention this problem. The reason is that most people regard GR as a field theory of gravity, and so talk of the problem of "quantizing gravity". It is taken as if Special Relativity has no conflict with Quantum Theory. As I have tried to point out, the problem is far deeper, as is apparent when one regards both of them as theories of motion: one takes it as smooth; and the other as quirky and jerky. The fact is that space and time are present in the language for *both* theories and the two do not agree what they mean by them. (The problem reminds one of Churchill's talking of America and Britain divided by a common language.) One could equally well talk of the difference of how they view *space*. What appears "near" or "far", or "left" or "right", may depend on the quanta involved. Having made this point, let me shift to the more usual discussion of the problem.

There was another unification that Einstein wanted — of Electromagnetism and Gravitation. Einstein did not believe that the nuclear forces were fundamental and expected that they would be effective consequences of the strong curvature of spacetime due to charges. Whereas Electromagnetism had been at the base of his Special Theory it was not incorporated in his General Theory. Since Gravity was contained in Geometry and Electromagnetism was not, there was a fundamental dichotomy that Einstein believed could not be there in Nature. In 1921 Theodor Kaluza [99] suggested enlarging the spacetime to 5 dimensions to incorporate Electromagnetism. The idea is that the 10 component metric tensor of 4 dimensions is fully used for the gravitational potential. If Electromagnetism is to be incorporated into Geometry, there is need to enlarge the space by adding a 5^{th} dimension. It turned out that the 5-dimensional Einstein equations incorporated the Maxwell equations as well. Einstein liked the idea at first, but later concluded that it did not do what was needed physically. Einstein went on trying to find a "unified field theory" to the end of his life. He even tried going to a complex four dimensional space by putting the Maxwell tensor as the complex part of the metric tensor (an extremely bad idea not worthy of Einstein, as it mixed the *potential* of Gravity with the *field* of Electromagnetism.) There is a further problem with the attempt, on account of the one extra slot available in the metric tensor, that must be filled with an additional term leading to an extra field equation [100]. It becomes necessary to take that into account and it leads to a prediction that does not work. Going to higher dimensions to incorporate the weak and strong nuclear forces avoids this problem as the constraint for the one extra term is no longer there. One

of Einstein's concerns was that one could not see the 5^{th} dimension. Oskar Klein [101] proposed that the extra dimension could be curled up into a very small size and there would be an isometry in that direction, so that it could not be seen. This idea became crucial for the higher dimensional reincarnations of Kaluza-Klein theories.

Going back to the problem of quantization of gravity, Dirac had developed a canonical procedure to convert any classical field theory to a quantum version. This was called "quantizing the field". He used this procedure to "quantize electrodynamics", giving QED. However, as mentioned above, the theory gave meaningless answers as the wave function diverged, yielding infinite probability amplitudes for occurrences. Dirac regarded this as a fundamental shortcoming of his theory. It was pointed out that the wave function, giving the quantum state is described by a normalized vector in a Hilbert space. As such, by definition, if the wave function changed it would have to be renormalized. To calculate results cut-offs at low and high energies that would go to zero and infinity had to be inserted. Before taking the limit for the cut-off, the numerator and denominator leading terms could be taken and the cut-off cancelled. Dirac was not impressed. He said that one can throw away small numbers compared with big numbers, but throwing away infinity is bad Mathematics. Nevertheless, the results led to testable predictions which were verified to high accuracy. Einstein had hoped that the singularities of GR and of Quantum Theory would take care of each other if one developed a General Relativistic Quantum Theory. Dirac tried to use his canonical quantization procedure for GR but it did not work.

The problem came from the fact that Quantum Field Theory needs to use equal time commutation relations for creation and annihilation operators on the same spacelike hypersurfaces. Relativity requires that time be like space, and so singling out one coordinate from the others is not valid. For Special Relativity, since the spacelike hypersurfaces are flat, the transformations are simply complex rotations and the quantization procedure carries over smoothly from one class of hypersurfaces to another. When the spacetime is curved this is no longer the case. It was possible that there was too much freedom of choice of selecting a foliation procedure. Maybe, Dirac's simple choice was not good enough. Arnowitt, Deser and Misner provided a canonical Hamiltonian formulation of GR [31]. This could be used for the attempted "canonical quantization". When it did not work, it was argued that this may be due to the fact that when the spacetime was broken into space and time, it was not manifestly covariant. Thus one of the attempts made was for "covariant quantization" [32]. The problem remained.

On the basis of a general argument that if the renormalization procedure is to work, it will be required that as one goes on applying it, the results continue to lead back to the starting value after re-scaling. This means that there will be invariance under the renormalization procedure, thus providing us with a group. This is the renormalization group. An equation was formulated to check if the renormalization group is applicable. This is the renormalization group equation. A classical field theory is renormalizable if it satisfies the renormalization group equation and a quantum version of the theory can be constructed by using that procedure. Gerard 't Hooft and Martinus Veltman [102] were awarded the 1999 Nobel Prize for their work that determines if a theory can be renormalized. Unless the theory gives finite answers to start with, it must be renormalizable. The usual 4-dimensional GR is neither finite nor renormalizable. *That* is the

problem.

Instead of relying on the old procedures one could try to take a more fundamental approach. Wess and Zumino had proposed that there may be a symmetry between fermions (particles of spin $(2n+1)\hbar/2$ and bosons (particles of spin $n\hbar$). They called this "supersymmetry". The theory allowed an anomalous term to arise in the calculations. Gervais and Sakita [103] showed that if reality is 26 dimensional and one assumes there are no point particles but little bosonic strings, the anomalous term would be removed. This change of the view of matter and energy as consisting of strings instead of particles would obviously be of great significance for GR. The question arises how one could mistake strings for particles. For this purpose two old ideas were appealed to. Planck had pointed out that one could construct quantities of dimensions of length, time and mass from the fundamental constants of Nature, $c, G, \hbar$: namely $l_P = \sqrt{G\hbar/c^3}, t_P = \sqrt{G\hbar/c^5}, m_P = \sqrt{c\hbar/G}$. At these dimensions one could not take the classical limit for Quantum Theory or GR. Putting in the values one finds that $l_P = 1.6 \times 10^{-32}$ cm, $t_P = 5.4 \times 10^{-43}$ s and $m_P = 2.2 \times 10{-}5$ g. The Superstring was taken to be of Planck length and the Kaluza-Klein argument was used to explain why one sees strings as if they are particles.

In the usual supersymmetry (called SUSY) it was shown that the infinities of Quantum Field Theory for bosons and fermions cancelled when applied to gravity, at least at the lowest order of calculation. The quantum of gravity was dubbed a "graviton" and was then treated as a spin $2\hbar$ (bosonic) field that would be massless (so that the gravitational force should be long range). The corresponding fermion, that would provide the cancellation, was called a "gravitino" (small graviton in the Italian construction). It would be a spin $3\hbar/2$ that would have a very large mass. This theory was called "supergravity" (SUGRA). Various unitary SUSY groups were explored for providing the "correct" SUGRA theory. It turned out that when the calculation was taken to the next order of interaction the infinities returned with a vengeance. Since the infinities for the graviton had been cancelled by one gravitino, a second gravitino could be used to cancel the infinities at the next order of calculation. This was called an "extended SUSY". For any $SU(n)$ SUSY, the extended group was denoted by $SU(n|2)$. At the next order three "gravitini" were needed. The hope was that for some choice of basic SUSY there would be an extension that would cancel all infinities at further levels of calculation. In fact there was great hope that the ultimate theory had been obtained with $SU(8|8)$ SUGRA in an 11 dimensional spacetime, using Kaluza-Klein reduction of dimensions [104]. In fact, this theory was touted as "the theory of everything", but it turned out that there were various problems with it.

A major problem with Superstring theory was that it dealt only with bosons and, at the fundamental level, these gave interactions (forces) while fermions seemed to be what would normally be regarded as matter. Besides, one needed to cancel the infinities of bosons with those of fermions. What was needed was a theory that contained both. This was provided by Green and Schwarz [105, 106] with the basic group $SO(32)$ in 10 dimensions. Once again, this was claimed to be a theory of everything. Superstrings are supposed to be in a 1+1 dimensional spacetime in one sense and then in a 26 or 10 dimensional spacetime in another sense. The new Superstrings were regarded as being the basic unit and the "particles" as vibrational modes of the Superstring. Among the modes is the spin $2\hbar$ boson and its "super-partner" a spin $3\hbar/2$ fermion. These would be a

graviton and a gravitino. There is also a scalar (spin 0) mode that is called a dilaton. Further, there is an unwanted "tachyonic" mode, The dilaton, though not really desirable is not unwanted like the tachyon, as that violates causality. The theory was later subsumed into a theory of super-membranes and that led to M-theory which claimed that through some "duality relations" the various most popular theories were all essentially one and the same, and that the 11-d SUGRA is also the same. Various experiments have been proposed from time to time to test these theories, *provided some conditions hold*. For example, if there are some larger extra dimensions, one may be able to see them. None has led to a valid prediction. That could be because the theory is essentially right but that condition does not hold, or because the theory itself is wrong. "Tests" of this type are not very useful.

Somehow the belief arose that since both Quantum Theory and GR play a role at Planck scale, Quantum Gravity *only* becomes important at that scale. There is absolutely no reason to believe that. We only know that *by* that scale we need Quantum Gravity. However, it could arise at a much earlier stage. As a *reductio ad absurdum*, I know that I cannot lift the Earth. However, that does not mean that I *can*, then, lift the Moon. A variant of this belief is that since this is the one natural scale, it must be the one at which Quantum Gravity starts. Pushed, the proponents say that Occam's razor requires that there be just the one scale. However, I do not know that Nature shaves with Occam's razor. Maybe she does not shave at all.

Instead of going to higher dimensions to resolve the problem of unifying GR and Quantum Theory, one might note that the metric tensor plays a double role: defining the gravitational potential; and measuring distances. Thus points in the spacetime will also be quantized, and hence "smeared out". However, to be able to use differential calculus, we need a continuum. One way of dealing with this would be to try to construct geometry on discrete spaces and use that as a model for spacetime. There have been various attempts for this, the best known of which, due to Sorkin [107], is "causal sets". In this approach space and time come into existence in discrete units. However, I have not seen how it would explain the results of the Aspect experiment any better than the usual spacetime does. The canonical quantization attempt has been taken farthest by Abhay Ashtekar [108] with his "loop quantum gravity". The smearing out has been taken into account but my problem with the different meanings of space and time persist here as well.

A totally different approach had been taken by Roger Penrose with his twistors [109, 35]. In some sense twistors were derived from his idea of spin networks [110], even though the first published paper on twistors appeared a few years before his paper on spin networks. (In fact, when I joined Penrose for the PhD in 1968, he gave me to read a *draft* of a paper on spin networks and a *reprint* of the paper on twistors.) The idea expressed in the work on spin networks was that spacetime may not be a fundamental structure in itself, but may be built up by fundamental entities that carry spin. Two of them, one with N units and the other with M units and $M < N$, and they exchange one unit of spin, they will continue with the same number of units as they came in with but will be deflected, with the one with less spin being deflected more than the one with greater spin. The net effect for a whole lot of such entities will be to build up a system of angles. For the whole network there would then be angles and distances built up. Thus the space would be quantized as the spin is quantized.

In many ways this idea lies at the base of loop quantum gravity as well. (As it happens, I had a similar idea and when I was applying for the PhD to different places, they would ask me if I had any ideas of what I wanted to do. When I started talking of this they would say "You are welcome to come here but we have no one to supervise you in this and you would have to work on your own. But have you talked to Roger Penrose ? He has similar ideas.)

A twistor is a pair of 2-component spinors that represents a null ray in a complexified Minkowski space that allows a construction of solutions of zero rest mass equations for arbitrary spin, including gravitation [111]. Being a null ray, it appears to be the only approach that has the seeds of being able to deal with nonlocality in space and time. The twistor lies in a 4 complex dimensional space, or 8 real dimensional space. As such the projective space of twistors has 6 real dimensions. It can be broken into two parts with the spin coefficient for twist being positive or negative, with a 5 dimensional boundary between them of twist zero, which corresponds to the space of null lines in real Minkowski space, called the space of null twistors. Using contour integrals for poles in the space of null twistors, we can obtain results for the twistor equivalent of Feynman graphs, which I call "Penrose graphs" [44, 34]. The results are finite and do not contain any divergences. As such, one has as much success with these as with any other attempts. However, as with all other attempts at Quantum Gravity, there is a hitch. Correctly, they only work in the high energy limit. There are *ad hoc* procedures of obtaining results for finite energies, but they do not give adequate predictions to be taken as providing a genuine theory of Quantum Relativity. Though it can be hoped that it provides for the basic nonlocality of Quantum Theory in space and time, the null line *does* have an arrow of time.

Another approach to the problem of putting Relativity and Quantum Theory together was to try taking the spacetime points themselves to be operators that do not commute. The idea being that the point does not exist in itself but only as it is observed. This approach is called "Noncommutative Geometry" and has been used in contexts other than the problem of Quantum Relativity as well [112]. When used with Superstrings or M-theory, the non-commutativity is taken only at the Planck scale. When used with loop quantum gravity it would appear that it remains at Planck scale. In this way of seeing the non-commutativity the problem of the difference between the Quantum and Relativity views of space and time would not be addressed by it.

It is my belief that more experiments to understand quantum behaviour, which provide a proper understanding of what brings about the onset of Quantum Mechanics, are needed. We need to know where the so-called "quantum level", starts. It is not a matter of short distance, as quantum effects spread at least over thousands of kilometers, and perhaps over light years. It is not a matter of short times, as quantum interference for single electron or photon experiments, can spread over days. Nor is it a matter of small energies, as solar mass objects show quantum behaviour. Only when we know where Quantum Theory starts can we hope to understand where Quantum Gravity starts. Moreover, we need to understand Quantum Theory in accelerated frames as well. Finally, only Nature, and not Mathematics, can tell us where and how to look for the onset of Quantum Gravity.

Bibliography

[1] A. Qadir, *Relativity: An Introduction to the Special Theory*, World Scientific 1989.

[2] A. Einstein and M. Grossmann, "Entwurf einer verallgemeinerten Relativitätstheorie und eine Theorie der Gravitation. I. Physikalischer Teil von A. Einstein II. Mathematischer Teil von M. Grossmann", translation "Outline of a Generalized Theory of Relativity and of a Theory of Gravitation. I. Physical Part by A. Einstein II. Mathematical Part by M. Grossmann", *Zeitschrift für Mathematik und Physik* **62** (1913) 225-244, 245-261.

[3] A. Einstein and M. Grossmann, "Kovarianzeigenschaften der Feldgleichungen der auf die verallgemeinerte Relativitätstheorie gegründeten Gravitationstheorie" tr. "Covariance Properties of the Field Equations of the Theory of Gravitation Based on the Generalized Theory of Relativity", *Zeitschrift für Mathematik und Physik* **63** (1914) 215-225.

[4] A. Einstein, "Grundgedanken der allgemeinen Relativitätstheorie und Anwendung dieser Theorie in der Astronomie", tr. "Fundamental Ideas of the General Theory of Relativity and the Application of this Theory in Astronomy", *Preussische Akademie der Wissenschaften, Sitzungsberichte*, **1915 (part 1)**, 315; "Zur allgemeinen Relativitätstheorie", tr. "On the General Theory of Relativity", *Preussische Akademie der Wissenschaften, Sitzungsberichte*, **1915 (part 2)**, 778-786, 799-801; "Erklärung der Perihelbewegung des Merkur aus der allgemeinen Relativitätstheorie", tr. "Explanation of the Perihelion Motion of Mercury from the General Theory of Relativity", *Preussische Akademie der Wissenschaften, Sitzungsberichte*, **1915 (part 2)**, 831-839; "Feldgleichungen der Gravitation", tr. "The Field Equations of Gravitation", *Preussische Akademie der Wissenschaften, Sitzungsberichte*, **1915 (part 2)**, 844-847.

[5] D. Hilbert, "Die Grundlagen der Physik", tr. "Foundations of Physics", *Mathematische Annalen* **92** (1915) 1-32.

[6] A. Einstein, "Grundlage der allgemeinen Relativitätstheorie", tr. "The Foundation of the General Theory of Relativity", *Annalen der Physik (ser. 4)* **49** (1916) 769–822.

[7] A. Einstein, "Näherungsweise Integration der Feldgleichungen der Gravitation Approximative Integration of the Field Equations of Gravitation", tr. "Approximative Integration of the Field Equations of Gravitation",

Preussische Akademie der Wissenschaften, Sitzungsberichte, **1916 (part 1)**, 688–696; **1918 (part 1)**, 154–167.

[8] A. Einstein, "Kosmologische Betrachtungen zur allgemeinen Relativitätstheorie", tr. "Cosmological Considerations in the General Theory of Relativity", *Preussische Akademie der Wissenschaften, Sitzungsberichte*, **1917 (part 1)**, 142–152.

[9] A. Einstein, "Zur Elektrodynamik bewegter Körper", tr. "On the Electrodynamics of Moving Bodies", *Annalen der Physik*, **18** (1905) 639-641.

[10] A. Einstein, "Die vom Relativätsprinzip geforderte Trägheit der Energie", tr. "On the Inertia of Energy Required by the Relativity Principle"; *Annalen der Physik*, **23** (1907) 371-384.

[11] P.A. Schlipp, *Albert Einstein: Philosopher-Scientist, Volume II*, Harper and Brothers Publishers 1951.

[12] A. Pias, *Subtle is the Lord...: The Science and Life of Albert Einstein*, Oxford University Press (2005).

[13] M. Karim and A. Qadir (eds.), *Experimental Gravitation: Proceedings of the International Symposium on Experimental Gravitation*, Plenum Press 1994; *Classical and Quantum Gravity* **11** (1994) A1-A242.

[14] L.D. Landau and E.M. Lifshitz, *The Classical Theory of Fields*, Edition 4, Butterworth-Heinemann 1987.

[15] C.W. Misner, K.S. Thorne and J.A. Wheeler, it Gravitation, W.H. Freeman 1973.

[16] M. Giaquinta and S. Hildebrandt, "Calculus of Variations 1: The Lagrangian Formalism", Springer-Verlag 1996.

[17] I. Ciufolini and J.A. Wheeler, *Gravitation and Inertia*, Princeton University Press 1996.

[18] J. Barbour, *The Discovery of Dynamics*, Oxford University Press 2001.

[19] A.Z. Petrov, *Einstein Spaces*, Pergamon Prress 1969.

[20] N. Straumann, *General Relativity and Relativistic Astrophysics*, Springer-Verlag Texts and Monographs in Physics, 1988.

[21] H. Stephani, D. Kramers, M.A.H. MacCallum, C. Hoenselaers, E. Herlt, *Exact Solutions of Einstein's Field equations* Second Edition, Cambridge University Press 2009.

[22] R.V. Pound and G.A. Rebka Jr., "Gravitational Red-Shift in Nuclear Resonance", *Physical Review Letters* **3** (1959) 439 – 441.

[23] H. Reissner, "Über die Eigengravitation des elektrischen Feldes nach der Einsteinschen Theorie", *Annalen der Physik* **50** (1916) 106 - 120.

[24] G. Nordström, "On the Energy of the Gravitational Field in Einstein's Theory", *Verhandl. Koninkl. Ned. Akad. Wetenschap., Afdel. Natuurk., Amsterdam* **26** (1918) 1201 - 1208.

[25] R.P. Kerr, "Gravitational Field of a Spinning Mass as an Example of Algebraically Special Metrics", *Physical Review Letters* **11** (1963) 237 – 238.

[26] E.T. Newman et al., "Metric of a Rotating, Charged Mass", *Journal of Mathematical Physics* **6** (1965) 918 – 919.

[27] A.N. Aliev, Y. Nutku, "Impulsive Spherical Gravitational Waves", *Classical & Quantum Gravity* **18** (2001) 891 - 906.

[28] A. Einstein and N. Rosen, "On gravitational waves", *Journal of the Franklin Institute* **223** (1937) 43 – 54.

[29] H. Bondi and I. Robinson, "Gravitational Waves in General Relativity. III. Exact Plane Waves", *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* **251** (1957) 519 - 533.

[30] R.T. Jantzen, P. Carini, and D. Bini, "The Many Faces of Gravitoelectromagnetism", *Annals of Physics* **215** (1992) 1 - 50.

[31] R. Arnowitt, S. Deser and C.W. Misner, "Dynamical Structure and Definition of Energy in General Relativity", *Phys. Rev.* **116** (1959) 1322 – 1330; "Canonical Variables for General Relativity", *Phys. Rev.* **117**(1960) 1595 – 1602.

[32] B.S. DeWitt, "Quantum Theory of Gravity: 3. Applications of the Covariant Theory", *Phys. Rev.* **162**(1967) 1239 – 1256; "Quantum Field Theory in Curved Space-Time", *Physics Reports* **19** (1975) 295 – 357.

[33] R. Penrose and M.A.H. MacCallum, "Twistor Theory: An Approach to the Quantisation of Fields and Space-Time", *Physics Reports* **6** (1973) 241 - 315.

[34] A. Qadir, "Penrose Graphs", *Physics Reports* **39** (1978) 133 - 170.

[35] R. Penrose and W. Rindler, *Spinors and Space Time Volume 1: Two-Spinor Calculus and Relativistic Fields; Volume 2: Spinor and Twistor Methods in Space-Time Geometry*, Cambridge University Press 1984; 1986.

[36] J.A. Wheeler, "Superspace and the Nature of Quantum Geometrodynamics", *Battelle Rencontres* (eds. C.M. DeWitt and J.A. Wheeler), pp. 242–307, Benjamin 1968.

[37] S.M. Mahajan A. Qadir and P.M. Valanju, "Reintroducing the Concept of Force into Relativity Ttheory", *Nuovo Cimento B* **65** (1981) 404 - 417.
A. Qadir and J. Quamar, "Relativistic Generalization of Newtonian Forces" *Proceedings of the Third Marcel Grossmann Meeting* pp. 189 - 220, ed. Hu Ning, North Holland Publishing Company 1983.
A. Qadir, "Reissner-Nordström Repulsion", *Physics Letters A* **99** (1983) 419 - 420.
A. Qadir, "General Relativity in Terms of Forces", *Proceedings of the Third*

Regional Conference on Mathematical Physics pp. 481-490, eds. F. Hussain and A. Qadir, World Scientific 1990.
A. Qadir, "The Gravitational Force in General Relativity", *M.A.B. Beg Memorial Volume*, eds. A. Ali and P.A. Hoodbhoy, World Scientific 1991.

[38] A. Qadir and M. Sharif, "The Relativistic Generalization of the Ggravitational Force for Arbitrary Spacetimes", *Nuovo Cimento B* **107** (1992) 1071 - 1083.
A. Qadir and M. Sharif, "General Formula for the Momentum Imparted to Test Particles in Arbitrary Spacetimes", *Physics Letters A* **167** (1992) 331 - 334.

[39] B. Carter, "The complete analytic extension of the Reissner-Nordström metric in the special case $e^2 = m^2$", *Physics Letters* **21** (1966) 423 - 424.

[40] A. Qadir and A.A. Siddiqui, "The cosmic censorship hypothesis and the naked Reissner-Nordström singularity", *Proceedings of the Eleventh Regional Conference on Mathematical Physics*, eds. S. Rahvar, N. Sadooghi and F. Shojai, World Scientific 2005.

[41] A. Qadir and M. Rafique, "A relativistic explanation for pulsar drift", *Proc. Fourth Marcel Grossmann Meeting*, Ed. R. Ruffini (Elsevier Science Publishers 1986), pp. 1597-1606; "Pulsar supernova remnant correlation", *Chinese Physics Letters* **3** (1986) pp. 189 - 191;
A. Qadir, M. Rafique and A. W. Siddiqui, "Pulsar drift and rotational collapse", *Chinese Physics Letters* **4** (1987) 177 - 180;
Z.U. Mian, A. Qadir, J. Quamar and H.A. Rizvi, "On the inclination of planetary orbits", *Commun. Theoret. Phys.* **11** (1989) 115 - 120;
A. Qadir and H.A. Rizvi, "A relativistic explanation for the inclination of planetary orbits", *Proc. Fifth Marcel Grossmann Meeting*, Eds. D.G. Blair and M.J. Buckingham (World Scientific 1989) pp. 1385-1405.

[42] A. Qadir and A.A. Siddiqui, "Use of Carter's coordinates for the extreme Reissner-Nordstrom spacetime to construct free-fall geodesics and a flat foliation", *Nuovo Cimento B* **117** (2002) 909 - 916; "Foliation of the Schwarzschild and Reissner-Nordstrom spacetimes by flat spacelike hypersurfaces", *Int. J. Mod. Phys. D* **15** (2006) 1419 - 1440;
V. Hussain, A. Qadir and A. A. Siddiqui, "A note on flat foliations of spherically symmetric spacetimes", *Phys. Rev. D* **65** (2002) 027501, 1-2.

[43] A. Pervez, A. Qadir and A. A. Siddiqui, "Foliation by constant mean-curvature hypersurfaces of the Schwarzschild spacetime", *Phys. Rev. D* **51** (1995) 4598 - 4599;
A. Qadir and A.A. Siddiqui, "K-slicing the Schwarzschild and Reissner Nordstrom spacetimes", *J. Math. Phys.* **40** (1999) 5883 - 5889;
A. Qadir, M. Sajid and A. A. Siddiqui, "K-slicing of the Reissner-Nordstrom spacetime: some new observations", *Nuovo Cimento B* **122** (2007) 333 – 341. '

[44] P. Morse, *Thermal Physics* (Second Edition) W.A. Benjamin 1971.

[45] A. Qadir, "On the reality of Hawking radiation", presented at the *Sixth Italian-Pakistani Workshop on Relativistic Astrophysics* January 24 - 26,

2019, Islamabad, Pakistan, Proceedings to be published by it Int. J. Mod. Phys. D Special Issue.

[46] R. Penrose and R.M. Floyd, "Extraction of Rotational Energy from a black hole", *Nature Physical Science* **229** (1971) 177 - 179;
R. Penrose, "Aspects of General Relativity", in *Physics and Contemporary Needs, Vol. 1*, ed. Riazuddin, Plenum Press 1977, pp 383 - 420.

[47] D. Christodoulou and R. Ruffini, "Reversible transformations of a charged black hole", *Phys. Rev. D* **4** (1971) 3552 – 3555.

[48] J.D. Bekenstein, "Black holes and entropy" *Phys. Rev. D* **7** (1973) 2333 - 2346.

[49] S.W. Hawking and G.F.R. Ellis, *The Large Scale Structure of Space-Time*, Cambridge University Press 1973.

[50] S.A. Fulling, "Nonuniqueness of Canonical Field Quantization in Riemannian Space-Time", *Phys. Rev. D* **7** (1973) 2850 – 2862.

[51] S.W. Hawking, "Black hole explosions?", *Nature* **248** (1974) 30–31; "Particle creation by black holes ", *Communications in Mathematical Physics* **43** (1975) 199-220.

[52] S.W. Hawking, "Breakdown of probability in gravitational collapse", *Phys. Rev. D* **14** (1976) 2460 - 2473.

[53] L. Susskind, *The Black Hole War: My Battle with Stephen Hawking to Make the World Safe for Quantum Mechanics*, Little and Brown 2008.

[54] V. Rubin, "A Century of Galaxy Spectroscopy", *Ap. J.* **451** (1995) 419 - 428.

[55] V. Hussain and B. Qureshi, "Ground State of the Universe and the cosmological constant. A Nonperturbative Analysis", *Phys. Rev. Lett.* **116** (2016) 061302.

[56] M. Milgrom, "A modification of the Newtonian dynamics as a possible alternative to the hidden mass hypothesis", *Ap. J.* **270** (1983) 365 – 370.

[57] H.A. Buchdahl, "Non-linear Lagrangians and cosmological theory", *MNRAS* **150** (1970) 1 – 8.

[58] E.M. Lifshitz and I.M. Khalatnikov, "On the singularities of cosmological solutions of the gravitational equations, I; II", *JETP* **39** (1961) 149 - 157; 800 - 808.
V.A. Belinskii, I.M. Khalatnikov and E.M. Lifshttz, "Oscillatory Approach to a Singular Point in the Relativistic Cosmology", *Advances in Physics* **19** (1970) 525–573.

[59] R. Penrose, "Gravitational collapse and space-time singularities", *Phys. Rev. Lett.* **14** (1965) 57 - 59;
S.W. Hawking, "Occurrence of Singularities in Open Universe Models", *Phys. Rev. Lett.* **15** (1965) 689 - 690; "Singularities in the Universe", *Phys. Rev. Lett.* **17** (1966) 444 - 445;

S.W. Hawking and R. Penrose, "The singularities of Gravitational Collapse and Cosmology", *Proc. Roy. Soc. Lond. A* **314** (1970) 529 - 548.

[60] L.C. Shepley and M.P. Ryan, *Homogeneous Relativistic Cosmologies*, Princeton University Press 1975.

[61] C.W. Misner, "Mixmaster Universe", *Phys. Rev. Lett.* **22** (1969) 1071 - 1074.

[62] R. Penrose, *Road to Reality: A Complete Guide to the Laws of the Universe*, Alfred A. Knopf 2004.

[63] R. Brout, F. Englert and E. Gunzig, "The Creation of the Universe as a Quantum Phenomenon", *Ann. Phys.* **115** (1978) 78 - 106.

[64] A.H. Guth, "Inflationary Universe: A Possible Solution to the Horizon and Flatness Problems", *Phys. Rev. D* **23** (1981) 347 – 356.
The Inflationary Universe: The Quest for a New Theory of Cosmic Origins, Perseus Books 1997.

[65] A.D. Linde, "Scalar Field Fluctuations in Expanding Universe and the New Inflationary Universe Scenario", *Phys.Lett.* **116B** (1982) 335 - 339.

[66] A.D. Linde, "Chaotic Inflation" it Phys.Lett. **129B** (1983) 177 - 181.

[67] É.J. Cartan, *The Theory of Spinor*, Hermann 1966.

[68] E.T. Newman and R. Penrose, "An approach to gravitational radiation by a method of spin coefficients", *J. Math. Physics.* **3** (1962) 896 - 902; Errata **4** (1963) 998.

[69] G.S. Hall, *Symmetries and Curvature Structure in General Relativity*, World Scientific 2004.

[70] A. Qadir, "Killing vectors of spherically symmetric, static spacetimes", paper presented at *General Relativity 11*, Stockholm, Sweden 1986;
A.H. Bokhari and A. Qadir, "Symmetries of static, spherically symmetric space-times", *J. Math. Phys.* **28** (1987) 1019 - 1022;"Killing vectors of static, spherically symmetric spacetimes", *J. Math. Phys.* **31** (1990) 1463 - 1470.

[71] A.H. Bokhari and A. Qadir, "Erratum: Symmetries of static, spherically symmetric space-times [J. Math. Phys. 28, 1019 (1987)", *J. Math. Phys.* **29** (1988) 525.

[72] A. Qadir and M. Ziad, "The classification of spherically symmetric space-times", *Nuovo Cimento B* **110** (1995) 317 - 334; "Spherically symmetric spacetimes", PhD thesis Quaid-i-Azam University, Islamabad, 1990.

[73] R. Penrose, "Structure of Space-time", *Battelle Rencontres*, eds. C.M. De Witt and J.A. Wheeler, Benjamin 1968.

[74] K.A. Khan and R. Penrose, "Scattering of Two Impulsive Gravitational Plane Waves", *Nature*, **229** (1971) 185 - 186.

[75] J. Weber, "Gravitational radiation", *Phys. Rev. Lett.*, **18** (1967) 498 – 501; "Gravitational-wave-detector events", **20** (1968) 1307 – 1308; "Evidence for discovery of gravitational radiation", **22** (1969) 1320 – 1324.

[76] B. Bertotti and B.J. Carr, "The prospects of detecting gravitational background radiation by Doppler tracking interplanetary spacecraft", *Ap. J.* **236** (1980) 1000 - 1011.

[77] A. Qadir and A.A. Mufti, "Gravitational waves from the Big Bang", *Lett. al. Nuovo Cimento* **29** (1980) 528 - 532.

[78] B. Bertotti, B.J. Carr, M.J. Rees, "Limits from the timing of pulsars on the cosmic gravitational wave background", . *Mon. Not. Roy. Astron. Soc.* **203** (1983) 945 - 954.

[79] M. Karim and A. Qadir (Eds.), *Experimental Gravitation*, IOP Press 1994; Special Issue of *Classical and Quantum Gravity* bf 11 No. 6A 1994.

[80] S. Chandrasekhar, *An Introduction to the Study of Stellar Structure*, Dover Publications 2010.

[81] S. Chandrasekhar, "The Maximum Mass of Ideal White Dwarfs", *Ap. J.* (1931) 81 - 82.

[82] L.D. Landau, "On the Theory of Stars", *Phys. Z. Sowjet.* **1** (1932) 285 - 286.

[83] A. Hewish, J.S. Bell, J.D.H. Pilkington, P.F. Scott and R.A. Collins, "Observation of a Rapidly Pulsating Radio Source", *Nature* **217** (1968) 709 - 713;
J.D.H. Pilkington, A. Hewish, J.S. Bell and T.W. Cole, "Observations of Some Further Pulsed Radio Source", *Nature* **218** (1968) 126 - 129.

[84] T. Gold, "Rotating neutron stars and the nature of pulsars", *Nature*, **221** (1969) 25 – 27.

[85] C.E. Rhoades and R. Ruffini, "Maximum Mass of a Neutron Star", *Phys. Rev. Lett.* **32** (1974) 324 - 327.

[86] R. Leach and R. Ruffini, "On the Masses of X-Ray Sources", *Ap. J. Lett.* **180** (1973) L15 - L16.

[87] R. Ruffini, "Astrophysics, General Relativity and Cosmology", *Physics and Contemporary Needs Vol. 1*, Ed. Riazuddin, Plenum Press 1977.

[88] D.E. Holz and J.A. Wheeler, "Retro-Machos: π in the Sky?", *Ap. J.* **578** (2002) 330 - 334.

[89] R.A. Hulse and J.H. Taylor, "Discovery of a Pulsar in a Binary System", *Ap. J.* **195** (1974) L51 - L55.

[90] B.P. Abbott et al., "Observation of Gravitational Waves from a Binary Black Hole Merger", *Phys. Rev. Lett.* **116** (2016) 061102 (16 pages).

[91] B.P. Abbott et al., "GW151226: Observation of Gravitational Waves from a 22-Solar-Mass Binary black hole Coalescence", *Phys. Rev. Lett.* **116** (2016) 241103 (14 pages).

[92] B.P. Abbott et al., "GW170817: Observation of Gravitational Waves from a Binary neutron star Inspiral", *Phys. Rev. Lett.* **119** (2017) 161101 (18 pages).

[93] A. Einstein, B. Podolsky and N. Rosen, "Can Quantum-Mechanical Description of Physical Reality be Considered Complete?", *Phys. Rev.* **57** (1935) 777 - 280.

[94] D. Bohm and Y. Aharonov, "Discussion of Experimental Proof for the Paradox of Einstein, Rosen, and Podolsky", *Phys. Rev.* **108** (1957) 1070 - 1076.

[95] J.S. Bell, "On the Einstein Podolsky Rosen Paradox", *Physics* **1** (1964) 195 - 200.

[96] A. Aspect, P. Grangier and G. Roger, "Experimental Tests of Realistic Local Theories via Bell's Theorem", *Phys. Rev. Lett.* **47** (1981) 461 - 463.

[97] J.A. Wheeler, "The 'Past' and the 'Delayed-Choice Double-Slit Experiment' ", *Mathematical Foundations of Quantum Theory*, pp 9 – 48, Ed. A.R. Marlow, Academic Press 1978.

[98] V. Jacques et al., "Experimental Realization of Wheeler's delayed-choice Gedanken Experiment", *Science* **315** (2007) 966 - 968.

[99] Th. Kaluza, "Zum Unitätsproblem der Physik" ("On the Problem of the Unification of Physics"), *Proceedings of the Prussian Academy of Sciences* **1:** (1921) 966 - 972.

[100] A. Kheyfets, L. K. Norris and A. Qadir, "Internal Consistency of Kaluza-Klein Ttheories", *Nuovo Cimento A* **101** (1989) 367 - 383.

[101] O. Klein, "Quantentheorie und fünfdimensionale Relativitätstheorie", ("Quantum Theory and Five-Dimensional Relativity"), *Zeitschrift für Physik A* **37** (1926) 895 – 906.

[102] G. 't Hooft and M.J.G. Veltman, "Regularization and Renormalization of Gauge Fields", *Nuclear Physics B* **44** (1972) 189 – 219.

[103] J.-L.Gervais and B.Sakita, "Field Theory Interpretation of Supergauges in Dual Models", *Nuclear Physics B* **34** (1971) 632 - 639.

[104] E.J. Cremmer, B. Julia and J. Scherk, "Supergravity theory in 11 dimensions", *Phys. Lett. B* **76** (1978) 409 - 412.

[105] M.B. Green and J.H. Schwarz, "Anomaly Cancellations in Supersymmetric D = 10 Gauge theory and Superstring Theory", *Phys. Lett. B* **149** (1984) 117 - 122; "Infinity Cancellations in SO(32) Superstring Theory", **151** (1985) 21 - 25.

[106] M.B. Green, J.H. Schwarz and E. Witten, *Superstring Theory: Vol.1 Introduction; Vol. 2 Loop Amplitudes, Anomalies and Phenomenology*, Cambridge University Press 1988.

[107] R.D. Sorkin, "A Finitary Substitute for Continuous Topology", *Int. J. Theoret. Phys.* **30** (1991) 923 - 947.

[108] A. Ashtekar and C. Rovelli, "A Loop Representation for the Quantum Maxwell Field", *Class. and Quant. Grav.* **9** (1992) 1121 - 1150.
A. Ashtekar, "Physics in Loop Space", *Proc. Cochin Advanced School on Gravitation and Cosmology*, eds. B.R. Iyer, S.V. Dhurandhar and K.B. Joseph, (IUCAA, Pune, 1993), Lecture Notes by R.S. Tate.

[109] R. Penrose, "Twistor Algebra", *J. Math. Phys.* **8** (1967) 345 – 366.

[110] R. Penrose, "Angular Momentum: An Approach to Combinatorial Space-time", *Quantum Theory and Beyond*, Ed. T. Bastin, Cambridge University Press 1971; "Applications of Negative Dimensional tensors", *Combinatorial Mathematics and its Applications: Proc. Conf. Oxford, 1969* Ed. D.J.A. Welsh , Academic Press 1971.

[111] R. Penrose, "Solutions of the Zero-Rest-Mass Equations", *J. Math. Phys.* **10** (1969) 38 - 39.

[112] A. Connes, "Noncommutative Differential Geometry", *Inst. Hautes Etudes Sci. Publ. Math.* **62** (1985) 257 - 360; *Essay on Physics and Noncommutative Geometry: The Interface of Mathematics and Particle Physics*, pp. 9 - 48, Oxford University Press 1990.

Index