STRUCTURES IN GENERAL RELATIVITY

by

Steven Tieu
B.Sc. Simon Fraser University 1995
M.Sc. Simon Fraser University 2001

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY
in the Department of Physics and Astronomy

© Steven Tieu, 2006
University of Victoria

All rights reserved. This thesis may not be reproduced in whole or in part, by photocopy or other means, without the permission of the author.
Supervisory Committee

Structures in General Relativity

by

Steven Tieu
B.Sc. Simon Fraser University 1995
M.Sc. Simon Fraser University 2001

Supervisory Committee

____________________________
Dr. F. I. Cooperstock, Supervisor (Department of Physics and Astronomy)

____________________________
Dr. M. Lefebvre, Departmental Member (Department of Physics and Astronomy)

____________________________
Dr. M. Pospelov, Departmental Member (Department of Physics and Astronomy)

____________________________
Dr. G. G. Miller, Outside Member (Department of Mathematics and Statistics)

____________________________
Dr. P. S. Wesson, External Examiner (Department of Physics and Astronomy, University of Waterloo)
Abstract

Structures within general relativity are examined. The differences between man-made structures and those predicted by the Einstein differential equations are very subtle. Exotic structures such as the Gödel Universe and the Gott cosmic string are examined with emphasis on closed time-like curves. Newtonian models are seen to also have an exotic aspect in that a vast halo consisting of unknown matter dominates
the galaxy. We introduce a model for galaxies based on a general relativity framework with the goal of excluding such artifacts from the system while describing the flat-rotation curves. Structures within this model were speculated to be exotic but after close scrutiny, their nature is shown to be benign. Numerical approaches are applied to model four galaxies: The Milky Way, NGC 3031, NGC 3198 and NGC 7331. Density and mass are deduced from these models and compared to the Newtonian models. Within the visible/HI region, there is 30% reduction in total mass. Extending the model to 10 times beyond the HI region using various fall-off scenerios, it is shown that there is only modest increase of the accumulated mass. In comparison to the Newtonian approach to galactic dynamics, the massive halos are not required to account for the flat velocity profiles.
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervisory Committee</td>
<td>ii</td>
</tr>
<tr>
<td>Abstract</td>
<td>iii</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>v</td>
</tr>
<tr>
<td>List of Figures</td>
<td>viii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>x</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>xi</td>
</tr>
<tr>
<td>Dedication</td>
<td>xii</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2 Relativity</td>
<td>4</td>
</tr>
<tr>
<td>2.1 Mathematical Foundation</td>
<td>7</td>
</tr>
<tr>
<td>2.1.1 Differential Geometry</td>
<td>9</td>
</tr>
<tr>
<td>2.1.2 The Action Principle</td>
<td>16</td>
</tr>
<tr>
<td>2.1.3 Noether’s Theorem</td>
<td>18</td>
</tr>
<tr>
<td>2.1.4 Variational Formulation</td>
<td>20</td>
</tr>
<tr>
<td>2.2 Axial Symmetry</td>
<td>23</td>
</tr>
</tbody>
</table>
3 Non-Causal Structures

3.1 Introduction ................................................. 25

3.2 Gödel’s Closed Time-Like Curves ........................... 31

3.3 Gott’s Closed Time-Like Curves ............................. 37

3.4 Discussions .................................................. 41

4 Galactic Structures .................................................. 45

4.1 Newtonian Astrophysics ...................................... 46

4.1.1 Dark Matter ................................................. 46

4.2 Departure From Newtonian Gravitation ....................... 51

4.3 General Relativistic Astrophysics ......................... 55

4.4 A Trial Model ................................................ 56

4.5 The Model Galaxy ............................................. 58

4.5.1 The Approximate Einstein Equations .................... 62

4.5.2 Dust Particle Velocities ................................... 63

4.6 Modeling Observed Galactic Rotation Curves ............. 64

4.7 Boundary Layer Conditions ................................... 67

5 Numerics .......................................................... 70

5.1 Least Squares .................................................. 70

5.1.1 Linear Square Fit ......................................... 71

5.1.2 Non-Linear Least-Square Fit ............................. 73
5.1.3 Newton's Method ................................................. 75
5.1.4 Minimization ..................................................... 76
5.1.5 Multi-dimensional Minimization ............................. 78
5.2 Numerical Solutions to PDE ................................. 80
5.3 Numerical Application and Results ....................... 83
  5.3.1 Curve Fitting ................................................... 83
5.4 Numerical Solution to The Boundary Layer ............. 88
  5.4.1 Numerical PDE ................................................ 89
  5.4.2 Matching Series .............................................. 98
  5.4.3 Non-Existence of a Boundary Layer ................... 100

6 Conclusions .......................................................... 102

Bibliography ............................................................. 104

Appendix ........................................................................ 109
# List of Figures

<p>| Figure 2.1 | Twin paradox. | 5 |
| Figure 2.2 | Tangent space. | 11 |
| Figure 2.3 | Curvature. | 15 |
| Figure 3.1 | Accelerated Reference frame. | 27 |
| Figure 3.2 | Closed timelike curve. | 34 |
| Figure 3.3 | Curve in Minkowski spacetime. | 35 |
| Figure 3.4 | Identification of events. | 38 |
| Figure 3.5 | Worldline across identified events. | 39 |
| Figure 3.6 | Boosted spacetime events. | 40 |
| Figure 3.7 | Lightcones in pq-coordinates. | 42 |
| Figure 4.1 | Singularity of a massive point particle. | 66 |
| Figure 5.1 | Newton’s method. | 74 |
| Figure 5.2 | Newton’s method failure. | 74 |
| Figure 5.3 | Gamma function. | 77 |
| Figure 5.4 | Minimization. | 79 |</p>
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.5</td>
<td>Milky Way velocity curve fit.</td>
<td>89</td>
</tr>
<tr>
<td>5.6</td>
<td>Milky Way density.</td>
<td>90</td>
</tr>
<tr>
<td>5.7</td>
<td>Milky Way log density contour.</td>
<td>91</td>
</tr>
<tr>
<td>5.8</td>
<td>NGC 3031 velocity curve fit and density.</td>
<td>92</td>
</tr>
<tr>
<td>5.9</td>
<td>NGC 3198 velocity curve fit and density.</td>
<td>93</td>
</tr>
<tr>
<td>5.10</td>
<td>NGC 7331 velocity curve fit and density.</td>
<td>94</td>
</tr>
<tr>
<td>5.11</td>
<td>Log density of the Milky Way and NGC 3031.</td>
<td>95</td>
</tr>
<tr>
<td>5.12</td>
<td>Log density of NGC 3198 and NGC 7331.</td>
<td>95</td>
</tr>
<tr>
<td>5.13</td>
<td>Milky Way velocity fall-off scenerios.</td>
<td>96</td>
</tr>
<tr>
<td>5.14</td>
<td>Milky Way accumulated mass.</td>
<td>96</td>
</tr>
<tr>
<td>5.15</td>
<td>PDE solution.</td>
<td>97</td>
</tr>
<tr>
<td>5.16</td>
<td>Matching $N$ function and its derivative.</td>
<td>99</td>
</tr>
<tr>
<td>5.17</td>
<td>Sign of $N_x N_y$.</td>
<td>100</td>
</tr>
</tbody>
</table>
List of Tables

Table 4.1 Order of magnitude of variables. .................................. 61
Table 5.1 Milky Way coefficients. ............................................. 85
Table 5.2 NGC 3031 coefficients. ............................................. 86
Table 5.3 NGC 3198 coefficients. ............................................. 87
Table 5.4 NGC 7331 coefficients. ............................................. 88
Table 5.5 Galactic mass comparisons. ....................................... 88
Acknowledgements

I would like to thank Dr. Fred I. Cooperstock for his support, his tremendous patience and his great wisdom. I would also like to thank Dr. Michel lefebvre and Dr. Werner Israel for being such great teachers and providing much inspirations. The support from my family has been invaluable and greatly appreciated. I would like to thank my fellow students, Ian Bailey, Tayfun Ince, Greg King, Aaron Saunderson, Daniel Vanderster and Jinkoo Yun for all the discussions in physics and other topics. And least but not last, I would like to thank Nikola Brabenec for being so understanding of my good intentions and for seeing the humour in the situation for all the times I said the wrong thing.
Dedication

To Helen, my little sister.
Chapter 1

Introduction

In 1949, Gödel solved the Einstein equations which describes a universe filled with a rotating fluid. He was surprised to discover that the solution allowed fluid particles to connect to their initial four-dimensional coordinates; that is, the fluid particles can travel into their past. While this was the first instance of serious consideration in associating time-travel with general relativity, it was not the first general relativistic solution which includes such a feature. It is noted that van Stockum’s solution[4] in 1937 of an infinitely long spinning cylinder of dust particles apparently allows for a pathway to travel into the past. In fact, more often than not, axially-symmetric solutions to the Einstein equations such as in the Kerr metric apparently will exhibit this feature.

Conditions for time-travel are usually associated with one of the following: negative energy, stitching of spacetime and discontinuity in spacetime coordinates (resulting in faster-than-light travel). In the first category, arguments against negative energy being unphysical can rule out many models which use such an approach. We believe that Gödel’s solution falls into the second category because there is no evidence of faster-than-light travel nor is there any evidence of negative energy. However, its
association with stitching of spacetime is not completely obvious and we will demonstrate the connections. An example which falls into the last category would be the Gott cosmic string where angle deficits from strings allow one to traverse spacelike separated events from one to another instantaneously. This faster-than-light travel was exploited to create spacetime structures in which time travellers are allowed to exist.

These exotic spacetime structures go very much against our own intuitions. One can easily imagine a host of paradoxes which arise from such conditions. Clearly there must be some inconsistencies which would disprove the notion of time-travel. However, trying to disprove time-travel is an unusually difficult task. This is due to the ambiguous nature of the laws of physics associated with each step in creating or negating the existence of such exotic spacetime structures. That is, the laws of physics do not favour one approach over another.

In this thesis, we shall first examine systems which have such exotic structures. In particular, the Gödel universe and the Gott cosmic string will be closely scrutinized. We find an underlying theme for the introduction of these unusual exotic features.

While general relativity may be complicated, allowing room for such exotic structures to exist, classical Newtonian systems are also afflicted with their own set of exotic features. Newtonian models of galaxies lead to vast massive halos which, from all measured results, seem to indicate that they consist of exotic matter as their constituent. To remedy this, we propose a model galaxy in a general relativistic framework where no exotic features exist. It usurps the classical model by resolving the flat galactic rotation curves without any introduction of exotic dark matter.

To begin in Chapter 2, general relativity is introduced and a mathematical foundation in differential geometry is established. In Chapter 3, non-causal structures in relativity are examined where the features in the Gödel spacetime and the Gott string are hypothesized to be man-made structures rather than requirements derived from general relativity. In Chapter 4, historical galactic structure in the framework of Newtonian mechanics which leads to the requirement for a dark matter halo is
introduced. Motivation for a departure from the Newtonian approach to a relativistic paradigm is given. We argue that the non-linearity in general relativity creates the observed flat rotation curves. Density and total mass are deduced from the model which does not require a vast halo of exotic dark matter. In Chapter 5, numerical schemes are applied to the partial differential equations generated from the model of the galaxy. Numerical results from the density and total mass are computed. Finally, Chapter 6 ends with discussions and conclusions.
Chapter 2

Relativity

Relativity is far from being an intuitive subject. One’s first encounter with it typically includes unusual phenomena such as length contraction and time dilation. When relativity is mixed with intuitive thinking, many paradoxes seem to arise such as the twin paradox or the ladder paradox. While these are easily explained, others such as the Ehrenfest paradox are not so clear-cut. In 1909 Ehrenfest[1], proposed a thought experiment about a relativistic spinning disk. The ratio of the circumference of a disk to its diameter is not \( \pi \) according to length contraction. Until recently, new resolutions[53] continue to be proposed and new interpretations are made to resolve this paradox based on different assumptions. After almost a century since relativity was introduced, new paradoxes[32][33] are still being introduced such as [52]Matsas’s submarine paradox in 2003. These paradoxes serve as exemplary cautions in dealing with a very counter-intuitive theory.

Given the non-intuitive nature of relativity, one must maintain a good grasp of control through the mathematics. When described in proper mathematical language, none of these paradoxes occurs and each can be easily explained. For example, with the twin paradox if one specifies the exact coordinates and the time slices of simul-
Figure 2.1: In the twin paradox, one twin will age 21.3 years while the other ages 12 years. (a) The time-slices of simultaneity of the departed twin during uniform motion are shown in the left figure. (b) The right figure includes the time-slices for the accelerated parts of trip.
taneity as shown in figure 2.1(a), one can see exactly where the paradox of difference in clock rates arises. The resolution to the paradox is in the twin’s non-symmetric reference frames during the intermediate accelerations as shown in Figure 2.1(b). One can easily see how one twin will age faster than another. In more subtle systems such as in the Ehrenfest paradox, it requires more than just a space-time diagram to explain. It requires mathematical statements about the elasticity of the disk to have a well-posed problem. While all of these problems arise from our lack of intuitive grasp of a four-dimensional spacetime (along with its transformations), once stated in a mathematical language, the systems become easier to handle. The mathematics underlying relativity is a powerful tool.

One must remember that the mathematics used in relativity is just a model as any other model in science. It can tell us about certain aspects of the system which leads to some very interesting research. But some features must not be taken literally. If one carries the model too far, interpretation can be misleading. One such example concerns Schwarzschild’s event-horizon. It was the focus of much attention in the decades following the discovery of the Schwarzschild solution. It was argued that something physically special occurred at the event horizon and may possibly be a break-down of the laws of physics. Others argue that it is merely a mathematical artifact. That is, while the Riemann curvature tensor components become singular in one coordinate frame, the Kretchman scalar remains well-behaved, showing that it is only a coordinate singularity. In any case, the general consensus is that the Schwarzschild solution outside the event horizon is a very good model but its interior domain may have limitations just as any other model in science.

Just as with all science, once a model is provided, boundaries are determined by exploring the extremes of the model. In experimental physics, the limitations may be the amount of energy supplied, the length-scale of the devices, the accuracy of measurements, etc. For example, in experimental relativity based on the accuracy of current technology, the extreme boundary of length-scale is in the micrometre regime[50]. But in theoretical physics where “devices” are ideal and measurements
are perfect, the length-scale can extend down to the Planck regime. Just as Newton did not know the limitations of his theory, we do not know the limitations on Einstein’s theory except for the extreme boundary encroaching upon quantum mechanics. Once a relativistic model is provided, examining the boundaries and determining what is a man-made mathematical artifact and what is predicted by general relativity is not a trivial task.

An example of such a situation is the feature associated with Gödel’s Universe. What made Gödel’s 1949 solution to the Einstein equations unusual was that the solution contained closed time-like curves (CTC). This led to controversies in the existence of the CTCs. Some argued that a universe being completely filled with rotating fluid is unnatural which leads to it containing unnatural CTCs. Others argued that it is very natural and is predicted by general relativity because CTCs show up in other systems as well. The CTC feature of Gödel’s Universe cannot easily be determined whether or not it is man-made.

2.1 Mathematical Foundation

In Euclidean geometry, the distance \( l \) which separates two points \( A \) and \( B \) is defined as

\[
l^2 = (x_A^1 - x_B^1)^2 + (x_A^2 - x_B^2)^2 + (x_A^3 - x_B^3)^2.
\]

This was expressed by Pythagoras. While this holds true for a cartesian coordinate system, it will not hold true in general for other coordinate systems such as polar coordinates. A slight sophistication is required. Instead of referring to a finite distance, one may only consider an infinitesimal distance, namely \( dl \), defined as

\[
dl^2 = (dx^1)^2 + (dx^2)^2 + (dx^3)^2.
\]
From this, a simple\textsuperscript{1} mathematical generalization to special relativity arises. Einstein may have discovered the theory of special relativity, but it was Minkowski who placed it into a mathematical context by introducing his metric

$$ds^2 = (dx^0)^2 - (dx^1)^2 - (dx^2)^2 - (dx^3)^2.$$  

This may seem like a trivial extension to the Euclidean metric, but there is much involved. The transformations from one inertial reference frame to another via a Lorentz boost

$$\bar{x} = x' + \frac{1}{\sqrt{1 - v^2}}(x_{\parallel} - vt),$$

$$\bar{t} = \frac{1}{\sqrt{1 - v^2}}(t - x \cdot v)$$

will generate numerous terms but once combined, many will cancel out, leaving the metric invariant in the form,

$$ds^2 = (dx^0)^2 - (dx^1)^2 - (dx^2)^2 - (dx^3)^2.$$  

The introduction of the metric was the first step in putting relativity into a mathematical context but there are many underlying subtleties. The Minkowski metric is merely a special form of

$$ds^2 = g_{ij}dx^i dx^j$$

where $g_{ij}$ is the metric tensor, in particular $g_{ij} = \eta_{ij} = \text{diag}(1, -1, -1, -1)$. Other metrics such as

$$ds^2 = dt^2 - dr^2 - r^2 d\phi^2 - dz^2$$

associated with a 4-dimensional cylindrical coordinate system will also describe the same flat spacetime geometry. A flat spacetime will remain flat regardless of which

\textsuperscript{1}Of course, much physics is required but the resulting mathematical step is rather simple.
coordinate system is used. In this example, we can use

\[ x^0 = t \]
\[ x^1 = r \cos \phi \]
\[ x^2 = r \sin \phi \]
\[ x^3 = z \]

to transform from one coordinate system to another. The transformation from one coordinate system to another which preserves its geometry is called a \textit{diffeomorphism}. In general, we can transform the Minkowski spacetime to an infinite set of \( \mathbb{R}^4 \) spaces through a \textit{class of diffeomorphisms} which preserves its flat geometry. This leads us into the topic of differential geometry.

\subsection{Differential Geometry}

A single coordinate system generally cannot cover all of a space of interest. For example, in the polar coordinate system, one cannot work along the line, \( \phi = 2\pi \) for \( 0 \leq r < \infty \). There is a lack of uniqueness associated with that line and differentiations do not exist. This is a coordinate singularity where all the mathematics associated with \textit{this} particular coordinate system breaks down. The coordinate system simply cannot cover this “patch”. In general, multiple coordinate systems are required to cover the whole space.

To resolve such an issue, a generalization of the coordinate system is required. An \textit{n}-dimensional \textit{manifold} consists of an abstract space and a collection of mappings in which the union of its domains covers the entire space. Each mapping takes a point from an open domain in the space into \( \mathbb{R}^n \). \footnote{When we write \( \mathbb{R}^n \), we make no assumptions on the \( \mathbb{R}^n \) space; we do not require it to be flat} These are called \textit{charts}, which are
generalizations of coordinate mappings. A collection of charts makes an atlas. An example of an atlas for a sphere of radius $R$ as a 2-dimensional manifold (embedded in $\mathbb{R}^3$) is the set of mappings:

Chart 1:
\[
\begin{align*}
  x &= f \\
  y &= g \\
  z &= \sqrt{R^2 - f^2 - g^2}
\end{align*}
\quad\leftrightarrow\quad
\begin{align*}
  f &= x & \text{for } z > 0 \\
  g &= y & \text{and } x^2 + y^2 < R^2
\end{align*}
\]

Chart 2:
\[
\begin{align*}
  x &= m \\
  y &= n \\
  z &= -\sqrt{R^2 - m^2 - n^2}
\end{align*}
\quad\leftrightarrow\quad
\begin{align*}
  m &= x & \text{for } z < 0 \\
  n &= y & \text{and } x^2 + y^2 < R^2
\end{align*}
\]

Chart 3:
\[
\begin{align*}
  x &= \sqrt{R^2 - p^2 - q^2} \\
  y &= (p - q)/2 \\
  z &= (p + q)/2
\end{align*}
\quad\leftrightarrow\quad
\begin{align*}
  p &= (y + z)/2 & \text{for } x > 0 \\
  q &= (y - z)/2 & \text{and } y^2 + z^2 < R^2
\end{align*}
\]

Chart 4:
\[
\begin{align*}
  x &= -\sqrt{R^2 - r^2 - s^2} \\
  y &= (r - s)/2 \\
  z &= (r + s)/2
\end{align*}
\quad\leftrightarrow\quad
\begin{align*}
  r &= (y + z)/2 & \text{for } x < 0 \\
  s &= (y - z)/2 & \text{and } y^2 + z^2 < R^2
\end{align*}
\]

or otherwise. In fact, we do not even require a metric to exist. We must think of $\mathbb{R}^n$ as merely an $n$-tuple set of real numbers.
Figure 2.2: This illustrates the tangent space associated with one point on the surface of a sphere. The vectors emerging from the point leave the surface of sphere so one cannot think of a vector existing on the manifold.

Chart 5:
\[
\begin{aligned}
\begin{cases}
x = \sqrt{R^2 - \mu^2 - \nu^2} \\
y = \mu \\
z = \nu
\end{cases}
\end{aligned}
\leftrightarrow
\begin{aligned}
\begin{cases}
\mu = y & \text{for } x > 0 \\
\nu = z & \text{and } y^2 + z^2 < R^2
\end{cases}
\end{aligned}
\]

Chart 6:
\[
\begin{aligned}
\begin{cases}
x = -\sqrt{R^2 - \xi^2 - \zeta^2} \\
y = \xi \\
z = \zeta
\end{cases}
\end{aligned}
\leftrightarrow
\begin{aligned}
\begin{cases}
\xi = y & \text{for } x < 0 \\
\zeta = z & \text{and } y^2 + z^2 < R^2
\end{cases}
\end{aligned}
\]

When one writes down a metric for the surface of a sphere of radius $R$
\[
ds^2 = R^2 du^2 + R^2 \sin^2 u dv^2
\]
(2.1)

with $0 < u < \pi$ and $0 < v < 2\pi$, it is understood that $du$ indicates, in a sense,
an infinitesimal displacement in the \( u \)-direction. If we are given the metric (2.1) for the \( uv \)-coordinates alone (without any embedding picture), we may be misled into thinking that \( du \) exists in the \( uv \)-plane. The term \( du \), called a one-form, is more abstract than just an infinitesimal displacement. To understand what a one-form is, we must introduce its dual, a vector where all the properties of a vector will reflect upon the one-form. As hinted with one-forms, vectors also do not exist on the manifold. This is illustrated in figure 2.2. One can see that the basis vectors \( e_u \) and \( e_v \) exist in their own tangent space. Vectors are not part of the \( uv \)-plane nor are they part of the surface of the sphere. The tangent space in which vectors exists is associated with the point \( p \), denoted as \( T_p(\mathbb{R}^2) \). It should be emphasized that a vector associated with a point \( p \) has no relation\(^3\) to another vector associated with another distinct point \( q \). One cannot add or subtract a vector in \( T_p(\mathbb{R}^2) \) with another vector in \( T_q(\mathbb{R}^2) \). (Inner products are also forbidden but that will become clear upon definition).

We first consider only the inner product between vectors and one-forms. It is clear that one-forms transform like contravariant tensors

\[
du^i = \frac{\partial u^i}{\partial x^j} dx^j.
\]

As mentioned before, vectors are duals of one-forms. So, in order for vectors to transform in a similar manner, we define a vector as a differential operator

\[
\frac{\partial}{\partial u^i} = \frac{\partial x^j}{\partial u^i} \frac{\partial}{\partial x^j}.
\]

A one-form linearly maps a vector to a real number,

\[
du : T_p(\mathbb{R}^2) \mapsto \mathbb{R},
\]

in particular,

\[
du^i \left[ \frac{\partial}{\partial u^j} \right] = \delta^i_j.
\]

\(^3\)without defining "connections".
One can think of the one-form $du^i$ as a tool used to extract the $j^{th}$ component of a vector

$$V = v^j \frac{\partial}{\partial u^j}.$$ 

That is written as

$$du^i[V] = v^j \delta^i_j$$

$$= v^i.$$ 

Inner products of vectors and one-forms are simply an application of the one-form mappings on the vectors.

From the definition of a one-form, it is easy to construct the metric as a bi-linear mapping from a pair of vectors into a real number as

$$g : T_p(\mathbb{R}^2) \otimes T_p(\mathbb{R}^2) \mapsto \mathbb{R}.$$ 

The definition of an inner product between two vectors is through this mechanism: the metric.

To establish a relationship between two “nearby” basis vectors (in different tangent vector spaces but) on the same manifold, we introduce the concept of a connection. A connection $\nabla$ is a rule that assigns a vector $U$ to a differential operator $\nabla_U$ which maps a vector $V$ into another vector $\nabla_U V$. It must follow

$$\nabla_{(fU+gV)} W = f \nabla_U W + g \nabla_V W$$

(2.2)

$$\nabla_U(\alpha V + \beta W) = \alpha \nabla_U V + \beta \nabla_U W$$

(2.3)

$$\nabla_U(fV) = U(f)V + f \nabla_U V$$

(2.4)

where $f$ and $g$ are scalar functions and $\alpha$ and $\beta$ are constants. For a fixed $j$ and $k$, the term

$$\nabla_{e_j} e_k$$
tells us how the vector $e_k$ evolves as it is parallely transported in the direction of $e_j$. This is expressed in terms of another vector, $V = \nabla_{e_j} e_k$. We can decompose this into its components by applying a one-form $e^i$ mapping on $V$. These are real numbers,

$$\Gamma^i_{jk} = e^i \left[ \nabla_{e_j} e_k \right] \tag{2.5}$$

called connection coefficients where $e_j$ are the basis vectors and $e^i$ are their duals. Another way to write (2.5) is

$$\nabla_{e_j} e_k = \Gamma^i_{jk} e_i. \tag{2.6}$$

The coefficients $\Gamma^i_{jk}$ tell us how the basis vectors are connected along coordinate curves.

Given an arbitrary vector

$$W = w^k e_k$$

we can apply the operator $\nabla_{e_j}$ to it, resulting in

$$\nabla_{e_j} W = (e_j \left[ w^k \right]) e_k + w^k \nabla_{e_j} e_k$$

by rule (2.4). In the first term, the basis vector $e_j$, being the operator $\partial/\partial u^j$ is applied to $w^a$ and in the second term, we use (2.6)

$$\nabla_{e_j} W = \left( \frac{\partial w^k}{\partial u^j} \right) e_k + w^k \Gamma^i_{jk} e_i$$

$$= \left( \frac{\partial w^k}{\partial u^j} + w^l \Gamma^i_{jl} \right) e_k.$$  

We denote the term in bracket on right-hand side as the covariant derivative,

$$w^k \gamma_j = \frac{\partial w^k}{\partial u^j} + w^l \Gamma^i_{jl} \tag{2.7}$$

With the definition of a connection established, we can ask how connections can dictate the geometry of a manifold. While a set of diffeomorphism preserves the inherent geometric structure of a manifold, it is not always obvious whether two given metrics describe the same geometry. A more rigorous comparison between the geometries is required based on a derivation of the metrics or the connection coefficients.
Figure 2.3: Curvature is defined through the difference between a vector parallely transported along two different paths ending at the same position.

In other words, we need to define precisely what it means for two manifolds to be geometrically different. As a starting position, consider a one-dimensional curve in a two-dimensional Euclidean plane. It is known that a curve can be characterized solely by the derivative of the tangent vector with respect to its length,

\[ \kappa = \left| \frac{dT}{ds} \right|. \]

For example, if a curve has a constant $1/a^2$ second derivative with respect to its length, it must be a circle of radius $a$, regardless of which coordinate one uses or how the circle is embedded in the 2 dimensional space. This invariant property is universal for all circles in Euclidean space. A generalization to the notion of a "tangent vector derivative with respect to its length" into an n-dimensional manifold is contained in the Riemann curvature.

The Riemann curvature is defined in terms of the difference between a vector parallely transported along one path and the same vector parallely transported along another path, both ending at the same position as seen in figure 2.3. Of course, the
results should only define the local geometry and thus the paths must be infinitesimal in length. As we take the limit as the lengths go to zero, we recover the geometric behaviour of the space. In rigorous terms, this difference in path is is the commutation between two covariant derivatives,

\[ V_{[ij]} = V_{ij} - V_{ji}. \]

The most surprising fact is that the derivatives of \( V \) all cancel each other, leaving only the linear \( V^b \) terms

\[ V_{[ij]} = R^a_{bij} V^b e_a \tag{2.8} \]

where \( R^a_{bij} \) is the Riemann curvature tensor. The value of \( R^a_{bij} \) represents the \( a \)th component of the \( b \)th basis vector parallel-transported along the \( i \)th direction in combination with the \( j \)th direction. One can use (2.7) to write the right-hand side of (2.8) as

\[ R^a_{bij} = \Gamma^a_{b,k} - \Gamma^a_{b,i} + \Gamma^a_{m i} \Gamma^m_{b i} - \Gamma^a_{m i} \Gamma^m_{b j}. \]

The trace of the Riemann tensor is the Ricci tensor defined as

\[ R^a_{baj} = R_{baj}. \]

These definitions and tools of differential geometry form the basis upon which the foundation of general relativity rests. The underlying behaviour of coordinates are dictated by the metric tensor. One does not simply "measure a coordinate" but rather, one measures the results from the metric operation (of bi-linearly mapping of four-vectors). The connection coefficients can be demonstrated physically through their influence upon geodesic paths. These coefficients dictate how spacetime points are connected to each other and thus how geodesics must follow. While the Riemann tensor provides us with a feel as to how much space is curved, it does not influence the entities on the manifold directly. The closest direct physical observable resulting from curvature of spacetime is not through the Riemann curvature tensor, but rather, through the trace of its components, namely the Ricci tensor. It is the Ricci
tensor which has a direct association with the energy-momentum tensor, dictating how spacetime is curved. To define how much spacetime curves will lead us into the next section. But for now, we see that each of these mathematical entities lends itself to some physical consequences of spacetime in the theory of general relativity.

2.1.2 The Action Principle

In the early development of physics as a science, the approaches had been to observe how a physical system evolves and develop a set of hypotheses which may become physical laws. Based on these laws, a set of differential equations can be determined. Solving these exactly will give rise to predictive ability of the model. The classical example would be Brahe’s observations of planetary motion around the sun. From his data\(^4\), Kepler, his student, was able to develop a set of hypotheses which later became known as Kepler’s laws. These laws could be applied to other planets, giving predictive ability.

In modern times such intuitive approaches face overwhelming obstacles. The first obstacle is due to how experimental data are collected and interpreted. The second obstacle is in the lack of knowledge of how parts of the system would react to forces.

In the first problem, data are usually collected in a very indirect way and it becomes less trivial as to how to interpret the data. For example, the mass of the quark is not intuitively defined as the ratio of force to acceleration as applied to it. One can define it as the invariant-mass threshold which particle colliders must exceed in order to produce that particular flavour of quark. However, within the confines of a hadron, there are enormous amounts of virtual sea quarks along with substantial binding energy. Even more complicated, the top quark does not exist long enough to hadronize to form any meson. The fundamental problem is that the mass of a quark does not have a clear-cut interpretation. While one can view a quark mass as

\(^4\)Brahe did not trust his assistant, Kepler, in fear that he may gain prominence over him. So, he only allowed him the data for Mars which was giving him the most trouble.
a parameter within the theoretical model, there is a loss of intuitive development of the system.

The second obstacle is in the complexity of multiple interactions. To combine several types of interactions in an ad hoc manner requires knowledge of how parts of the system will react to forces and pressure in a dynamical environment. A simple example would be a roller coaster sitting on a frictionless platform. Determination of how it moves demands knowing exactly how much force the tracks exert upon the cart at each point in time. An enormous amount of physical insight is required to develop a set of differential equations to describe how the cart will accelerate and how the rails would react. All these complexities diminish the ability of classical intuitive approach in deducing laws and differential equations.

To solve such a dilemma, a structural approach in developing the differential equations was adopted by Lagrange. His approach was a reformulation of classical mechanics based on the extremizing of the action. In Lagrangian mechanics, the action is defined as the time-integral of the difference between the kinetic energy and the potential energy,

\[ A = \int (T - V) \, dt. \]

This action is a functional which maps a set of functions \( x_1(t), x_2(t), \ldots, x_n(t) \) to a real number. That is,

\[ A = A[x_1(t), x_2(t), \ldots, x_n(t)] \]

\[ A : C^2(\mathbb{R}^n \to \mathbb{R}) \to \mathbb{R}. \]

To determine the equations of motion on \( x_1(t), x_2(t), \ldots, x_n(t) \), it is required that any small variation in the functions will keep that action stationary as

\[ \delta A = A[x_1(t) + \delta x_1(t), \ldots, x_n(t) + \delta x_n(t)] - A[x_1(t), \ldots, x_n(t)] \]

\[ = 0 + O(\delta x^2). \]
This condition results in a set of differential equations for $x_1(t), x_2(t), \ldots, x_n(t)$ which dictates its time-evolution. The generality of such an approach provides the structure for more sophisticated systems.

To apply Lagrange's approach to other systems, a more general definition for the action is required. The *action* is defined as a functional which maps the motion of an object or a field to an invariant quantity or scalar quantity. It is done in such a way that the equations of motion can be determined. The principle of stationary action states that the object or field will follow the function(s) which yields a stationary value in the action. That is, upon small variations of the functions, the value of the action remains fixed. The approach in using the action provides a richer structure for determining how a system evolves, not just a classical mechanical system, but for a host of new systems.

**2.1.3 Noether’s Theorem**

Conserved quantities have always played a central role in theoretical physics and one of the most important tools related to these quantities is Noether’s theorem. It provides a systematic approach in determining how conserved quantities are related to the symmetries within the Lagrangian. The theorem states that if the Lagrangian remains invariant\(^5\) under a transformation,

\[
\mathcal{L}(\psi_\alpha(x), \psi_{\alpha,\beta}(x)) \rightarrow \mathcal{L}(\psi_\alpha'(x'), \psi_{\alpha,\beta}'(x'))
\]

then conserved currents exist in the form,

\[
J^\mu_{,\mu} = 0
\]

where

\[
J^\mu = \mathcal{L}\delta x^\mu + \frac{\partial \mathcal{L}}{\partial \psi_{\mu,\alpha}} \left( \delta \psi_\alpha - \psi_{\alpha,\beta} \delta x^\beta \right).
\]

\(^5\)We also require the Lagrangian to be form-invariant as well as scale-invariant; that is, $\int \mathcal{L}dV^4 = \int \mathcal{L}'dV'^4$.\hspace{1cm}
This powerful mathematical tool has been used directly in many fields of theoretical physics such as electromagnetism and classical mechanics. In electromagnetism, it provides us with the means of determining the conserved four-current, \( J^\mu \). Instead of postulating conservation of charge, one postulates the form of the Lagrangian for electromagnetism and Noether’s theorem leads to both the equations of motion and the conservation of charge. In classical mechanics, the theorem provides the continuity equation for energy and momentum based on the four symmetries of a four-dimensional Euclidean geometry: time-translation and the three spatial translations invariance. Before the introduction of Noether’s theorem, the latter was added as a postulate based on empirical facts.

To apply the Noether’s theorem to general relativity is subtle. The fact that spacetime is curved means that generally there may not be any time-translation symmetry and like-wise for any particular space-like symmetries. This creates many problems in determining the conserved quantities in general relativity. While there are various proposed definitions of mass such as the Bondi mass, the ADM mass and the Komar mass, none has resulted in a clear-cut definition of what energy or mass is. \(^6\) The lack of what mass or energy is, has created skepticism in solutions in relativity which includes “negative mass”. This problem has spread to many areas from Hawking radiation to causal structure of spacetime to gravitational waves. For systems which exhibit Killing vectors \( \xi \), a conserved quantity can be written as

\[
E = \xi^\mu p_\mu
\]

where \( p_\mu \) is the four-momentum. The symmetries of general relativity should result in the conservation of the energy-momentum tensor,

\[
T^{\mu\nu} = 0
\]

\(^6\)In a tongue-in-cheek remark, M. Weiss and J. Baez’s answer to the question “Is Energy Conserved in General Relativity?” is “In special cases, yes. In general it depends on what you mean by energy, and what you mean by conserved”.

which is actually the Bianchi identity in disguise. In summary, the concept of energy in general relativity is an active area of research.

### 2.1.4 Variational Formulation

The action for general relativity is assumed to consist of a matter action combined with a gravitational action. The pure gravitational action is added to the matter action as

\[
I = I_M + I_G.
\]

Using the Ricci scalar, defined as

\[
R = R_{ij}g^{ij},
\]

Hilbert, independent of Einstein, was the first to propose the form of the gravitational action as

\[
I_G = \int \sqrt{g} R \, dx^4
\]

where \( g = |\det[g_{ij}]| \).

The principle of least action requires that this is stationary for variations of the metric, written as

\[
\delta I_G = 0
\]

Assuming a Dirichlet boundary condition, we can write this as

\[
\begin{align*}
\delta I_G &= \int \delta(\sqrt{g} R_{ij}g^{ij}) \, dx^4 \\
&= \int \left\{ R(\delta \sqrt{g}) + \sqrt{g}g^{ij}(\delta R_{ij}) + R_{mn} \sqrt{g} (\delta g^{mn}) \right\} \, dx^4
\end{align*}
\] (2.9)  

(2.10)

The Jacobi identity

\[
d(\det[A]) = \text{Tr}[\text{adj}(A)dA] \\
= \det(A)\text{Tr}[A^{-1}dA],
\]
leads to

\[ \delta \sqrt{g} = \frac{1}{2} \sqrt{g} g^{ij} \delta g_{ij}. \]  

(2.11)

Applying the variation to \( g^{im} g_{mn} = \delta^i_n \), we can write

\[ \delta g^{mn} = -g^{mi} g^{nj} \delta g_{ij}. \]  

(2.12)

To find \( \delta R_{ij} \) we first write

\[ R_{ij} = \Gamma^a_{ai,j} + (-\Gamma^a_{ij,b} + \Gamma^a_{ib} \Gamma^b_{j} - \Gamma^m_{ij} \Gamma^a_{bm}) \delta_a^b \]

\[ \delta R_{ij} = \delta \Gamma^a_{ai,j} + (-\delta \Gamma^a_{ij,b} + (\delta \Gamma^a_{ib}) \Gamma^b_{j} - (\delta \Gamma^m_{ij}) \Gamma^a_{bm} \delta_a^b - (\delta \Gamma^a_{am}) \Gamma^m_{ij} \]

\[ = \{ (\delta \Gamma^a_{ai})_{,j} - (\delta \Gamma^a_{am}) \Gamma^m_{ij} \} \}

\[ = \{ (\delta \Gamma^a_{ai})_{,j} - (\delta \Gamma^a_{im}) \Gamma^m_{ij} \} \}

\[ = (\delta \Gamma^a_{ai})_{,j} - (\delta \Gamma^a_{im}) \Gamma^m_{ij} \}

Because the metric tensor can pass through the covariant derivative, it is possible to write

\[ \sqrt{g} g^{ij} \delta R_{ij} = (\sqrt{g} g^{ij} \delta \Gamma^a_{im})_{,j} - (\sqrt{g} g^{ij} \delta \Gamma^a_{im})_{,a} \]  

(2.13)

The two terms on the right-hand side look like a divergence of the form

\[ v^i_{,j} = \frac{1}{\sqrt{g}} (\sqrt{g} v^i)_{,j}. \]  

(2.14)

We can write (2.13) as

\[ g^{ij} \delta R_{ij} = \frac{\partial}{\partial x^j} (\sqrt{g} g^{ij} \delta \Gamma^a_{im}) - \frac{\partial}{\partial x^a} (\sqrt{g} g^{ij} \delta \Gamma^a_{ij}) \]  

(2.15)

Using (2.11), (2.12) and (2.15), equation (2.10) becomes

\[ \delta I_G = \int \{ R \delta \sqrt{g} + \sqrt{g} g^{ij} \delta R_{ij} + R_{mn} \sqrt{g} \delta g^{mn} \} \ dx^4 \]

\[ = \int \{ R \left( \frac{1}{2} \sqrt{g} g^{ij} \delta g_{ij} \right) + \sqrt{g} \left( \frac{\partial}{\partial x^j} (\sqrt{g} g^{ij} \delta \Gamma^a_{im}) - \frac{\partial}{\partial x^a} (\sqrt{g} g^{ij} \delta \Gamma^a_{ij}) \right) \]

\[ + R_{mn} \sqrt{g} \left( g^{mi} g^{nj} \delta g_{ij} \right) \} \ dx^4 \]
Recall that we have assumed a dirichlet boundary condition so the total-divergence terms in the middle vanish, leaving

$$\delta I_G = \int \left( \frac{R}{2} g^{ij} - R_{mn} g^{m_i g_{n^j}} \right) \delta g_{ij} \sqrt{g} \, dx^4.$$

Through proof by contradiction\(^7\), the integrand must be zero, thus resulting in the vacuum Einstein equations,

$$\tilde{R}^{ij} - \frac{1}{2} g^{ij} R = 0.$$

The left-hand side is known as the Einstein tensor, written as

$$G^{ij} = R^{ij} - \frac{1}{2} g^{ij} R. \quad (2.16)$$

### 2.2 Axial Symmetry

General relativity’s set of non-linear partial differential equations are enormously complicated with the exception for a few special cases. Axially-symmetric systems play an important role in this thesis. Assuming such a symmetry reduces the Einstein equations tremendously. In 1917, Weyl\(^2\) considered a class of the Einstein equations which is static and has axial symmetry. In natural units where c, the speed of light is 1, starting from the Weyl metric of the form

$$ds^2 = e^{2\nu} dt^2 - e^{2(w-v)} (dr^2 + dz^2) - e^{-2\nu} r^2 d\phi^2$$

\(^7\)This is also known as the Reymond-duBois lemma.
where \( v \) and \( w \) are functions of only \( r \) and \( z \), the components of the (covariant) Einstein tensor (2.16) become

\[
G_{00} = e^{4v-2w} \left( w_{zz} + w_{rr} + v_r^2 + v_z^2 - 2v_{zz} - 2v_{rr} - \frac{2}{r}v_r \right)
\]

\[
G_{11} = v_r^2 - v_z^2 - \frac{1}{r}w_r
\]

\[
G_{22} = -r^2 e^{-2w} (w_{zz} + w_{rr} + v_r^2 + v_z^2)
\]

\[
G_{33} = \frac{1}{r} w_r - v_r^2 + v_z^2
\]

\[
G_{13} = 2v_r v_z - \frac{1}{r}w_z
\]

where subscripts mean partial derivative and we have designated \( x^0 \equiv t \), \( x^1 \equiv r \), \( x^2 \equiv \phi \) and \( x^3 \equiv z \). In vacuum, they simplify to

\[
v_{zz} + v_{rr} + \frac{1}{r}v_r = 0
\]

\[
w_r + r(v_z^2 - v_r^2) = 0
\]

\[
w_z - 2rv_r v_z = 0.
\]

The function \( v(r, z) \) acts like a gravitational potential because the first equation is simply the Laplace equation in an axially symmetric form. The next two equations state that the potential, \( v(r, z) \) is the generator of the function \( w(r, z) \).

One possible solution is a pair of gravitating particles along the \( z \)-axis

\[
v = -\frac{m_1}{r_1} - \frac{m_2}{r_2}
\]

where \( r_1^2 = r^2 + (z - z_0)^2 \) and \( r_2^2 = r^2 + (z + z_0)^2 \). The other function, \( w \), can be integrated as

\[
w = \int r(v_z^2 - v_r^2) \, dr + \int 2v_r v_z \, dz
\]

\[
= -\frac{m_1^2 r}{2r_1^4} - \frac{m_2^2 r^2}{2r_2^4} + \frac{m_1 m_2}{2z_0^2} \left( \frac{1}{r_1 r_2} (r^2 + z^2 - z_0^2) - 1 \right)
\]
Initially, this solution to the Einstein equations was thought to indicate a failure of general relativity because it describes two gravitating masses that are static with respect to each other; that is, they are not falling towards each other. However, upon closer inspection, particularly at the axis of symmetry \( r = 0 \), there is a lack of *elementary flatness*. That is, just as with a tip of a cone, if one computes the ratio between a circle’s circumference to its radius, the limit of the ratio (as \( r \to 0 \)) does not converge to \( 2\pi \). The failure of elementary flatness shows that the two particles are held apart by a singular strut which shows up as a singularity in the energy-momentum tensor.
Chapter 3

Non-Causal Structures

A closed time-like curve (CTC) is a spacetime curve that is always future-directed, proceeding into the forward lightcone and finally connecting to the point where it reconnects with the spacetime point of an earlier event in its history. In this chapter, we will examine the concept of a closed time-like curve and the nature of causality in general relativity. In the first section, we will introduce the history of causality and how it remains steadfast through the introduction of relativity and quantum mechanics. In the second section, we will examine the classical Gödel spacetime and determine its essential property leading to the existence of CTCs. In the third section, we will consider the Gott spacetime, another configuration which contains CTCs. While these two examples are based on different mechanisms for producing CTCs, we tie them together in a unifying theme on how CTCs come about.

3.1 Introduction

The idea of time-travel has been around since ancient times. When philosophers speculated about the notion of time, they eventually arrived at the question of whether
time-travel is possible and the paradox associated with it. Before and during the era of Newton, the idea of time was simply an axis in one direction with no peculiarity associated with it. There were no physical principles which would prevent time-travel. However, the understanding of the "time" axis changed drastically when Einstein discovered relativity. The same questions that the ancients had asked applies to the new structure of space and time.

In the introduction of special relativity, there is a mixing of space and time via the Lorentz boost,

\[
\bar{x} = \gamma (x - vt)
\]
\[
\bar{t} = \gamma (t - vx).
\]

This mixing creates many paradoxes through thought experiments. One may think that time travel into the past might be possible by some prescription of transforming a time-directed segment into a spacetime mixture, traversing the space “backward” and then transforming back to the original time direction. However, it is easily shown that a time-like segment stays time-like no matter how one applies the transformation. Other paradoxes relating to time such as the twin paradox arise but once placed in a mathematical context, each paradox can be easily resolved.

An interesting example which shows an apparent causality-violating structure in flat spacetime is the uniform accelerated reference frame. [41] It can be shown that applying varying instantaneous Lorentz boosts for a uniformed accelerated observer results in the coordinate transformation

\[
\bar{x} = -\frac{1}{\kappa} + \left( x + \frac{1}{\kappa} \right) \cosh \kappa \tau
\]
\[
\bar{t} = \left( x + \frac{1}{\kappa} \right) \sinh \kappa \tau
\]
Figure 3.1: This shows the many instances of the accelerated observer’s $x$-axis. It pivots around the origin, a distance $1/a$ from the observer. All events to the left of the pivot point occurs in the reverse order.

where $\kappa = a$. The metric becomes

$$ds^2 = dt^2 - dx^2$$

$$= (1 + \kappa x)^2 dt^2 - dx^2.$$  

As an observer accelerates to the right shown in figure 3.1, his $x$-axis pivots around a point a distance $1/a$ away causing all events to the left of that point\(^1\) to occur in reverse order, in a sense, traveling backward in time. However, causality has not been violated because all events beyond that point are not within the past light-cone of

\(^1\)In fact, the distance between this pivot point and the accelerated observer remains constant as he continues to accelerate away from it! One can think of this effect as being due to length-contraction counteracting the distance traveled by the observer. For example, accelerating at 9.8 m/sec\(^2\) the pivot point is $9.18 \times 10^{15}$m or 0.97 light years in the opposite direction. This can be calculated by comparing the classical velocity of an accelerated observer $v + at$ and the change in length-contraction, namely $d(L/\gamma)/dt \approx -avL$. For the leading-order terms to counteract each other, the condition $L = 1/a$ must hold.
the observer.

Even with the introduction of non-trivial behaviour of time through special relativity, our intuition of the nature of time has remained firm. Special relativity has preserved causality in the thought experiments we invoked. Incidentally, all of these thought experiments are conducted at the macroscopic level in flat spacetime. At this macroscopic level, we cannot see or sense forward into the future. One may argue that our senses are too limited to detect information coming from the future in a microscopic level. To examine the causal structure of spacetime at a microscopic level, we turn to quantum mechanics.

The same questions of whether information flows from the future can be asked in the framework of quantum mechanics. Because quantum mechanics blurs the spatial as well as the temporal position of a particle, it becomes the focus of attention. The example of possible causality-violation in flat spacetime relates to the Heisenberg Uncertainty Principle. Due to the uncertainty in a particle’s position, two events in its world-line may be space-like separated. For example, one observer may see an electron being emitted at event $p_1$ and then later absorbed at $p_2$ with the two having a space-like separation. While a time-like separation will retain its ordering for any Lorentz-boosted inertial reference frame, a space-like separation will not have ordering preserved. This means different observers may see the events $p_1$ and $p_2$ in different order. One observer may see an electron being absorbed before it is emitted! The only resolution[17] to this paradox is that the second observer detects a positron emitted at point $p_2$ and then absorbed at $p_1$. Furthermore, if we had a particle in a plane-wave configuration in which the momentum is known almost precisely but the position of the particle can jump from one location in the universe to another almost instantaneously, causality is still preserved.

Another test of causal nature of spacetime was questioned when Feynman introduced his path-integral approach to quantum field theory. [31] He originally believed that a particle such as the electron “does anything it likes. It goes in any direction at any speed, forward or backward in time, however it likes and then you add up the
amplitudes and it gives you the wave-function. However, in the calculation of the transition amplitude

$$\langle q_f, t_f | q_i, t_i \rangle = N \int \mathcal{D}q \exp \left( \frac{i}{\hbar} S[q] \right)$$

where

$$\mathcal{D}q = \lim_{n \to \infty} \prod_{j=1}^{n} dq_j$$

the action is

$$S = \int \mathcal{L}(q, \dot{q}) \, dt$$

and upon expansion of the propagator contained in terms on the right-hand side, each term admits a time-ordered operator which enforces causal structure of space-time. Particle accelerators and high precision detectors produce results which agree extremely well with Quantum Field Theory calculations based on this. Had information flowed from the future, these machines would have shown much different characteristics. In short, we simply cannot see the "thickness" of any time-slices at the microscopic level.

These examples show that even though causal structure clearly seems to be violated, some other mechanism comes into play to preserve it. It strongly suggests that causality is a principle which we intuitively feel should always hold true just like the conservation of charge even though there is nothing (except for empirical fact) that forces it. Common sense has taught us about cause and effect even before the introduction of quantum mechanics and relativity.

With the introduction of general relativity, the idea of curvature, being the dominant mechanism, seems like a likely candidate to create a universe in which time-travel was possible. One of earliest solutions of the Einstein equations which contains closed time-like curves was that of Van Stockum's solution[4] in 1937. It describes an infinitely long cylinder of dust rotating rapidly with the centrifugal force being balanced exactly by its gravitational attraction. The interior metric for \( r < r_0 \) is

$$ds^2 = dt^2 - e^{-\omega^2 r^2} (dr^2 + dz^2) - r^2(1 - \omega^2 r^2)d\phi^2 - 2\omega r^2 d\phi dt.$$
A closed time-like curve is defined as curve which has a real space-time interval,

\[ ds^2 > 0 \]

that eventually runs back into its own starting point. In Van Stockum's solution, with \( \phi = 0 \) and \( \phi = 2\pi \) being identified and the component \( g_{\phi\phi} > 0 \) for \( 1 < \omega^2 r^2 \), the \( \phi \) curve becomes a CTC.

In 1949, Gödel introduced a solution to the Einstein equation in which one has the ability to "travel into the past". His announcement was made at an IAS lecture in Princeton where renowned scientists such as Einstein, Oppenheimer and Chandrasekhar were in the audience. Chandrasekhar was influenced by his presentation and followed up on calculations of the geodesics. [9] He and Wright independently found that the geodesics were not CTCs. In 1970, [14] Stein published a paper in defense of Gödel, stating that Gödel had never claimed that his CTCs were geodesics, but that they could exist if there was a simple mechanism for changing the trajectory such as using a rocket. The renewed spark of interest arose from the inclusion of the Gödel solution in [19] Hawking and Ellis's famous book called "The Large Scale Structure of Spacetime" in 1973. It contained an intuitive diagram showing the tilting of light cones in spacetime with null and closed timeline curves. Subsequently, many [42][22][39] authors wrote about the concepts of CTC and time machines. Some have even taken the CTC notion so seriously as to propose experiments to search for the presence of CTCs in nature.

Many paradoxes will arise upon the existence of CTCs. One can travel into the past and affect the outcome of the future leading to logical contradictions. Unlimited computing resources will suddenly become available through the use of CTC. A computer which had a limited amount of computing capability can compute anything through trial-and-error and suddenly have an unlimited computing capability. Furthermore, the issue of entropy flow would violate the second law of thermodynamics, one of the holy grails of physics. Such paradoxes fly in the face of the laws of physics that we know to hold true.
One could trivially create a CTC in flat spacetime by identifying certain spacetime points\cite{58}. Because it is so trivial, no one has ever tried to establish a connection from this clearly artificial and purely mathematical phenomenon with the supposedly physical CTCs of the Gödel spacetime and of others such as Gott. \cite{59}It is our contention that the connection is very direct indeed, that the CTCs of Gödel and others simply follow from the identification of spacetime points, that they are in effect man-made rather than the consequences of exotic gravity via general relativity.

### 3.2 Gödel’s Closed Time-Like Curves

The difficulty in disallowing the existence of a CTC is the fact that spacetime at $\theta = 0$ and $\theta = 2\pi$ are identical and that there is nothing preventing one from stitching one part to another. For the time being, consider the simple case of a particle tracing a path of a circle in flat space. At a specified time, say, 1PM, the particle is at $\theta = 0$ but arrives back at $\theta = 2\pi$ at a later time. One can simply identify for the particle not only its spatial coordinates $\theta$ at the two extremities as one does automatically but also the time coordinates at the extremities 1 PM and 2 PM which one would normally never do. However, the latter will certainly lead to a CTC being produced. Spatial points are identified upon such a cyclical path in conformity with our physical experience, but to do so with temporally separated points goes against our physical intuition.

The Gödel spacetime describes a spacetime of a rotating universe containing a cosmological constant. The metric is in a generic class\cite{51} given by

$$\begin{align*}
ds^2 = -f^{-1}[\epsilon^{\nu}(dz^2 + dr^2) + r^2 d\phi^2] + f(d\tilde{t} - w d\phi)^2
\end{align*}$$

(3.1)

where $f$, $\epsilon$, and $w$ are functions of $r$ and $z$ with the coordinates having the ranges

$$\begin{align*}
-\infty < z < \infty, \quad 0 \leq r, \quad 0 \leq \phi \leq 2\pi, \quad -\infty < \tilde{t} < \infty
\end{align*}$$

(3.2)
and with $\phi = 0$ and $\phi = 2\pi$ being identified as usual. The metric component

$$g_{\phi\phi} = -f^{-1}\left(r^2 - f^2w^2\right)$$ (3.3)

has different signs for certain fixed $r$ and $z$ values. There is no difficulty in deducing the nature of the curve

$$\tilde{t} = \tilde{t}_0, \quad r = r_0, \quad \phi = \phi, \quad z = z_0$$ (3.4)

with $z_0$, $r_0$, $\tilde{t}_0$ held fixed with the condition

$$f^2w^2 < r^2$$ (3.5)

as being a closed spacelike curve. However, for the condition that

$$f^2w^2 > r^2$$ (3.6)

$\phi$ becomes a timelike coordinate. In such case, the curve is seen to be a CTC as a result of the now-timelike coordinate $\phi$ having $\phi = 0$ and $\phi = 2\pi$ still being identified as was the case when $\phi$ was spacelike. The condition (3.6) implies that the metric has two timelike coordinates $\tilde{t}$ and $\phi$. Bizarre as it may appear, there is nothing mathematically wrong in coordinatizing a normal spacetime with more than one timelike coordinate (Synge[11] describes a situation where a normal spacetime is described with four timelike coordinates).

As we start to see the theme of connecting mathematics to physics, the essence is in the physical interpretation. For $\phi$ being a spacelike coordinate, the identification of spacelike separation does not contradict logic from our understanding of the azimuthal spatial symmetry. However, the continuation of identifying the $\phi$ values of 0 and $2\pi$ when it becomes timelike is in question; our experience with time is that it is non-periodic. The interpretation becomes suspect when a timelike coordinate does not advance in the description of a timelike curve for which the physical proper time must necessarily advance.
For a better grasp of the interpretation at hand, we present a simple flat spacetime in cylindrical polar coordinates with the metric

\[ ds^2 = dt^2 - dr^2 - r^2 d\phi^2 - dz^2 \]  

(3.7)

using the standard coordinate ranges and where \( \phi = 0 \) and \( \phi = 2\pi \) are identified as usual, i.e.

\[ (t, r, 0, z) = (t, r, 2\pi, z). \]  

(3.8)

Retaining the identification in \( \phi \) for 0 and \( 2\pi \) as we effect the transformation

\[ \bar{t} = t + a\phi, \quad \bar{\phi} = \phi, \quad \bar{r} = r, \quad \bar{z} = z \]

where \( a \) is a constant, the metric becomes

\[ ds^2 = d\bar{t}^2 - d\bar{r}^2 - 2a d\bar{t} d\bar{\phi} - (r^2 - a^2) d\phi^2 - dz^2. \]  

(3.9)

which is in the same form as (3.1) but with constant values for \( f, w \) and \( \nu \). The usual approach is to consider (3.9) and identify in the following manner:

\[ (\bar{t}, r, 0, z) = (\bar{t}, r, 2\pi, z). \]  

(3.10)

A peculiarity becomes apparent as the sign of the \( g_{\phi\phi} \) component of (3.9) changes from negative to positive for \( r^2 < a^2 \), shifting its characteristic to a timelike curve. Furthermore, the imposed closure condition given by (3.10) creates a CTC.

By the analysis of the lightcones, it is straightforward to develop the standard figure 3.2 depicting the transition from closed spacelike to null to timelike curves. The result is a diagram similar to that which displays the curves of the Gödel universe as shown in the standard texts (see for example [19]). Figure 3.2 would indicate that for \( r_0 < a \), there are closed timelike curves.

It is indeed seen that the \( \phi \)-curve is a CTC for a fixed \( r_0 < a \). The curve never reverses direction into the past lightcone, remaining timelike and hence the proper time flows monotonically and never becomes imaginary. If we transform the "cylindrical
coordinates” \((\bar{t}, r, \phi, z)\) into the more familiar “cartesian coordinates” \((\bar{t}, \bar{x}, \bar{y}, \bar{z})\), we find that the \(\phi\) curve follows the trajectory

\[
\begin{align*}
\bar{t} &= \bar{t}_0 \\
\bar{x} &= r_0 \cos \phi \\
\bar{y} &= r_0 \sin \phi \\
\bar{z} &= z_0 \\
\end{align*}
\]

\(ds^2 > 0 \) (time-like) \(\forall \phi \in [0, 2\pi]\)

and this timelike curve returns to the original location in spacetime as a CTC.

However, we recall that the original spacetime with metric (3.7) is simply ordinary flat spacetime. The metric (3.9) was derived simply from a coordinate transformation. One may wonder as to how flat spacetime can contain CTCs. Upon closer
Figure 3.3: The boxes in the figure are now at constant $t$. In the $(t, \phi)$ coordinate system, the spacelike, null and timelike curves are seen as a unified family of curves advancing monotonically in time $t$. Evolving curves never close in terms of $t$ and so there are no CTCs with the periodic time restriction removed. The fixed $\bar{t} = \bar{t}_0$ surface is actually helicoidal in this case.
inspection we see that the key to the characteristic change in the spacetime lies in the identifications. The identification (3.8) is not equivalent to (3.10). To understand how one choice is more natural than the other, consider the transformation of the light cones of figure 3.2 back into the original fiducial \((t, r, \phi, z)\) coordinates. This is illustrated in figure 3.3. The reason for the apparent tilting of these lightcones in figure 3.2 with respect to the \((\bar{t}, \phi)\) coordinates as \(r\) varies is clear: The curves \(t + a\phi = \bar{t}_0, r = r_0, z = z_0\) being helices, are inside and outside the lightcone for \(r_0 < a\) and \(r_0 > a\) respectively.

It would not be too far-fetched to believe that Gödel's intention in deriving his solution to the Einstein equations was to introduce CTCs. He may be employing a system that, based on its underlying structure, forces the identifications. For example, it is known that a surface of constant positive curvature must uniquely be a sphere and thus identifications do not allow room for choices. But the conditions in which local geometry dictates global structure is far and few. It is only in rare instances where such conditions occur. In this particular case, if we compare the metric of the \(t - \phi\) sub-manifold of the Gödel spacetime,

\[ ds^2 = A_0d\phi^2 + B_0d\phi dt + dt^2, \]

where \(A_0 = (\sinh^4 r_0 - \sinh^2 r_0)\) and \(B_0 = 2\sinh^2 r_0\), to a spherical surface,

\[ ds^2 = r_0^2d\theta^2 + r_0^2\sin^2 \theta d\phi^2, \]

paying particular attention to the \(\theta\)-dependence, we find that it closely resembles our metric

\[ ds^2 = (a^2 - r_0^2)d\phi^2 - 2ad\phi d\bar{t} + d\bar{t}^2 \]

more than it does, the spherical case. We know that our cylindrical-like metric does not have any abstract structure which enforces identifications. This suggests that the identification that Gödel made was not forced by the geometry.

In fact, in Gödel’s paper, the original metric was written in the form of

\[ ds^2 = a^2 \left( d\bar{t}^2 - d\bar{r}^2 + \frac{1}{2}e^{2\sigma}d\phi^2 + 2e^{\sigma}d\bar{t}d\phi - d\bar{z}^2 \right) \]

(3.11)
and is expressed with timelike coordinates $\tilde{t}, \tilde{\phi}$ globally. Here, the underlying 3+1 character of the spacetime is hidden. It is advantageous to have the metric expressed in a form that displays the 3+1 character explicitly. This can be achieved with the transformation

$$
\tilde{t} = t + \frac{r\phi}{2} (1 - \ln r) + \frac{1}{2} \ln r,
$$

$$
\tilde{r} = r\phi, \quad \tilde{z} = z,
$$

$$
\tilde{\phi} = -\frac{1}{2} e^{-r\phi} \ln r
$$

and the metric becomes

$$
\frac{ds^2}{a^2} = dt^2 - \left[ \phi^2 + \frac{1}{8r^2} (r\phi \ln r - 1)^2 \right] dr^2 - \left[ \frac{3}{4} r^2 + \frac{1}{8} (r \ln r)^2 \right] d\phi^2 - dz^2
$$

$$
- \frac{1}{4} (8r\phi + r\phi (\ln r)^2 - \ln r) dr d\phi + r dt d\phi.
$$

It is to be noted that in the process, $\phi$ dependence in the metric appears\(^2\). The identification

$$(\tilde{t}, \tilde{r}, 0, \tilde{z}) = (\tilde{t}, \tilde{r}, 2\pi, \tilde{z})$$

is transformed to

$$(t, 1, \phi, z) = (t + 2\pi (1 - \phi) e^\phi, e^{-4\pi e^\phi}, \phi e^{4\pi e^\phi}, z)$$

and in this form, there is no suggestion of any identification of spacetime points that would yield closure in this explicitly 3+1 coordinate system.

\(^2\)Because of this $\phi$ dependence, it cannot be considered a periodic coordinate. Forcing an identification would result in a severe discontinuity of the metric. The main advantage of this form is its explicit 3+1 nature.
Figure 3.4: In the left figure, the identifications after the removal of the strip \(-1 < \bar{x} < 1\) are shown using horizontal line-segments. The right figure illustrates the same identification of points after the Lorentz boost.

### 3.3 Gott's Closed Time-Like Curves

In the Gott spacetime, two cosmic strings are boosted in opposite directions via the Lorentz transformation but with an offset so as not to collide. In certain configurations, a curve which encloses both strings is seen as a CTC. The peculiar situation starts with the fact that a cosmic string creates a deficit in the azimuthal angle. This creates a wedge of void space in the spacetime region which requires identifications on either side of the wedge to enclose the void. While the wedge itself is not important, what is important is the identifications.

To understand how this works, we only consider the strip of void space where identifications are made on either side. We start with a flat 1+1 spacetime in which a strip \(-1 < \bar{x} < 1\) is removed and the points identified are \((\bar{t}, -1)\) and \((\bar{t}, 1)\). For this example, we shall call the region \(\bar{x} < -1\) the negative side and the \(\bar{x} > 1\) the positive side. If one applies a Lorentz boost from \((\bar{t}, \bar{x})\) to \((t, x)\) before identifying the
points, the two edges of the cut as seen in the new coordinate system will "slip" as shown in figure 3.4.

With this slip in spacetime point identification, any object from the positive side crossing the strip to the negative side will be displaced back in $t$ value as seen in figure 3.5. While one can think of this as a transition strip for traveling through time, causality is not violated because any attempts to close the worldline would require another transition in the opposite direction. Traveling through this strip in the opposite direction would have the $t$ value increased by the same amount. Thus, no events in the future of $e_2$ coincide with $e_1$, i.e. it is impossible to return to the initial event via a timelike trajectory. The identification of points will close the gap leaving no evidence of tampering.

A being living in this spacetime should not be able to find any such evidence but we will see that this rests fully on the order of operation. As we have seen, the identifications were made after the Lorentz transformation, creating the slip. If the identifications were made before the Lorentz transformation is applied, the spacetime would be continously connected particularly in the $t$ direction and there would be no slip. That is, all events would be mapped smoothly from one coordinate to another without any jump in "time". For now, the choice of whether to close the gap before or after applying the Lorentz boost seems inconsequential but this choice in the order of operation will be the key feature to be employed in the next example.

To see how a CTC is created using the two boosted strings, we introduce a new $\tilde{y}$ dimension. In this 2+1 system $(\tilde{t}, \tilde{x}, \tilde{y})$, instead of having one strip $-1 < \tilde{x} < 1$, $-\infty < \tilde{t} < \infty$ removed for $\tilde{y} = 0$, we will consider two parallel strips removed. We call the "front" strip

$$-1 < \tilde{x} < 1, \quad -\infty < \tilde{t} < \infty, \quad \tilde{y} = y_1$$

where $y_1$ is a positive constant and the "back" strip

$$-1 < \tilde{x} < 1, \quad -\infty < \tilde{t} < \infty, \quad \tilde{y} = -y_1.$$
Figure 3.5: This shows a possible worldline of a massive object as it crosses the identified strip. Events $e_1$ and $e_2$ are identified. In this coordinate, the $t$-value of $e_2$ is less than that of $e_1$ so one can say that $e_2$ occurred before $e_1$.

The Lorentz boost with velocity $+\beta_s$ is applied in the positive $\bar{x}$-direction for half of the space $\bar{y} \equiv y \geq 0$. The other half has a boost in the negative $\bar{x}$-direction with the same velocity magnitude. To put it simply, each strip will be Lorentz-boosted in opposite directions and the half-spaces are pasted together. This is possible because the two half-spaces are flat and hence there are no jumps at the stitch-plane.

We construct the CTC by considering the path of a traveller starting at event $E_1$ on the right side of the front strip shown in figure 3.6. As the traveller crosses the Lorentz-boosted strip to event $E_2$, he travels “back in time”. With $y_1$ sufficiently small, he could proceed via a timelike path to the back strip at event $E_3$. To return to his initial starting point, he crosses over the back strip, to event $E_4$. From $E_4$, he takes timelike trajectory to return to event $E_1$. His worldline is a CTC. This illustration captures the essential mechanism of the Gott-produced CTC[39].

The jump conditions in $t$ are used to create the CTC in the Gott spacetime. As noted above, the order of operation dictates whether there are jumps in $t$-values which ultimately determine the existence of CTC in this spacetime. The choice of
identifying the events across the strips before or after the boost can be made to suit the goal. Gott chose the latter in order to realize the CTC. But the question is which order reflects the more natural approach in dealing with such a system. To answer that, consider a cosmic string at rest. One expects continuity and axial symmetry even though there is an angle-deficit. That is, the "wedge" that is removed should not be detectable and there should be no discontinuity either in $t$ or in $(x, y)$. Next, consider the results of a Lorentz boost of this cosmic string. The continuity should be maintained even though the axial symmetry is lost. This expectation of maintained continuity requires that the gap be closed before the Lorentz boost is applied. That is, the natural choice is to identify the events across the strips before the boost. Even if continuity was lost, the direction of the wedge should be in a symmetric alignment, either towards or away from the direction of boost, rather than perpendicular as Gott has chosen. In either case, it rules out the closed timelike curve as envisaged by Gott.
Figure 3.7: The lightcones in the pq-plane.

3.4 Discussions

In both the Gödel spacetime and the Gott spacetime, a choice was made to create the CTC. In the former, identification of endpoints of a time-like curve was the source of the CTC. In the latter, the choice in the order of operations created a configuration which allows for the realization of the CTC. The question of which choice is more natural becomes the issue. In both cases, we had argued that the choices were improper.

The problem with such arguments concerning which choice is more natural is that each must be examined on a case-by-case basis and that there is no underlying principle to go by. Once we are given a metric, it is not clear as to whether a periodic
condition being imposed upon a coordinate is proper or not. Take for example the standard $2 + 1$ Minkowski spacetime with metric

$$ds^2 = dt^2 - dx^2 - dy^2$$

(3.12)

with $-\infty < t < \infty$, $-\infty < x < \infty$ and $-\infty < y < \infty$. Applying the transformation

$$p = x \cos t,$$

$$q = x \sin t,$$

$$y = y$$

(3.13)

the metric becomes

$$ds^2 = \left( \frac{q^2}{(p^2 + q^2)^2} - \frac{p^2}{p^2 + q^2} \right) dp^2 + \left( \frac{p^2}{(p^2 + q^2)^2} - \frac{q^2}{p^2 + q^2} \right) dq^2$$

$$-2pq \left( \frac{1}{(p^2 + q^2)^2} + \frac{1}{p^2 + q^2} \right) dp dq - dy^2.$$ 

(3.14)

Consider the parametric curve $(p, q, y) = (\cos \tau, \sin \tau, 0)$ with $\tau$ running from 0 to $2\pi$. In the $pq$-plane this curve is closed because it starts at $(p, q, 0) = (1, 0, 0)$ and returns back to the same point $(1, 0, 0)$. Furthermore, one can substitute this into (3.14) or simply match it with (3.13) to show that $ds^2 = +d\tau^2$ making it a CTC. One can see this easily in figure 3.7. Suppose we were not given (3.12) but only the metric (3.14) in which we can deduce figure 3.7. There is no indication\(^3\) of a choice of identifying or not identifying points being made. In this case, we cannot even argue which “choice” is more natural.

Now, the question is whether we truly have a CTC or not. Imagine two people solving the Einstein field equations. One person finds

$$ds^2 = dt^2 - dr^2 - r^2 d\phi^2$$

(3.15)

---

\(^3\)The singularity at $(p, q) = (0, 0)$ may become suspect, but every coordinate which has a periodic nature produces a coordinate singularity.
as a solution while the other finds (3.14) as another solution. The first person concludes that there are no CTCs at all while the second concludes that all curves which enclose the $y$-axis are CTCs. One might argue that the second system is peculiar because the metric has a singularity on the $y$-axis; that is, at $(p, q, y) = (0, 0, y_0)$. However, the metric (3.15) also has a singularity. Therefore, who is right? Because we have a very simple system in this case, we can easily see that both are correct. The fact is that we are dealing with two different spaces. The first person sees the second person doing an unnecessary identification at $t = 0$ and $t = 2\pi$. The second sees an unnecessary continuation of space; i.e., a curve $p = \cos \tau, q = \sin \tau$ should end at $\tau = 2\pi$ and should not continue. The essential point is that closure of time-like curves can be introduced if one wishes\textsuperscript{4} but the Einstein field equations do not require it to be either case.

It is essential not to lose sight of the fact that in the study of CTCs, our experience in nature has already been imposed prior to any analysis. The demand that curves of causally connected events always evolve into the forward lightcone and stay in that direction throughout, is intuitive. While it is very natural that a spatial coordinate may become periodic, it goes against our experience when a time-like coordinate becomes periodic through identification of the endpoints.

\textsuperscript{4}These arguments do not conflict with \cite{Novikov}\ Novikov's consistency principle.
Chapter 4

Galactic Structures

From small scale astronomical process such as planetary evolution, to medium scale stellar formation, to larger scale galaxy morphology and to the largest scales in creation of cosmological structures, gravitation plays the key role in all the dynamics. Planets are formed from gravitational attraction of gas, dust particles and remnants from previous supernova explosions. In stellar evolution, the two most significant events in a star's life are in its birth where gravity dictates whether sufficient pressure is available to trigger nuclear fusion and at the ending of its life in a nova or supernova explosion. While much nuclear energy is expended during its lifetime, the most significant amount of energy emitted occurs at the end through gravity-induced explosion. At the galactic scale, the kinematics in formation of spiral galaxies, nebulae, clusters, etc are all driven by gravitational interaction. On the cosmological scale, it is clear that gravity is the force which is responsible for the string-like structure of the cosmos. In all these processes, the most important driving force is gravity.

In this chapter, we will examine spacetime structures of matter distribution at the galactic scale. In section 4.1, galactic structure and dynamics based on Newtonian mechanics and gravitation is introduced. Then in the proceeding section, motivation
for departure from the Newtonian gravitation is presented followed by the model galaxies based on Einstein’s theory of general relativity. Peculiar aspects of the model are discussed.

4.1 Newtonian Astrophysics

Newton’s equation of gravity can be written as

\[ \nabla^2 \Phi = 4\pi \rho. \]  \hspace{1cm} (4.1)

The left-hand side contains the gravitational field of an astronomical system and the right-hand side contains the matter density. Almost all of our solar system can be described by this equation to an accurate degree. We know what is on the left-hand side and what is on the right-hand side and both agree very well. With this equation alone, we have landed vehicles and probes on distant planets. We have extremely good observational evidence this equation works in the scale of our solar system. To observe the dynamics beyond this scale, unknown quantities slowly creep in. The right-hand side becomes less known.

4.1.1 Dark Matter

In 1933, Zwicky, a Swiss astrophysicist, proposed the idea of unknown gravitating matter[3] which does not emit or absorb any electromagnetic radiation. He was led to this conclusion after observing dynamical discrepancies in the Coma Cluster. The Coma Cluster is a large cluster 140 Mpc away consisting of over a thousand galaxies, mostly elliptical. Being of the same type meant a general profile would match most galaxies. Upon applying the virial theorem to the cluster, he discovered that there should be much optically undetected mass within the cluster.

From the galaxies orbiting on the edge of the cluster, he deduced the total mass. He counted about 1000 interior galaxies and along with the total mass, his calculations
indicated the average mass of each galaxy was at least $4.5 \times 10^{10} M_\odot$. The average luminosity per galaxy was only $8.5 \times 10^7 L_\odot$. This gave him a mass-to-light ratio of

$$\Upsilon_\odot = 530 \frac{M_\odot}{L_\odot}.$$  

That is, he found about 500 times more mass than there would appear to be. He compared it to the known mass-to-light ratio $\Upsilon_\odot = 3 M_\odot / L_\odot$ for the local “Kapteyn stellar system". To explain such a large discrepancy between the Coma Cluster and the expected mass-to-light value, he deduced that there must be an enormous amount of mass hidden in the Coma Cluster which does not radiate any electromagnetic radiation but does reveal itself through its gravitational interaction. This was the first hypothesis of dark matter.

Zwicky’s idea of dark matter was not well-received during his time and his work was published in an obscure journal called “Helvetica Physica Acta”. After 40 years, with the introduction of radio telescopes, astronomers were finally convinced to take the existence of dark matter into consideration. In the early 1970s, with technology to observe galactic structures just emerging, there was much debate\cite{21}\cite{18}\cite{20}\cite{16} about the rotation curves at large radii (beyond the visible region), particular that of M31. There were questions of whether the rotation curves were in decline at the outer radii. The absence of decline in the rotation curves implied that there was much undetected mass. It was not until the late 1970’s with further refinement to radio telescope technology that dark matter became a serious contention in astrophysics.

Refinement in radio telescope technology was focused on detecting the 21 cm line emission from neutral hydrogen. In a neutral hydrogen atom, the electron’s spin can be aligned parallel or anti-parallel to the proton’s. Because of magnetic interaction, there is a hyperfine splitting in energy level between the two spin states. With such

\footnote{In 1910, Kapteyn was the first to use photographic plates try to to determine the density of the Milky Way. What he discovered (from the northern hemisphere) was an equal number of stars towards the galactic centre and towards the anti-centre. From this symmetry, his conclusion was that the sun was near the centre of the Milky way. In fact, what he thought was the Milky Way turned out to be an area of only a few kiloparsecs across.}
a small energy difference, the characteristic decay time is in the order of 10 million years. This implies that a tiny bit of energy can be stored in neutral hydrogen atoms for a very time. Furthermore, it is a source of emission in which the relative intensity gives an indication of the amount of hydrogen present in a region. Its first detection was made in 1951 by Ewen and Purcell[6] and refinements in the technology continued through until the mid 1970s when high-resolution mapping on the rotation curves of galaxies became possible.

One of the classical works supporting the dark matter hypothesis was that of Bosma[26] who, in 1978, made observations on the rotation curves for half a dozen galaxies. While there were previous works by other astronomers in observing galactic dynamics, they were limited to the inner region of each galaxy by optical observation and to limited spatial resolution of the Hα emission. With the development of technology such as the Westerbork Synthesis Radio Telescope to detect the 21 cm line emission in the HI region in detail, Bosma made observations to determine the dynamics of spiral galaxies. His work showed the lack of Keplerian fall-off of velocity in almost all spiral galaxies. In Newtonian gravitation, there is a one-to-one relation (in natural\(^2\) units) between enclosed mass and the rotational speed:

\[
\int \frac{\rho(R)}{R} \, dv = V^2(R).
\]

One can "invert" the velocity to deduce the enclosed mass. A lack of Keplerian fall-off in velocity implies that the total mass must be increasing as a function of \(R\). Integrated to the outer visible region, this meant that the mass-to-light ratio increased to very large numbers.

After dark matter was established at the galactic scale, much effort was put into searching for dark matter. In determining the shape of the galaxies, tracers are employed. The density profile can be traced using RR Lyrae stars. Using these are advantageous due to their distinctive large-amplitude variability and their luminous nature which allows for large distance observations. For example, they can also be

\(^2\)In SI units, it would be \(G \int (\rho/R) \, dv = V^2\).
used to determine the distribution of the Milky Way's central bulge. [38] Early works by Oort and Plaut in 1975 and then later by Wesselink in 1987 showed that the bulge had a more flattened distribution than a spherical bulge. To probe at larger scales outside our own galaxy, a single star is not sufficiently energetic to be detected. The use of dwarf galaxies are used as tracers. In 1997, Zaritsky et al. used these to determine the shape of the halo[44] for a sample of galaxies. As one would expect, the number of dwarf galaxies surrounding one particular parent galaxy is very low, the only possible approach in determining the shape is to look for a number of (inclined) spiral galaxies and put together a statistical sample. They concluded that the dark matter halo extended well beyond the HI region which is typically 20-40 kpc. Their results indicated that sizes of the halos around typical galaxies are in the order of 200 kpc. The overall picture of most galaxies is that they are surrounded by an enormous halo that is ten times wider than the galaxy itself.

Evidence of dark matter beyond the galactic scale were found in other observations. Gravitational lensing and cosmic microwave background observations provides estimates of the ratio of dark matter to visible matter in the order of ten to one. This implies that dark matter plays a critical role in galactic dynamics all the way up to the cosmological scales. Such an important role propels researchers into determining the nature (and even the location) of the dark matter constituent. The search effort can be classified into two main categories for which we shall provide examples.

In the first category, one assumes that the underlying Newtonian theory of gravitation is correct and that there are some entities yet to be detected. The undetected objects may be comprised of well-known matter but in unknown states or configurations in such a way which makes it difficult to detect. On the other hand, the object may be composed of a newly undiscovered gravitating exotic matter which has eluded detectors in particle-accelerator experiments on Earth. The approaches in this category are typically experimental-based and are targeted towards refinements of detection of the unknown entity.

In the second category of approaches, one questions the underlying gravitational
theory and assumes that observational tools are correct in the quantities that are interpreted. Modification to gravitational theory is the natural path of this approach. Introduction of extended theories beyond general relativity and the standard model such as string theory would also fall into this category. These approaches are based mainly on theoretical grounds and hopes to refine the model, rather than developing better detection apparatus.

One of the most conservative approaches to solving the dark matter problem is to assume that that the dark matter halo consists of baryonic matter in a different configuration as compact objects. These are known as Massive Compact Halo Objects (MACHO). Unfortunately, they are very difficult to detect because these objects are not dispersed like gaseous hydrogen clouds and thus they cannot be detected by the 21 cm emission using radio telescope. Furthermore, these objects do not emit any radiation since there are insufficiently massive to provide enough pressure to induce nuclear fusion. However, the fact that they are very difficult to detect makes them a good candidate for the dark matter composition. The only way to detect them is through their gravitational interactions. Besides their contribution to the gravitational field of a galaxy, they also interact gravitationally through microlensing. As a MACHO passes through the line-of-sight between a constant source of light, usually a bright star, and an observer, the light intensity increases and decreases symmetrically (with respect to time). While such events occurs very infrequently, one can still provide some constraint on the distribution based on a statistics. After several years of survey, [45]Alcock et al. concluded in 1998 that MACHOs can, at most, only account for a small fraction of the mass of the dark matter halo.

Also in the category for better detection of unknown entities is the search for Weakly Interacting Massive Particles also known as WIMPs. These are hypothetical particles such as the neutralino, a super-symmetric partner to the neutrino predicted by supersymmetry. Projects such as the Cryogenic Dark Matter Search (CDMS) are underway in search of WIMPs. As of 2006, no events were found beyond the background level. [55]Upper limits for the crossss-section were deduced: $1.6 \times 10^{-43}$
cm² for Ge, and 3 × 10⁻⁴² cm² for Si for a 60GeV WIMP. Further time-integration may possibly yield some results.

While most of the approaches to the problem are based on detecting and probing the dark matter distribution through its gravitational interaction, in the second category of approaches, Milgrom[28] introduced a Modified Newtonian Model Dynamics (MOND). His proposal was a modification of the Newtonian theory of gravity itself at large scales. Instead of having the standard

\[ F = \frac{Mm}{R^2} = ma, \]

he proposed

\[ F = \frac{Mm}{R^2} = ma \mu \left( \frac{a}{a_0} \right) \]

where \( \mu \) is a newly introduced ad hoc function. Under slow acceleration, \( \mu \approx 1 \), but at higher acceleration, its value drifts away from unity, providing a constant velocity curve profile. One can think of MOND as a modification to the

\[ F = ma \]

equation or to the Poisson equation (4.1). Effectively, the gravitational potential becomes

\[ \psi(R) = -\frac{M}{R} + \sqrt{\alpha_0 M} \ln \left( \frac{R}{R_0} \right) \]

where \( \alpha_0 \) is its only measurable³ free parameter[57]. There are two problems with such a solution. MOND is not based on first principles and with its empirical approach, one can only consider it as an effective model. Furthermore, the theory violates general relativity. To solve this problem, a relativistic version of MOND was introduced by Bekenstein[54] in 2005 called tensor-vector-scalar gravity where a scalar field and a vector field are introduced to resolve the galactic rotation profile. While the theory is mathematically sound, it flies in the face of Occam’s Razor.

³The other parameter is \( R_0 \), but it is not measurable. In \( \ln(R/R_0) = \ln(R) - \ln(R_0) \), the extra term creates an unmeasurable shift in the potential energy.
4.2 Departure From Newtonian Gravitation

One of Einstein’s first steps in the introduction of general relativity was to impede its future success. To be accepted as a valid theory of gravitation, his first step was to show that Newton’s theory is a subset of general relativity. This is a very natural first step. If the non-linear Navier-Stokes equations were derived from a mathematical principle rather than on physical arguments such as conservation of energy, the first step in its introduction is a demonstration that the wave equation$^4$ results in certain conditions. For general relativity, based on linearity arguments, one can show that Newton’s equation of gravity is indeed a special case of weak gravity and slow speed. We know “slow speed” means slow in comparison to the speed of light. While we have a sense of what weak gravity is, we will re-examine what “weak gravity” really means. To begin, we must understand how we detect gravity in the first place.

We cannot “turn off” the gravitational field nor can we justly compare a system with and without curvature. Because the gravitational field affects the natural reference frame, much more care must be placed upon the reference frame used. We cannot simply leave a scale on a “fixed” surface and measure the weight or force of gravity on it because this would imply a strong underlying assumption on the coordinate system used: the observer is at spatial infinity, the surface is held static$^5$ with respect to the observer and the source of the gravitational field is static with respect to this coordinate system. In other words, the way we detect the gravitational field must be independent of whether we are in a stationary or an accelerating elevator. While the equivalence principle states that there is no difference locally between an accelerated reference frame and a local inertial reference frame in a gravitational field, the keyword here is local, being sufficiently close together. Mathematically, there is a

$^4$To continuously dwell on the wave equations severely limits its usefulness.

$^5$It is held fixed for this particular coordinate system. However, just because one coordinate system sees two objects held fixed, it may not be seen as so in other coordinates system. Take for example, two objects near or at the event horizon. An observer out at infinity will see two fixed objects whereas an observers near the horizon would see a dynamical picture.
clear distinction between a space devoid of and one with a gravitational field. The latter contains intrinsic spacetime curvature, regardless of the coordinate system used. To properly measure the gravitational field, one would have to measure (indirectly) the difference in potential at two distinct points sufficiently far apart.

When we “feel the gravitational force” on Earth, this force is not directly from the gravitational potential. The gradient of the field is, in fact, too small to detect. The force which we feel on Earth is an indirect response from the repulsive electromagnetic force of the matter under our feet. The repulsive force pushing up at our feet is from all the atoms and molecules under our feet, all the way down to the Earth’s core. Each contribution is very small, but the combined force adds up significantly. And this is what we experience when we “feel the gravitational force”. Without the help of the Earth’s atoms and molecules to detect the gravitational field, we would be floating in orbit wondering if there was a gravitational field at all. In other words, a blind astronaut would not know if he or she is in orbit around Jupiter or our moon, even though the gravitational fields are vastly different in strength. One can think of the enormous packs of atoms and molecules as the device to probe the gravitational potential difference between the surface of the earth and its core. When we feel the gravitational force, it turns out that we are measuring the field at these two locations.

The difficulty in measuring a gravitational field on the surface of Earth is that the metric is quantitatively very close to flat spacetime. One can write the gravitational potential field as a small perturbation away from a flat spacetime metric. That is,

\[
[g_{ij}] = \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & -1 & 0 & 0 \\
0 & 0 & -R^2 & 0 \\
0 & 0 & 0 & -R^2 \sin^2 \theta
\end{bmatrix} + \begin{bmatrix}
2\psi(R) & 0 & 0 & 0 \\
0 & -2\psi(R) & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{bmatrix}
\]

where \(\psi(R) = -M/R\) is very small. As a dimensionless quantity, the potential \(\psi(R)\) can naturally be viewed as a measure of departure away from flat spacetime. On the
surface of the Earth, this quantity is

$$1 - 2 \frac{M_\odot}{R_\odot} = 1 - 1.39 \times 10^{-9}$$

which is dimensionless. With such a small quantity associated with the gravitational field, it is extremely difficult to measure its strength, let alone detect it. The difference in the gravitational potential creates a small time dilation effect which can be measured in two ways which requires much ingenuity.

One approach is to directly send an atomic clock to a high altitude and compare it to a ground-based atomic clock. This was done in 1976 with the Gravity Probe A[27]. As a rocket carrying the probe reached an altitude of 10,000km, it continuously transmitted its own clock signal along with a reflected signal from Earth. The frequency shift,

$$\frac{\Delta f}{f} = (\psi_s - \psi_e) - (r_s - r_e) \cdot a_e - |v_e - v_s|^2$$

where the first term is the general relativistic term, the second is the first-order Doppler effect and the last is the second-order Doppler effect. After accounting for these variables\(^6\) as well as the slight change in the electrical path, the shift agreed to 1 part in 10\(^4\). As shown, measuring gravity is much more difficult than expected. If we can detect the earth’s gravitational field directly, we would also be able to sense the gravitational time-dilation effect directly as well.

Now, we are in the position to answer the question of what “weak gravity” means. First, it should be re-emphasized that, when we feel the force of gravity on Earth, we are not merely probing the gravitational field on the surface\(^7\), but rather, we are probing the difference in the gravitational field from the surface to the core. It is only with the great size of our Earth-probe that we can “feel” gravity. Therefore, even though weak gravity means the small order of magnitude of the metric away from

---

\(^6\)The first order Doppler effect was 3 part in 10\(^5\) and the second order Doppler effect was 6 parts in 10\(^10\).

\(^7\)Also, we are not just probing its gradient on the surface.
flat spacetime, its effect can be severely amplified through large distances. This is why astronomical observations, which targets objects of vast size, must take into effect the small spacetime curvature and regard the dynamics as general relativistic effects rather than just a classical mechanical system with Newtonian gravity as a separable ingredient.

4.3 General Relativistic Astrophysics

The problem of accounting for the observed essentially flat galactic rotation curves has been a central issue in astrophysics. There has been much speculation over the question of the nature of the dark matter that is believed to be required for the consistency of the observations with Newtonian gravitational theory. Clearly the issue is of paramount importance given that the dark matter is said to constitute the dominant constituent of a galactic mass [34]. The dark matter enigma has served as a spur for particle theorists to devise acceptable candidates for its constitution. While physicists and astrophysicists have pondered over the issue, other researchers have devised new theories of gravity to account for the observations (see for example [60]). However the latter approaches, imaginative as they may be, have met with understandable skepticism, having been devised solely for the purpose of the task at hand.

General relativity remains the preferred theory of gravity with Newtonian theory as its limit. General relativity has been successful in every test that it has encountered, going beyond Newtonian theory where required.

It is understandable that the conventional gravity approach has focused upon Newtonian theory in the study of galactic dynamics as the galactic field is weak (apart from the deep core regions where black holes are said to reside) and the motions are non-relativistic ($v \ll c$). It was this approach that led to the inconsistency between the theoretical Newtonian-based predictions and the observations of the vis-
ible sources. To reconcile the theory with the observations, researchers subsequently concluded that dark matter must be present around galaxies in vast massive halos that constitute the great bulk of the galactic masses\(^8\). However, in dismissing general relativity in favor of Newtonian gravitational theory for the study of galactic dynamics, insufficient attention has been paid to the fact that the stars that compose the galaxies are essentially in motion under gravity alone ("gravitationally bound"). It has been known since the time of Eddington that the gravitationally bound problem in general relativity is an intrinsically non-linear problem even when the conditions are such that the field is weak and the motions are non-relativistic, at least in the time-dependent case. Most significantly, we have found that under these conditions, the general relativistic analysis of the problem is also non-linear for the stationary (non-time-dependent) case at hand. Thus the intrinsically linear Newtonian-based approach used to this point has been inadequate for the description of the galactic dynamics and Einstein’s general relativity should be brought into the analysis within the framework of established gravitational theory\(^9\). This is an essential departure from conventional thinking on the subject and it leads to major consequences as we discuss in what follows. [64] We will demonstrate that via general relativity, the generating potentials producing the observed flattened galactic rotation curves are necessarily linked to the mass density distributions of the flattened disks, obviating any necessity for dark matter halos in the total galactic composition. We will also present the indicator that the threshold for luminosity occurs at a density of \(10^{-21.75}\) kg m\(^{-3}\)

\(^8\)See however [49] who argues for a much less massive halo based upon gravitational lensing data.

\(^9\)Actually within the framework of Newtonian theory, it is possible to define an "effective" potential (see for example [34] page 136) to incorporate the centrifugal acceleration in a rotating coordinate system with a given angular velocity. Since this contains the square of the angular velocity of the rotating frame, there is already the hint of non-linearity present. However, in what follows in general relativity, we will see the non-linearity related to the angular velocity as a \textit{variable} function. Moreover, for a system in rotation, this non-linearity cannot be removed globally.
4.4 A Trial Model

In an attempt to model a galaxy, the simplest model, much like the Mestel disk model, is employed. We start with the assumption that the galaxy consists of a static plane of dust which is pressureless. The assumed metric is of the form

$$ds^2 = a(z)dx^2 + b(z)dy^2 + c(z)dz^2 + e(z)dt^2.$$  

The energy-momentum tensor of a perfect fluid being

$$T^{ij} = (p + \rho)U^iU^j + pg^{ij}$$

would only have one component, namely $T^{00} = \rho_0\delta(z)$ because $p = 0$ and $U^i = \delta_0^i$. By symmetry argument, the lack of $x$ and $y$ dependence requires that $a(z) = b(z)$. The derived Einstein tensor is

$$G_{11} = G_{22} = \frac{a}{4c} \left( \frac{a'^2}{a} + \frac{a'e'}{ae} - 2\frac{a''}{a} - a'e' + \frac{e^2}{e} - 2\frac{e''}{e} + \frac{e'e'}{ce} \right) = 0$$

$$G_{33} = \frac{1}{4} \left( \frac{a'^2}{a} + 2\frac{a'e'}{ae} \right) = 0$$

$$G_{00} = \frac{ae}{4} \left( \frac{a'^2}{a^2} + 2\frac{a'e'}{ac} - 4\frac{a''}{a} \right) = \kappa T_{00}$$

where prime means derivative with respect to $z$. All other components are zero. The simplest equation to solve is $G_{33} = 0$. This leads to two solutions,

$$a(z) = \frac{\alpha_0}{e(z)^2} \quad \text{or} \quad a(z) = \alpha_1.$$  

The latter solution collapses to the trivial solution, namely $G_{00} \equiv 0$. We pick the former solution. Inserting this into the other three equations results in

$$G_{11} = G_{22} = \frac{\alpha_0}{ce^2} \left( 2\frac{e''}{e} - \frac{c'e'}{ce} - 5\frac{e'^2}{e^2} \right)$$

$$G_{00} = \frac{e}{c} \left( 2\frac{e''}{e} - \frac{c'e'}{ce} - 5\frac{e'^2}{e^2} \right).$$
This means if $G_{11} = G_{22} = 0$ then $G_{00}$ must also be zero because of the similarity in the terms in bracket. General relativity is telling us that with the assumed symmetry, the system cannot be pressure-less and static. This is very much expected because a gravitating dust cloud could not possibly remain static without some mechanism to keep it from collapsing. We must introduce pressure to balance the gravitational force. In Newtonian gravitation, one possible mechanism for maintaining the configuration would be the fact that, (assuming an infinite symmetric plane) each dust particle must feel forces equally from every direction within the plane and the net force would be zero, keeping it in a static equilibrium. Furthermore, in Newtonian gravitation, pressure is not important in determining the gravitational field. But in general relativity, pressure would play an important role as we have seen here.

Introducing pressure for a model of stars and interstellar dust is unrealistic. The material is not in contact and there cannot be any pressure. The only time when dust models exhibit pressure is during collapse. For example, in spherically symmetric collapse, the dust particles will converge into a smaller and smaller volume where at a certain instance, the dust particles will be in contact and thus a fluid model must be used[43]. The other option$^{10}$ is to introduce centrifugal force through rotation. We must require the system to be stationary and not just static.

$^{10}$We can also introduce time-dependence but that adds enormous amount of unnecessary complexity to the problem.

4.5 The Model Galaxy

We model a galaxy in the context of general relativity with the assumption that the system is axially symmetric and the galaxy is a uniformly rotating fluid without pressure. To introduce motion, we must also assume the whole system is in a stationary state as opposed to a static state. The metric form which contains these features can
be written as

\[ ds^2 = -e^{\nu} (udz^2 + dr^2) - r^2 e^{-\nu} d\phi^2 + e^{\nu} (cdt - N d\phi)^2 \]  

(4.2)

where \( u, \nu, w \) and \( N \) are functions of cylindrical polar coordinates \( r, z \). From the metric, we generate the Christoffel symbols, the Ricci tensor and the Riemann curvature tensor components shown in the Appendix. The Einstein tensor with some of the components are shown:

\[
\begin{align*}
G^{11} &= \frac{e^{4w-2\nu}}{4r^2} \left( N_r^2 - \frac{N_z^2}{u} \right) + \frac{e^{2w-2\nu}}{4} \left( w_r^2 - \frac{w_z^2}{u} \right) - \frac{e^{2w-2\nu}}{2r} \left( \nu_r + \frac{u_r}{u} \right) \\
G^{13} &= \frac{e^{2w-2\nu}}{2u} \left( w_r w_z - \frac{1}{r} \nu_z \right) - \frac{e^{4w-2\nu}}{2ur^2} N_r N_z \\
G^{22} &= -\frac{e^{4w-\nu}}{4r^4} \left( \frac{N_r^2 + N_z^2}{u} \right) + \frac{e^{2w-\nu}}{4ur^2} \left( \frac{u_z \nu_z}{u} - u_r \nu_r \right) \\
&\quad + \frac{e^{2w-\nu}}{2r^2} \left( \frac{u_r^2}{2u^2} - \frac{u_{rr}}{u} - \left( \frac{w_r^2 + \frac{w_z^2}{u}}{u} \right) - \left( \nu_{rr} + \frac{\nu_z}{u} \right) \right) \\
G^{33} &= \frac{e^{4w-2\nu}}{4r^2} \left( N_r^2 - \frac{N_z^2}{u^2} \right) + \frac{e^{2w-2\nu}}{4u} \left( \frac{w_z^2}{u} - w_r^2 + \frac{2}{r} \nu_r \right)
\end{align*}
\]  

(4.3)

where subscripts denote partial derivatives. The \( G^{02} \) and \( G^{00} \) are unnecessarily complicated and are omitted for the order-of-magnitude\(^{11}\) argument. Furthermore, it is unnecessary because, if conflict arises with the other components in the order-of-magnitude argument, this would mean the theory of general relativity is inconsistent.

From the above equations, we can determine the order of magnitude for each function. For this magnitude argument, we insert the constant \( G \) into the Einstein equation

\[ G^{ij} = 8\pi G T^{ij} \]

and compare terms in magnitudes of \( \sqrt{G} \). It is easy to show that to the order required, \( u \) can be taken to be unity. The position of \( u \) typically appears in the denominator

\(^{11}\)This is a loose notation favored by many relativists but adequate for our purposes here as a smallness parameter.
in terms such as
\[ \left( N_r^2 - \frac{N_z^2}{u} \right). \]

If we let \( u \) be of any order of \( G \) besides zero, it would push parts of the equations into a different order. In other words, \( N_r \) would be of different order than \( N_z \). Unless we are modelling an extremely flat disc or an extremely elongated cylinder, it is sensible to let \( N_r \) be in the same order as \( N_z \). By this argument, the only conclusion about \( u \) is that it is of order unity and \( N_r \) is the same order as \( N_z \). Because of scaling, we shall take \( u \) to be unity unless we require to change the model somewhat\(^{12}\).

The metric (4.2) is reduced to
\[ ds^2 = -e^{\nu-w}(dz^2 + dr^2) - r^2e^{-w}d\phi^2 + e^w(cdt - N d\phi)^2. \]  

(4.4)

It is most simple to work in the frame that is co-moving with the matter,
\[ U^i = \delta^i_0 \]  

(4.5)

where \( U^i \) is the 4-velocity. This is reminiscent of the standard approach in the Friedman-Robertson-Walker (FRW) cosmology. The geodesic equations for \( U^i \) is
\[ \frac{dU^i}{ds} + \Gamma^i_{jk}U^jU^k = 0. \]  

(4.6)

In particular, for \( U^i = \delta^i_0 \), we find the conditions for \( w(r, z) \) are
\[ \frac{dU^1}{ds} + \frac{1}{2}w_r e^{2w-v} = 0 \]
\[ \frac{dU^3}{ds} + \frac{1}{2}w_z e^{2w-v} = 0 \]

while the other two geodesic equations are trivially satisfied. That is,
\[ \frac{\partial w(r, z)}{\partial r} = 0 \]
\[ \frac{\partial w(r, z)}{\partial z} = 0. \]  

(4.7)

\(^{12}\)This leaves some room for future improvement in the model.
<table>
<thead>
<tr>
<th>Function</th>
<th>Magnitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w(r, z)$</td>
<td>0</td>
</tr>
<tr>
<td>$u(r, z)$</td>
<td>1</td>
</tr>
<tr>
<td>$N(r, z)$</td>
<td>$\mathcal{O}(G^{1/2})$</td>
</tr>
<tr>
<td>$\nu(r, z)$</td>
<td>$\mathcal{O}(G)$</td>
</tr>
</tbody>
</table>

Table 4.1: Order of magnitude for each function.

Thus, for a co-moving reference frame, we can set

$$w(r, z) = 0$$

without loss of generality because any constant solution $w(r, z) = C$ can be absorbed into the scaling of the coordinates. This usage of co-moving frame was done in the pioneering paper by van Stockum [4] who set $w = 0$ from the outset. Interestingly, the geodesic equations imply that $w = \text{constant}$ even for the exact Einstein field equations as studied in [4]. Normally, $w$ is used to determine the total mass of an isolated system by examining its fall-off characteristics with respect to the spherical radius, $R \equiv \sqrt{r^2 + z^2}$. However, $w$ is constant in this system of coordinates by (4.7) and we cannot use it to determine the total mass of a system. The $w$ constancy does not imply that that the mass is zero. In other non-co-moving coordinate systems, $w$ would be seen to be variable. With the field being weak and the system being non-relativistic, the mass is well-approximated simply by the integral of $\rho$ over coordinate volume. Moreover, we will choose solutions that are free of singularities and hence free of the ambiguities present in [24].

To further determine the orders of magnitude, we continue to examine (4.3). We can see that $N_r^2$ and $N_z^2$ are mixed with $\nu_r$ and $\nu_z$. With

$$G^{ij} = \frac{8\pi G}{c^2} T^{ij},$$

...
this leads to the requirement that $N$ is of order $\sqrt{G}$ and $\nu$ is of order $G$. A summary of this is shown in table 4.1.

4.5.1 The Approximate Einstein Equations

After the reductions, the Einstein equations become

$$2r\nu_r + N_r^2 - N_z^2 = 0,$$

$$r\nu_z + N_r N_z = 0,$$

$$N_r^2 + N_z^2 + 2r^2(\nu_{rr} + \nu_{zz}) = 0,$$  \hspace{1cm} (4.8)

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

$$\frac{3}{4r^2} (N_r^2 + N_z^2) + \frac{N}{r^2} \left( N_{rr} + N_{zz} - \frac{N_r}{r} \right) - \frac{1}{2} (\nu_{rr} + \nu_{zz}) = 8\pi \rho$$

which can be further reduced to

$$\nu_r = \frac{N_z^2 - N_r^2}{2r},$$  \hspace{1cm} (4.9)

$$\nu_z = -\frac{N_r N_z}{r},$$  \hspace{1cm} (4.10)

$$N_{rr} + N_{zz} - \frac{N_r}{r} = 0,$$  \hspace{1cm} (4.11)

$$\frac{N_r^2 + N_z^2}{r^2} = 8\pi \rho$$  \hspace{1cm} (4.12)

where $\rho$ is the mass density. It is to be noted that it is the freely gravitating motion of the source material (the stars) in conjunction with the choice of co-moving coordinates (4.5) that leads to the constancy of $w$ within the source. Had there been pressure appearing on the right-hand side of (4.8), $w$ would have been variable\(^{13}\). The

\(^{13}\)Even in freely gravitating motion, $w$ would have been variable had we opted for non-co-moving coordinates.
non-linearity of the galactic dynamical problem is manifest through the non-linear relation\(^{14}\) between the functions \(\rho\) and \(N\). Rotation under freely gravitating motion is the key here. By contrast, for time-independence in the non-rotating problem, there must be pressure present to maintain a static configuration, \(N\) vanishes for vanishing \(\omega\) and \(\nabla^2 w\) is non-zero yielding the familiar Poisson equation of Newtonian gravity.

While the first two equations (4.9) and (4.10) can be used to determine \(\nu\), the equations of interest are (4.11) and (4.12). It is easy to show that if we introduce a new potential,

\[
\Phi \equiv \int \frac{N}{r} \, dr
\]

and apply it to (4.11), the governing equation for this new function is

\[
\nabla^2 \Phi = 0 \tag{4.13}
\]

where \(\nabla^2\) is the laplacian flat-space operator using cylindrical coordinates. The profound statement associated with (4.13) is that harmonic functions are the generators of the axially symmetric stationary pressure-free weak fields. In fact Winicour [23] has shown that all such sources, even when the fields are strong, are generated by such flat-space harmonic functions. This is in sharp contrast to the Newtonian gravity where the governing equation is the Poisson equation.

### 4.5.2 Dust Particle Velocities

To determine the velocity of a dust particle, we follow its stationary motion around the axis of symmetry. In the same manner as in [24], we hold \(r\) and \(z\) fixed and apply the local transformation,

\[
\bar{\phi} = \phi + \omega(r, z) \, t.
\]

\(^{14}\)While we have eliminated \(w\) using (4.6) to get (4.12) by choice of co-moving coordinates, \(N\) cannot be eliminated and hence non-linearity is intrinsic to the study of the galactic dynamics.
Effectively, this locally diagonalizes the metric. The deduced local angular velocity $\omega$ and the tangential velocity $V$ is

$$\omega = \frac{N}{r^2 - N^2}$$  \hspace{1cm} (4.14)

$$\approx \frac{N}{r^2}$$  \hspace{1cm} (4.15)

$$V = \omega r.$$  \hspace{1cm} (4.16)

with the approximate value applicable for the weak fields under consideration. In the present case, it is the rotation

$$V = c \frac{N}{r}$$  \hspace{1cm} (4.17)

$$= c \frac{\partial \Phi}{\partial r}$$  \hspace{1cm} (4.18)

via the function $N$. This $N \approx \omega r^2$ can be interpreted as the angular momentum per unit mass. Furthermore, the Einstein equations connects the $N$ function directly to the density. The resulting non-linear equation (4.12) is in sharp contrast to the linear Poisson equation.

### 4.6 Modeling Observed Galactic Rotation Curves

Since the field equation (4.12) for $\rho$ is non-linear, the simplest way to proceed in galactic modeling is to first find the required generating potential $\Phi$ and from this, derive an appropriate function $N$ for the galaxy that is being analyzed. With $N$ found, (4.12) yields the density distribution. If this is in accord with observations, the efficacy of the approach is established. Every galaxy is different and each requires its own composing elements to build the generating potential. In cylindrical polar coordinates, separation of variables by assuming

$$\Phi(r, z) = R(r)Z(z)$$
yields

\[ \frac{1}{R} \frac{d^2 R}{d^2 r} + \frac{1}{r R} \frac{dR}{dr} = -\frac{1}{Z} \frac{d^2 Z}{d^2 z}. \]

With the left-hand side being a function of \( r \) and the right-hand side being a function of \( z \), they can be equated only if both sides are identical to a constant. We write

\[ \frac{1}{R} \frac{d^2 R}{d^2 r} + \frac{1}{r R} \frac{dR}{dr} = -k^2 \]

\[ -\frac{1}{Z} \frac{d^2 Z}{d^2 z} = -k^2 \]

which have well-known solutions. Together, they yield the following solution for \( \Phi(r, z) = R(r)Z(z) \):

\[ \Phi(r, z) = Ce^{-k|z|}J_0(kr) \]

where \( J_0 \) is the Bessel function \( n = 0 \) of Bessel \( J_n(kr) \) and \( C \) is an arbitrary constant. The absolute value of \( z \) has been inserted for reflection symmetry and will be discussed in the next section.

We use the linearity of (4.13) to express the general solution of this form as a linear superposition with different \( k_n \) values

\[ \Phi = \sum_n C_n e^{-k_n|z|}J_0(k_nr) \quad (4.19) \]

with \( n \) chosen appropriately for the desired level of accuracy. From (4.19) and (4.16), the tangential velocity\(^{15} \) is

\[ V = -\sum_n k_n C_n e^{-k_n|z|}J_1(k_nr) \quad (4.20) \]

There is still the freedom of \( k_n \) values to choose. Just as the sin \( kx \) functions are orthogonal for integer \( k \), the Bessel functions \( J_0(kr) \) have their own orthogonality relation:

\[ \int_0^1 J_0(k_nr)J_0(k_mr)rdr \propto \delta_{mn} \]

\[ ^{15}dJ_0(x)/dx = -J_1(x) \text{ from [7].} \]
where \( k_n \) are the zeros of \( J_0 \) at the limits of integration. This orthogonality condition is on \( \Phi \) rather than on \( V \) because the differential equation dictates the integral condition. Therefore \( k_n \)'s are chosen so that the \( J_0(k_n,r) \) terms are orthogonal to each other. We have found that only 10 functions with parameters \( C_n, n \in \{1 \ldots 10\} \) suffice to provide an excellent fit to the velocity curve for all the galactic models used.

It should be noted that unlike typical velocity curve fits that allow arbitrary velocity functions, our curve fits are required to be created from derivatives of harmonic functions. In most Newtonian models of galaxies, arbitrary distribution can be inserted into the model without any justification except for the fact that mathematically, they fit well. In Kent's model[35], the galaxy is separated into the bulge component, the disk component and the halo having distribution

\[
\rho = \frac{\sigma^2}{2\pi(r^2 + a^2)}
\]

with the four fitting parameters being the characteristic density of the halo \( \sigma^2 \), the halo length-scale \( a \) and the mass-to-light ratios for the bulge and the disk. There is much arbitrariness in the way each parameter is introduced. On the other hand, in our relativistic model, the parameters must all be introduced in the same manner, being coefficients of each of the Bessel terms. This aspect tremendously reduces the freedom in which general relativity can be used to construct a model galaxy.

### 4.7 Boundary Layer Conditions

For the classical Newtonian problem in cylindrical coordinates with axial-symmetry, the equation for the gravitational potential is

\[
\frac{\partial^2 \psi}{\partial z^2} + \frac{\partial^2 \psi}{\partial r^2} + \frac{1}{r} \frac{\partial \psi}{\partial r} = \rho(r, z). \tag{4.21}
\]

The mathematical solution to this Poisson equation is

\[
\psi(r, z) = \int \frac{\rho(r, z)}{\sqrt{r^2 + z^2}} dr dz. \tag{4.22}
\]
Figure 4.1: Solution for the Newtonian potential for a point mass exhibits a singularity.

For a point mass, the "measured" potential is

$$\psi(r, z) = \frac{\psi_0}{\sqrt{r^2 + z^2}}$$

(4.23)

as shown in figure 4.1. Here is where we arrive at the junction between mathematics and physics. Mathematics provides us with the solution (4.22) and physics provides us with the empirically measured result (4.23). In order for these two to match, the charge distribution must exhibit the singular property

$$\rho(r, z) \propto \delta(r)\delta(z).$$

There is nothing intuitive\(^{16}\) about the right-hand side of the above except for the mathematical statement about the nature of the singularity. i.e., the integral of the right-hand side is one. We cannot have deduced this by blindly relying on the

\(^{16}\)The Dirac delta is much like the concept of infinity. Infinity is not an actual number; it can only be viewed as a limiting process. We certainly cannot write $3\delta(x) > 2\delta(x)$, but we can write $\lim_{n \to \infty} (3\delta_n(x) - 2\delta_n(x)) > 0$. 
mathematics and simply inserting (4.23) into (4.21) because if we did, we would end up with

\[
\frac{3z^2}{(r^2 + z^2)^{5/2}} - \frac{1}{(r^2 + z^2)^{3/2}} + \frac{3r^2}{(r^2 + z^2)^{5/2}} - \frac{1}{(r^2 + z^2)^{3/2}} + \frac{-1}{(r^2 + z^2)^{3/2}}
\]

which simplifies to zero! This would tell us absolutely nothing about the nature of the singularity. The key point here is the physical interpretation of the singularity.

In our problem, we have the non-linear equation

\[
\frac{1}{r^2} \left( \frac{\partial N}{\partial r} \right)^2 + \frac{1}{r^2} \left( \frac{\partial N}{\partial z} \right)^2 = \rho(r, z)
\]

(4.25)

where the "measured potential" is

\[
N(r, z) = -\sum_{n=1}^{10} C_n k_n r e^{-k_n z} J_1(k_n r).
\]

However, unlike the linear case of the poisson equation, there is no exact mathematical solution for us to determine the nature of the singularity (if one exists). Again, just as with our previous example in (4.24), we simply cannot insert this potential into the equation (4.25) to determine the nature of the singularity. Simply following the mathematics blindly has led some to [61] deduce that the boundary layer contains negative mass which seems rather contradictory because the left-hand side of (4.25) contains sums of squares.

There are three possible approaches to resolving the ambiguity about the nature of the singularity[65]. The first involves avoiding any gradient-discontinuity of \(\rho\) by cutting a slice between, say \(z_0 = \pm 1\) kpc and solving the partial differential equation numerically with the boundary conditions\(^{17}\)

\[
\frac{\partial N_{(int)}(r, z)}{\partial z} \bigg|_{z=z_0} = \frac{\partial N_{(ext)}(r, z)}{\partial z} \bigg|_{z=z_0}
\]

\(^{17}\)The symmetric boundary condition will naturally provide \(N_{(int)}\) with mirror-symmetry without any gradient-discontinuity.
and

$$N_{(int)}(r, z_0) = N_{(ext)}(r, z_0).$$  \hspace{4cm} (4.26)

It should be noted that (4.26) implies that the derivative

$$\frac{\partial N_{(int)}}{\partial r} = \frac{\partial N_{(ext)}}{\partial r}$$

matches at the boundary as well. This numerical scheme will be implemented in the next chapter.

The second approach to the problem is similar in that we avoid the gradient-discontinuity of $\rho$ by using $\cosh(kz)$ instead of $e^{-k|z|}$. However, $\cosh(kz)$ will grow for large values of $z$; thus, to accomplish both tasks of maintaining asymptotic behaviour without a gradient-discontinuity, we keep the 10 parameters with the associated $k_n$ for the internal strip and introduce (depending on how many is necessary) hundreds of new parameters with associated $\kappa_m$ values externally. Basically, we match a series of $\cosh(k_n z)$ with a series of $e^{-\kappa_m |z|}$ at $z = \pm z_0$.

The final approach is to determine if the gradient discontinuity is creating some unusual feature within the model. That is, we would like to confirm whether or not the claim that the $z = 0$ layer does contain a negative mass layer. This is done through calculation of the geodesic equation of a test particle

$$\frac{dU^z}{ds} - \frac{\nu_z}{2} (U^r)^2 + \nu_r U^r U^z + NN_z(U^\phi)^2 e^\nu - N_z U^t U^\phi e^\nu + \frac{\nu_z}{2} (U^z)^2 = 0$$

where $U^r$ and $U^\phi$ are zero, leaving only

$$\frac{dU^z}{ds} = -\frac{\nu_z}{2} (U^z)^2.$$ 

We can determine whether the right-hand side is a positive or negative value using (4.10) and numerical values for the function $N$. All three approaches are done numerically in the next chapter.
Chapter 5

Numerics

To model the galaxy, we require two different tools. The first is the least-square fit which will be used extensively and the second is the numerical solution to partial differential equations.

5.1 Least Squares

Least-square fittings are classified in one of the two categories: Linear and Non-linear. The linearity is associated with how the parameters appear in the curve-fitting equation. In a linear fit, the parameters appear in the curve-fitting function in the separable form of

$$F(x) = \sum_{\alpha=1}^{n} p_{\alpha} f_{\alpha}(x)$$  \hspace{1cm} (5.1)

where

$$\{f_{\alpha}(x) | \alpha = 1, 2, \ldots, \eta\}$$
is the given set of basis functions and \( \{p_\alpha\} \) is the set of parameters to be found. The general linear least square fit is represented by

\[
\min_{\{p_\alpha\}} \sum_{i=1}^{N} \left( y_i - \sum_{\alpha=1}^{\eta} p_\alpha f_\alpha(x_i) \right)^2
\]

where the sets \( \{x_i|i = 1 \ldots N\} \) and \( \{y_i|i = 1 \ldots N\} \) are given data points. Here with the linear case, we can see how the parameters appear in (5.1). In contrast, a non-linear least square fit is represented by

\[
\min_{\{p_i\}} \sum_{i=1}^{N} (y_i - f(p_1, \ldots p_\eta, x_i))^2
\]

where the parameters "\( p_\alpha \)" are embedded into the function itself. However, both cases share the same approach towards accomplishing a least-square fit; that is, the derivatives with respect to the parameters are taken and the resulting algebraic expressions are equated to zero. The resulting algebraic equations may be linear or non-linear depending on how the parameters appear in the curve-fitting function. In the latter case, a solution may not exist whereas in the linear case, a unique solution is guaranteed to exist. This is the most distinct feature which will make the approach much easier to handle.

### 5.1.1 Linear Square Fit

The standard approach to minimizing a linear fit is to take the partial derivative of

\[
L(p_1, \ldots, p_\eta) = \sum_{i=1}^{N} \left( y_i - \sum_{\alpha=1}^{\eta} p_\alpha f_\alpha(x_i) \right)^2
\]

with respect to each \( p_\beta \), resulting in

\[
\frac{\partial L}{\partial p_\beta} = -\sum_{i=1}^{N} 2 \left( y_i - \sum_{\alpha=1}^{\eta} p_\alpha f_\alpha(x_i) \right) f_\beta(x_i)
\]

(5.2)
and then equate each to zero to find the critical point. It is easily seen that the second derivative
\[
\frac{\partial^2 L}{\partial p_\beta^2} = \sum_{i=1}^{N} 2f_\beta(x_i)f_\beta(x_i)
\]
(5.3)
is positive-definite\(^1\) which assures us that any critical point cannot be a maximum. This property makes linear curve-fits tremendously more straight-forward than non-linear fits. Equating the first derivative to zero results in
\[
\sum_{i=1}^{N} y_i f_\beta(x_i) - \sum_{i=1}^{N} \sum_{\alpha=1}^{\eta} f_\alpha(x_i)f_\beta(x_i)p_\alpha = 0.
\]
which can be written in a matrix form as follow:
\[
\begin{bmatrix}
M_{11} & \ldots & M_{\eta 1} \\
\vdots & \ddots & \vdots \\
M_{1\eta} & \ldots & M_{\eta \eta}
\end{bmatrix}
\begin{bmatrix}
p_1 \\
p_2 \\
\vdots \\
p_\eta
\end{bmatrix}
=
\begin{bmatrix}
Y_1 \\
\vdots \\
Y_\eta
\end{bmatrix}
\]
(5.4)
where
\[
Y_\beta = \sum_{i=1}^{N} y_i f_\beta(x_i)
\]
\[
M_{\alpha \beta} = \sum_{i=1}^{N} f_\alpha(x_i)f_\beta(x_i).
\]
The linearity in the problem is exhibited in (5.4) where the parameters \(p_1, \ldots p_\eta\) appear as linear terms. The simplicity of (5.4) means the standard approach to minimizing a linear square only involves inverting the square matrix.

In preparation to extend to a non-linear algorithm, we consider an simple one-dimensional or one-paramater system
\[
F(x) = p_1 x
\]
\(^1\)We can safely assume that \(f(x_n)\) is non-zero for most \(x_n\)'s. If it was zero for all \(x_n\)'s such as \(\sin(x_n)\) for \(x_n = n\pi\), then there would be an infinite number of possible fits.
There are only two quantities to compute: \( Y_1 = \sum_{i=1}^{N} y_i x_i \) and \( M_{11} = \sum_{i=1}^{N} x_i^2 \). We must solve
\[
Y_1 - M_{11} p_1 = 0. \tag{5.5}
\]
Define the left-hand side as a new function \( G(p_1) = Y_1 - M_{11} p_1 \) so that the basic premise is to solve \( G(p_1) = 0 \). This may seem trivial, but in the next sections, more structures will be built upon this starting point.

### 5.1.2 Non-Linear Least-Square Fit

In a non-linear least-square fit, the function to be minimized is
\[
L(p_1, \ldots, p_\eta) = \sum_{i=1}^{N} (y_i - f(p_1, \ldots, p_\eta, x_i))^2.
\]
To do so, we solve
\[
\frac{\partial L}{\partial p_\beta} = -\sum_{i=1}^{N} 2(y_i - f(p_1, \ldots, p_\eta, x_i)) \frac{\partial}{\partial p_\beta} f(p_1, \ldots, p_\eta, x_i) = 0.
\]
While this equation may look vastly different from (5.2), the basic idea is identical. We must solve for the roots of a system of equations.

Continuing with our simple one-parameter example from the previous section, we change our fitting function to a more complicated form:
\[
F(x) = J_0(p_1 x)
\]
where \( J_0 \) is the Bessel function of order zero. The function to be minimized is
\[
L(p_1) = \sum_{i=1}^{N} (y_i - J_0(p_1 x_i))^2
\]
by taking its derivative and solving
\[
-2 \sum_{i=1}^{N} (y_i - J_0(p_1 + x_i)) x_i J_0'(p_1 x_i) = 0 \tag{5.6}
\]
Figure 5.1: Newton's method is based on constructing a linear model around the current point to determine the next iterative point.

for $p_i$. This is the non-linear equivalent of (5.5) where

$$G(p) = -2 \sum_{i=1}^{N} (y_i - J_0(p_i + x_i)) x_i J_0'(p_i x_i).$$

Solving such a problem numerically requires the use of a root-finding algorithm. Newton's method will be employed.

5.1.3 Newton's Method

Newton's method is an algorithm for finding the roots of

$$G(p) = 0$$

for an arbitrary function $G(p)$. It is an iterative method based on constructing a linear\footnote{Other methods such as the Brent method goes beyond the linear model and calls for a construction of a second-order polynomial about the current iterative point.} model about the point $(p_{(n)}, G(p_{(n)}))$ and using this to take the next best
Figure 5.2: When the derivative is close to zero, the next iterative point can be grossly over-estimated leading to poor numerical conditions.

estimate of its root. Given the slope of the tangent line at \((p_n, G(p_n))\), the linear model is constructed as

\[
\frac{0 - G(p_n)}{p_{n+1} - p_n} = G'(p_{n+1})
\]

where the next point \(p_{n+1}\) is expected to converge closer to the solution as illustrated in figure 5.1. Solving for \(p_{n+1}\) in (5.7) results in

\[
p_{n+1} = p_n - \frac{G(p_n)}{G'(p_n)},
\]

In the best-case scenario, \(G(p)\) is linear in \(p\) and the root of \(G(p) = 0\) can be found within one Newton-step. In the worse case scenario the approach can fail if \(G'(p_n)\) is zero or close to zero, causing numerically unstable conditions. Such a condition is exemplified in figure 5.2. There are many modifications to the algorithm which can overcome such difficulties. One such approach is known as back-tracking where upon encountering such a scenario, \(G(p_{n+1}) \geq G(p_n)\), the next step is reduced by half as

\[
p_{n+1} = \frac{1}{2} \left( p_{n+1}^{\text{test}} + p_n \right)
\]

until the condition

\[
G(p_{n+1}) < G(p_n),
\]
is satisfied. This can ensure that each successive step will indeed approach the root instead of diverging away from it. However, with such an ad-hoc manner, it tremendously reduces the convergence rate of the algorithm and introduces more costs in function-evaluations.

In our example from (5.6), we have a function

$$G(p) = -2 \sum_{i=1}^{N} (y_i - J_0(p x_i)) x_i J_0'(p x_i).$$

for which we need to find the roots. We apply the Newton method, starting from the point \( p_{(0)} \) and use

$$G'(p_{(0)}) = -2 \sum_{i=1}^{N} (y_i - J_0(p_{(0)} x_i)) x_i^2 J_0''(p_{(0)} x_i) + 2 \sum_{i=1}^{N} x_i^2 J_0'(p_{(0)} x_i)^2,$$

which results in the next iterative point at

$$p_{(1)} = p_{(0)} - \frac{G(p_{(0)})}{G'(p_{(0)})}.$$

We can write this as

$$\{p_{(0)} G'(p_{(0)}) - G(p_{(0)})\} - \{G'(p_{(0)})\} p_{(1)} = 0$$

which is the same as (5.5) if we let \( Y_1 = p_{(0)} G'(p_{(0)}) - G(p_{(0)}) \) and \( M_{11} = G'(p_{(0)}) \).

Looking at the overall scheme, one can see that even with the vast change from a linear to a non-linear problem and the introduction of a different algorithm, at the heart of the problem still lies the simple equation, \( Y_1 - M_{11} p_1 = 0 \).

### 5.1.4 Minimization

The goal is more than just to find the roots of a system of equations, but rather to find the minimizer of the original function, \( L(p_1, \ldots, p_\eta) \). In the linear problem, the

---

3The secondary goal in most numerical methods is to reduce the number of function-evaluations. They tend to be expensive, computationally or otherwise.
second derivative (5.3) is positive-definite, but this is not guaranteed for the nonlinear problem. To contend with this lack of guarantee for an existence of a minimum requires more ingenuity, sometimes using ad-hoc methods. The basic problem stems from the distinction between the task of simply finding the roots of an equation and finding the minimum of a function. The latter has more structure and requirements.

Minimization of a function, $L(p)$, requires the use of two types of algorithms. The first is the global algorithm which is a course-grain sweep of the region of interest. It depends on the type of problem and the numerical characteristics at hand. There is never any set rules (or convergence-analysis) for any group of global algorithms. Usually, there is much human intervention in this step and one must get a feel for the problem at the global level before coming up with a simple algorithm. For example, if we wish to find a minimum of an ill-behaved function $G(p) = \Gamma(p)$, the Gamma function, we can initially pick 10 different places at equal intervals between -3 and 3. Upon closer inspection, we will find that this function diverges at negative integers as
shown in figure 5.3. We would limit our search to $0.1 \leq p \leq 3$ by picking 10 different locations in that smaller region and using the minimizer as the starting point to the next step: the local algorithm.

The local algorithms are typically variants of the Newtonian method where $G(p) = L'(p)$ and we iterate through the steps using (5.8) as its core-algorithm. Based on this setup, successive $p(n)$'s will drift towards points in $L(p)$ with the lowest derivatives (i.e., $L'(p) \approx 0$). However, there are problems associated with this "passive drifting". The two most obvious problems in using (5.8) are:

1. Stepping towards a point with lower derivative (in magnitude) of $L(p)$ does not necessarily mean it is approaching a minimum. The lack of distinction between a minimum and a maximum can lead the algorithm to step towards a maximum instead of a minimum. But one cannot simply just step in the opposite direction. Consider the point at $p = -2.5$ in figure 5.3. If we continue to step away from the local maximum, we will end up in an infinite loop, falling towards $-\infty$. We can determine if we are stepping in the correct direction by looking at the sign of the second derivative, $G''(p(n))$. If it is positive, it indicates that, the local model is a minimum. Otherwise, if it is a local maximum, we must re-evaluate the problem with the global algorithm possibly using a smaller region than the original global model.

2. The denominator in (5.8) may be zero or close to zero. This is much like the failure of the Newton algorithm shown in figure 5.2. One can avert this numerical instability by using back-tracking algorithm\(^4\) or stepping up to the global picture and using a global algorithm.
Figure 5.4: In the multi-dimensional Newton's method, one constructs a plane or a hyperplane as a local model to determine the next iterative step.

5.1.5 Multi-dimensional Minimization

Non-linear least-square curve-fits of $\eta$ number of parameters require a generalization from one to an $\eta$ dimensional system for Newton's method as shown in figure 5.4. Most of the generalization can be as simple as re-writing $p_{(n)}$ in place of $p_{(n)}$. However, other terms require closer examination. We cannot simply write (5.8), as

$$p_{(n+1)} = p_{(n)} - \frac{G(p_{(n)})}{G'(p_{(n)})}$$

because of the mixing of vectors and scalars. The appropriate notation is written with $L(p)$ as

$$p_{(n+1)} = p_{(n)} - \left[H(p_{(n)})\right]^{-1} \left[\nabla L(p_{(n)})\right]$$

where $[H(p_{(n)})]$ is the Hessian matrix of the function $L(p_{(n)})$ and the derivative combinations are taken with respect to $p_{(n)}$.

\footnote{One can think of the back-tracking method as a temporary global scheme which, like many other global algorithms, suffer from a poor rate of convergence.}
We can see that \([H(p_{\{n\}})]^{-1}\) may be numerically unstable after inversion. It means the determinant is almost zero, because it is close to a non-full-rank matrix. To handle such a situation requires computation of the eigenvectors and the associated eigenvalues. The lowest eigenvalues (close to zero) will determine the degenerate direction(s) in \(p\) space. In other words, the eigenvectors will show the direction(s) of the trough(s). Again, back-tracking can be employed to prevent these numerical instabilities.

### 5.2 Numerical Solutions to PDE

Just as second-order partial differential equations (PDE), in the form of

\[
\frac{\partial^2 f}{\partial x^2} + b \frac{\partial^2 f}{\partial x \partial y} + c \frac{\partial^2 f}{\partial y^2} + d \frac{\partial f}{\partial x} + e \frac{\partial f}{\partial y} = h,
\]

are classified into three different classes, numerical algorithms for solving PDE are also classified into the same distinct categories. The three classes\(^5\) of PDE are

\[
b^2 - 4ac > 0 \quad \Rightarrow \quad \text{hyperbolic}
\]

\[
b^2 - 4ac = 0 \quad \Rightarrow \quad \text{parabolic}
\]

\[
b^2 - 4ac < 0 \quad \Rightarrow \quad \text{elliptic}.
\]

An example of each is the wave equation, the heat equation and the Helmholtz equation, respectively. How the PDE is solved is dictated by which of these three classes it falls into.

What makes the distinction between the solutions in each class of PDEs is domain of dependence (DoD). The solution to the wave equation at a point \((x_0, t_0)\) has its

---

\(^5\)Generally the discriminant, \(a^2 - 4ac\), is not a constant so a PDE may have a region of one type and other areas of other types. Boundary conditions may become subtle when mixing domains in such a manner. Fortunately, most PDE's of interest do not vary in their characteristics across the whole domain.
DoD as \((x_0 \pm c\tau, t_0 \pm \tau)\) where \(c\) is the speed of propagation in the equations and \(\tau > 0\). For a (null-vector) wave in a Minkowski spacetime, the DoD is the past light cone\(^6\). For the heat equation, the DoD at \((x_0, t_0)\) is all the points before \(t_0\); that is, the region for which \(-\infty < x < \infty\) and \(t < t_0\) contributes to the solution at \((x_0, t_0)\). And finally, the DoD for the Helmholtz equation is the whole domain. Changes to one point in the solution will affect the whole solution. This is the extreme opposite from that which prevails in the case of the wave equation where information from only two initial points are required to determine solution at \((x_0, t_0)\).

To begin solving the PDE numerically, one discretizes the spatial region as well as the temporal direction. The three ways of writing the (first) derivative are

\[
\frac{\partial f}{\partial x} \approx \frac{f(x + \Delta x, t) - f(x, t)}{\Delta x},
\]

\[
\frac{\partial f}{\partial x} \approx \frac{f(x, t) - f(x - \Delta x, t)}{\Delta x}
\]

and

\[
\frac{\partial f}{\partial x} \approx \frac{f(x + \Delta x, t) - f(x - \Delta x, t)}{2\Delta x}
\]

called the forward, backward and central difference scheme, respectively. Choosing one form will place a constraint on how we discretize the domain. For example, we can write the first-order wave equation\(^7\),

\[
\frac{\partial f}{\partial t} + c \frac{\partial f}{\partial x} = 0
\]

as

\[
\frac{f_{j}^{n+1} - f_{j}^{n}}{\Delta t} + c \frac{f_{j}^{n} - f_{j-1}^{n}}{\Delta x} = 0
\]

\(^6\)This is not to be mistaken with the causal domain. The focus is on the light-cone itself because anything inside will propagate away from the point of interest to the left and to the right at the speed of light passing through the light-cone. Interior points of the lightcone will not contribute to the solution at the point \((x_0, t_0)\).

\(^7\)The wave equation, \(f_{tt} - c^2 f_{xx} = 0\), has two solutions, one propagating in the left direction and one in the right direction. The first-order wave equation, \(f_t + c f_x = 0\), will only propagate in the right direction.
where upper indices indicate temporal slices and lower indices, spatial grid-points. This is called the simple-upwind method and the explicit time-stepping form is

$$f_j^{n+1} = f_j^n - \frac{\Delta t}{\Delta x} (f_j^n - f_{j-1}^n)$$

which tells us exactly how to evolve each point forward in time. Other methods may have an implicit form which dictates how to evolve a whole row of points. If we start with a $k$-wave of $\lambda^1$ amplitude\(^8\), as an initial condition written as

$$f_j^1 = \lambda^1 e^{ikj\Delta x},$$

we must make sure the next iterative step will not amplify the wave and blow up the solution. That is, we must consider $|\lambda^{n+1}/\lambda^n|^2$. After some algebra, we find that

$$\left| \frac{\lambda^{n+1}}{\lambda^n} \right|^2 = 1 - 4 \left( \frac{c\Delta t}{\Delta x} \right) \left( 1 - \left( \frac{c\Delta t}{\Delta x} \right) \right) \sin^2 \left( k\Delta x / 2 \right)$$

which tells us that we require the domain discretization to satisfy

$$0 \leq \left( \frac{c\Delta t}{\Delta x} \right) \leq 1.$$

This means, using the finite difference equation (5.9), we cannot step forward in $\Delta t$ time too much or else the solution will blow up. This is also known as the Courant-Friedrichs-Lewy stability criterion or CFL condition. As a matter of fact, if we require no amplification and no damping, we are then forced to have $c\Delta t = \Delta x$. This is called the characteristic path of the PDE. However, one must not think that the characteristic path will always produce numerically stable results. For example, if we had used a central difference scheme as

$$\frac{f_j^{n+1} - f_j^n}{\Delta t} + c\frac{f_{j+1}^n - f_{j-1}^n}{2\Delta x} = 0$$

we would find that the amplification is

$$\left| \frac{\lambda^{n+1}}{\lambda^n} \right|^2 = 1 + \left( \frac{c\Delta t}{\Delta x} \right)^2 \sin^2 \left( k\Delta x \right)$$

\(^8\)The 1 in $\lambda^1$ is a superscript, not an exponent.
which is always greater than one. This shows that it is unconditionally unstable regardless of the choices of $\Delta t$. This is called von Neumann stability analysis which ensures that numerical results are valid.

Stability analysis is the key to whether the correct numerical solution results. It was found that if the PDE was in the form of

$$\frac{\partial f(x, t)}{\partial t} + \frac{\partial F(f(x, t))}{\partial x} = 0$$

where $F$ is called the flux function, then it will result in the correct solution. For example, the discretization of the inviscid Burgers’ equation in the form,

$$\frac{\partial f}{\partial t} + \frac{\partial (f^2/2)}{\partial x} = 0$$

will result in the correct shock-wave formation but the non-conservative form,

$$\frac{\partial f}{\partial t} + f \frac{\partial f}{\partial x} = 0$$

would not arrive at the correct solution with an offset in the shock-wave no matter how finely spaced the grid points were. This is called non-linear instability.

Once we have the assurance of stability, we can focus upon accuracy. To further increase accuracy, Lax and Wendroff introduced a method[8] in 1960 which increases the accuracy from first-order to second order. However, their method suffers from oscillation (of second order) and does not preserve monotonicity.

If we take the flux function from the simple-upwind method and the flux function from the Lax-Wendroff method, a hybrid can be used

$$F(f, j) = F_{SU}(1 - \phi(f, j)) + \phi(f, j)F_{LW}$$

where one would pick $\phi(f, j)$ between values of 0 and 1 depending on the smoothness of the current iterative profile. This method is called the flux limiter method and it provides the best of both worlds in terms of high accuracy and low oscillation.
5.3 Numerical Application and Results

5.3.1 Curve Fitting

In our application, with \( r \) and \( z \) in kpc, we have a velocity function in m/sec

\[
V(r, z) = -\frac{3 \times 10^8}{r} \sum_{n=1}^{10} C_n k_n e^{-k_n |z|} J_1(k_n r)
\]  
(5.10)

to fit this with data. However, there is a \( z \)-dependence in the function and the data are along the \( r \) direction only.

There are two possibilities for removing the \( z \)-dependence. The first is integrate the velocity for a layer, say 0.1kpc thick, in either \( z \)-direction. However, this creates much ambiguity in connecting the data to the curve. Exactly how thick was the measured disk? Or rather exactly how far should we integrate in the \( z \)-direction? Should there be a characteristic length in the \( z \) direction that would set the limits of integration? What should the criteria for these limits be? The second approach lacks much of these ambiguities. The step in removing the \( z \)-dependence is to simply set \( z = 0 \) and then apply the curve-fitting.

To see which is more natural, consider how the data are recorded. When radio telescopes detect sources from the HI region, a spread of velocities is detected at each radius. But the value which is recorded is the peak velocity. In our model, the velocity drops like \( C_n e^{-k_n |z|} \) which means we would like to match our peak value with the recorded peak value. Thus it is natural to set \( z = 0 \)

\[
V(r, 0) = -\frac{3 \times 10^8}{r} \sum_{n=1}^{10} C_n k_n J_1(k_n r)
\]  
(5.11)

and from there, curve-fit the data.

For each galaxy, 10 parameters were used to curve-fit (5.11) through the simple linear curve-fit algorithm. The coefficients for

\[
N(r, z) = -\sum_{n=1}^{10} C_n k_n r e^{-k_n |z|} J_1(k_n r)
\]
are tabulated in tables 5.1 to 5.4. Here, \( r \) and \( z \) are in kpc. The velocity in m/sec is given by (5.10) and the density in kg/m\(^3\) is given by

\[
\rho(r, z) = 5.64 \times 10^{-14} \frac{(N_r^2 + N_z^2)}{r^2}.
\]

<table>
<thead>
<tr>
<th>( C_n )</th>
<th>( k_n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.0171 ± 0.0032</td>
<td>0.0801608519</td>
</tr>
<tr>
<td>-0.00179 ± 0.00078</td>
<td>0.1840026036</td>
</tr>
<tr>
<td>-0.001033 ± 0.000018</td>
<td>0.2884575971</td>
</tr>
<tr>
<td>-0.000294 ± 0.00019</td>
<td>0.3930511480</td>
</tr>
<tr>
<td>-0.000378 ± 0.000024</td>
<td>0.4976972570</td>
</tr>
<tr>
<td>-0.000190 ± 0.000084</td>
<td>0.6023687989</td>
</tr>
<tr>
<td>-0.000089 ± 0.000079</td>
<td>0.7070545543</td>
</tr>
<tr>
<td>-0.0000097 ± 0.000029</td>
<td>0.8117490509</td>
</tr>
<tr>
<td>-0.000085 ± 0.000056</td>
<td>0.9164493042</td>
</tr>
<tr>
<td>-0.000051 ± 0.000014</td>
<td>1.021153549</td>
</tr>
</tbody>
</table>

Table 5.1: Curve-fitted coefficients for the Milky Way.

The Milky Way’s velocity curve fit is shown in figure 5.5 and the derived density is shown in figure 5.6 along the \( r \) direction and near the \( z \) axis. Using the known density, a contour plot can be constructed as shown in figure 5.7. The velocity curve-fits and the density profiles for NGC 3031, NGC 3198 and NGC 7331 are shown in figures 5.8, 5.9 and 5.10, respectively. All the derived galactic masses from the relativistic models are tabulated in table 5.5 along with comparisons to Newtonian models. As
<table>
<thead>
<tr>
<th>$C_n$</th>
<th>$k_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.0107</td>
<td>0.1093102526</td>
</tr>
<tr>
<td>-0.00174</td>
<td>0.2509126413</td>
</tr>
<tr>
<td>-0.000935</td>
<td>0.3933512687</td>
</tr>
<tr>
<td>-0.00277</td>
<td>0.5359788381</td>
</tr>
<tr>
<td>-0.000123</td>
<td>0.6786780777</td>
</tr>
<tr>
<td>-0.0000504</td>
<td>0.8214119986</td>
</tr>
<tr>
<td>-0.0000445</td>
<td>0.9641653013</td>
</tr>
<tr>
<td>-0.0000497</td>
<td>1.106930524</td>
</tr>
<tr>
<td>-0.000191</td>
<td>1.249703597</td>
</tr>
<tr>
<td>-0.0000932</td>
<td>1.392482112</td>
</tr>
</tbody>
</table>

Table 5.2: Curve-fitted coefficients for NGC 3031. The original data for curve-fitting lacked error bars, but the author estimates the velocity error was 20% which carries over to the $C_n$ coefficient.

As one can see, the relativistic models yield masses that are consistently lower than the Newtonian model by about 30%, a reduction which Balasin and Grumiller[62] also arrived at in their relativistic model.

**Luminosity Threshold**

Using the plot of the log-density, we introduce a tool for predicting luminosity threshold. In the log-density plots shown in figure 5.11, and 5.12, we compare one of the
<table>
<thead>
<tr>
<th>$C_n$</th>
<th>$k_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.01171 ± 0.00018</td>
<td>0.08016085193</td>
</tr>
<tr>
<td>-0.001083 ± 0.000010</td>
<td>0.1840026036</td>
</tr>
<tr>
<td>-0.000818 ± 0.000026</td>
<td>0.2884575971</td>
</tr>
<tr>
<td>-0.00018483 ± 0.0000039</td>
<td>0.3930511480</td>
</tr>
<tr>
<td>-0.000191 ± 0.000016</td>
<td>0.4976972570</td>
</tr>
<tr>
<td>-0.0000717 ± 0.0000059</td>
<td>0.6023687989</td>
</tr>
<tr>
<td>-0.0000609 ± 0.0000094</td>
<td>0.7070545543</td>
</tr>
<tr>
<td>-0.0000259 ± 0.0000032</td>
<td>0.8117490509</td>
</tr>
<tr>
<td>-0.0000381 ± 0.0000042</td>
<td>0.9164493042</td>
</tr>
<tr>
<td>-0.0000383 ± 0.0000056</td>
<td>1.021153549</td>
</tr>
</tbody>
</table>

Table 5.3: Curve-fitted coefficients for NGC 3198.

plots to Kent's luminosity data[35]. For example, in his NGC 3031 data, luminosity$^9$ terminates at 15 kpc. If we draw a horizontal line at -21.75 on our plot of log-density of NGC 3031 in figure 5.11, we find that the lines intersect at the radius where Kent's optical data stops. From the other figures provided by Kent for optical intensity curves and our log-density profiles for the other two galaxies, NGC 3198 and NGC 7331, we determine that the threshold density for the onset of visible galactic light as we probe in the radial direction is at $10^{-21.75}$ kg·m$^{-3}$. It would be of interest to explore as many sources as possible to test the indicated hypothesis that this density is the universal optical luminosity threshold for galaxies as tracked in the radial di-

$^9$for his instruments.
\begin{table}
\begin{tabular}{|c|c|}
\hline
$C_n$ & $k_n$ \\
\hline
-0.023037 ± 0.00043 & 0.06499528536 \\
-0.0021181 ± 0.0000062 & 0.1491913003 \\
-0.0017220 ± 0.000090 & 0.2338845382 \\
-0.0006174 ± 0.0000075 & 0.3186901200 \\
-0.0005263 ± 0.000023 & 0.4035383165 \\
-0.0001954 ± 0.000016 & 0.4884071344 \\
-0.0002167 ± 0.0000089 & 0.5732874765 \\
-0.0001097 ± 0.0000077 & 0.6581749063 \\
-0.0001201 ± 0.000011 & 0.7430670036 \\
-0.00005371 ± 0.0000010 & 0.8279623371 \\
\hline
\end{tabular}
\end{table}

Table 5.4: Curve-fitted coefficients for NGC 7331.

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

Extended Model

It should be noted that at the far edge of the galaxy, there is a relatively large degree of numerical fluctuation as seen in figures 5.11 and 5.12. Beyond 15-20 kpc, the log-density does not exhibit the downward trend and oscillates slightly. In a paper by
<table>
<thead>
<tr>
<th>Galaxy</th>
<th>HI Region (in kpc)</th>
<th>Derived Mass ($\times 10^{10} M_\odot$)</th>
<th>Newtonian Mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Milky Way</td>
<td>30</td>
<td>22.0</td>
<td>20-60</td>
</tr>
<tr>
<td>NGC 3031</td>
<td>22</td>
<td>11.6</td>
<td>13.3 [35]</td>
</tr>
<tr>
<td>NGC 3198</td>
<td>32</td>
<td>10.5</td>
<td>15.1 [35]</td>
</tr>
<tr>
<td>NGC 7331</td>
<td>36</td>
<td>32.1</td>
<td>43.3 [35]</td>
</tr>
</tbody>
</table>

Table 5.5: Derived mass from relativistic models in comparison to published Newtonian models.

Menzies and Mathews [63], they pointed out that this renders the model insufficiently accurate beyond the HI region. In order to complete the global picture, we examine a larger region beyond the HI region by appending fictitious data points to the velocity data for the Milky Way. The edge of the Milky Way’s HI region is only 30 kpc, but we introduce data points from 30 to 300 kpc using different fall-off scenerios: $1/\sqrt{r}$, $1/r$, $1/r^2$ and $1/r^4$. This is shown in figure 5.13.

In order to curve-fit this larger region and to capture the essential features (which leads to density) of the inner 0-30 kpc portion, a larger number of parameters was required. In total, 120 were used to model this extended area which is 10 times wider than the original. The comparison of density in this 120-parameters model to the 10-parameters model yields very little difference in the HI region. Therefore the plot in comparing density has been omitted.

The derived masses of these extended models are computed and shown in figure 5.14. As we can see, in comparison to the original $22.0 \times 10^{10} M_\odot$, for the inner region, there is relatively little accumulated mass beyond the HI region. This 25% to 300% mass increase is in sharp contrast to most Newtonian models which requires a vast dark matter halo adding more than 1000% of the original mass.

In our model, there are some points of contention. It should be noted that these
Figure 5.5: Velocity curve-fit for the Milky Way in units of m/s vs kpc. The data comes from a wide variety of sources[30][37][38][40] from CO emission, to emissions from HI and HII regions. Some data sets do not include any error bars. The error bars shown on the graph come from one set of data. The outer region beyond 15 kpc contains large errors.

extended models do not merge with vacuum. Instead, they taper off to extremely low density in that the contribution is insignificant even for large radii. This is a globally dust model and it does not have a Kerr metric-like property at larger distances so we do not expect any matching between our metric and the Kerr metric. There is still much ambiguity as to how the fall-off scenario should be modelled. One possible approach is the use of velocity dispersion[66]. At the outer radii, different scenarios exhibit different velocity profiles in the z-direction. The difference in profile equips us with a tool to distinguish between the scenarios and provides us with a test to match theory with experiment.

5.4 Numerical Solution to The Boundary Layer

As mentioned in the previous chapter, there are three approaches to resolving the ambiguity of the nature of the (possible) singularity at $z = 0$ for the model. The
Figure 5.6: Derived density profiles in units of kg/m$^3$ for the Milky Way at $z = 0$ (top) and $r = 0.001$ kpc (bottom).
first requires a numerical solution to a PDE. The second requires the matching of two series which will be done using least-square fits. And the last method is simply a numerical computation of the geodesic equation.

5.4.1 Numerical PDE

The first approach requires the removal of a slice of the \( N(r, z) \) function and solving it numerically. The function \( N \) must satisfy the the PDEs,

\[
\frac{\partial^2 N}{\partial z^2} + \frac{\partial^2 N}{\partial r^2} - \frac{1}{r} \frac{\partial N}{\partial r} = 0
\]  \hspace{1cm} (5.12)

\[
\frac{1}{r^2} \left( \left( \frac{\partial N}{\partial z} \right)^2 + \left( \frac{\partial N}{\partial r} \right)^2 \right) = \rho(r, z)
\]  \hspace{1cm} (5.13)
Figure 5.8: Velocity curve-fit and derived density for NGC 3031. Error bars were not included in the published velocity curve data but the author estimates the errors to be 20%.
Figure 5.9: Velocity curve-fit and derived density for NGC 3198
Figure 5.10: Velocity curve-fit and derived density for NGC 7331.
Figure 5.11: Log graphs of density for the Milky Way (left) and NGC 3031 (right) showing the density fall-off. The $-21.75$ dashed line provides a tool to predict the outer limits of visible matter. The fluctuations at the end are the result of limited curve-fitting terms.

Figure 5.12: Log graphs of density for the NGC 3198 (left) and NGC 7331 (right) showing the density fall-off. The $-21.75$ dashed line provides a tool to predict the limits of luminous matter. As before, there are fluctuations near the border.
Figure 5.13: Beyond the HI region, the velocity can be modeled in many different manners: here $V \propto 1/\sqrt{r}$, $V \propto 1/r$, $V \propto 1/r^2$ and $V \propto 1/r^4$ are illustrated.

Figure 5.14: The Milky Way’s accumulated mass as a consequence of velocity fall-off beyond the HI region.
Figure 5.15: This shows the comparison between the original function $N$ (shown as a mesh) which has the $|z|$ characteristics to the numerically-evolved PDE solution (shown as surface) for (5.14). The contour lines for each are used to emphasize the difference in characteristics.
in the “pancake” domain $0 < r < 30$ and $-1 \leq z \leq 1$ with boundary conditions
\[
\frac{\partial N_{(int)}(r, z)}{\partial z} \bigg|_{z=-1} = \frac{\partial N_{(ext)}(r, z)}{\partial z} \bigg|_{z=-1}
\]
\[
\frac{\partial N_{(int)}(r, z)}{\partial z} \bigg|_{z=1} = \frac{\partial N_{(ext)}(r, z)}{\partial z} \bigg|_{z=1}
\]
\[
N_{(int)}(r, 1) = N_{(ext)}(r, 1)
\]
\[
N_{(int)}(r, -1) = N_{(ext)}(r, -1).
\]

The focus here is in removing the gradient discontinuity nature of $\rho(r, z)$ by smoothing it out; thus we must not enforce its characteristic by using (5.13) where $\rho(r, z)$ is the “known source”. Instead, we must let $\rho(r, z)$ be unknown within this strip. That is, we shall deduce the $N$ function from (5.12).

The boundary condition is extremely unusual in that it leaves the solution at the edge of the pancake domain free to wander and there are two enforced conditions at the same location, namely at $z = \pm 1$. There is no known solution to such a boundary condition except for numerical solutions. Looking at the boundary conditions, one can “grow” the solution in the $z$-direction from both the top and from the bottom towards the center. However, this will generally cause a gradient discontinuity unless the solution is extremely fine-tuned. It was pointed out by B. Khoudier that one can view (5.12) as a stationary heat equation
\[
\frac{\partial N}{\partial t} = \frac{\partial^2 N}{\partial z^2} + \frac{\partial^2 N}{\partial r^2} - \frac{1}{r} \frac{\partial N}{\partial r}
\]
(5.14)

with a convection term. We start with an $N(r, z)$ exhibiting a gradient discontinuity and evolve it like a heat equation while maintaining the boundary conditions. Eventually, it will settle down to the stationary solution and the left-hand side of (5.14) becomes zero. This was done for $-1 \leq z \leq 1$ and the solution is plotted in figure 5.15 showing the gradient nature of the original $N(r, z)$ in comparison to the evolved $N(r, z)$.

We can reduce the thickness of our pancake domain and repeat for thinner and thinner domains. With the the range $-1 \leq z \leq 1$, there may be significant changes
Figure 5.16: These graphs show the matching conditions for $N$ and $\partial N/\partial z$ at $z = 1$ kpc. Because of symmetry, the matching conditions at $z = -1$ kpc is identical, except for mirror reflection. In both graphs, $r$ is measured in kpc and $N$ has units of kpc$^{-1}$ while $N_z$ is dimensionless.

to the total mass of the galactic model, but eventually, as the pancake is reduced to a very small region, the contribution will not change the total mass by any significant amount.

5.4.2 Matching Series

In the second approach to resolving the problem, instead of solving the numerical PDE, the sum of the series of 10 known $\cosh(k_n z)$ terms are matched with another series of $e^{-\kappa_m |z|}$. The approach was implemented using the curve-fitting scheme.
However, in this particular case, we do not have the $\kappa_m$ values defined and thus, they will be left as parameters.

To match

$$N_{(\text{int})}(r, z) = - \sum_{n=1}^{10} C_n k_n r \cosh(k_n z) J_1(k_n r)$$

with

$$N_{(\text{ext})}(r, z) = - \sum_{l=1}^{L} C_l k_< r e^{k_l |z|} J_1(k_l r)$$

we start with $L = 20$ which requires 40 parameters: $C_1, \ldots, C_{20}$ and $\kappa_1, \ldots, \kappa_{20}$. This non-linear square fit was accomplished with minimal success. That is, while the matching was within range, it was not precise. It was found that a series of $\kappa_m$ evenly distributed between 0.0 and 2.0 did not result in much drift in $\kappa_m$ after applying the non-linear least square fit. In other words, we start with a set $\kappa_1 = 0.1, \kappa_2 = 0.2, \ldots, \kappa_{20} = 2.0$, apply the non-linear least square fit allowing $\kappa_m$ to be free parameters, we find that each $\kappa_m$ did not change in value afterward. This proves to be extremely useful because it allowed the estimates of the $\kappa_m$ values and reduced the problem from a non-linear to a linear curve-fit. To get a better fit, hundreds of parameters were used. A result of 1000 parameter-fit was used in the matching shown in figure 5.16.

The problem with this large number of parameters is that the total mass computation is not possible due to the large storage space required.

5.4.3 Non-Existence of a Boundary Layer

In the last approach, we compute

$$\frac{dU^z}{ds} = - \frac{\nu_z}{2} (U^z)^2.$$  

where

$$\nu_z = - \frac{N_r N_z}{r}$$

Basically, we need to determine the sign of $N_r N_z$ for different $r$ and $z$ values. If $N_r N_z$ is positive, it means a test particle is repelled away from the boundary layer.
Figure 5.17: Determination of sign of $N_r N_z$ for the Milky Way. It is clearly negative except for the boundary in which $N$ was forced to be zero.

and if it is negative, it means it is attracted towards it. The plot of function $N_r N_z$ is shown in figure 5.17.

As we can see, the function is completely negative, which means there is no negative mass boundary layer. Had there been an enormous amount of negative mass as speculated by some, there would be repulsion away from the $z = 0$ plane. What was found was that if one pushes a test particle in the direction of rotation, the geodesic equation would force the particle into a region where it is co-moving with the dust particles. This creates a stabilizing effect, quite contrary to a violent disintegration of the whole system. These points of evidence clearly show the nature of the “singularity” at $z = 0$. 
Chapter 6

Conclusions

General relativity works within the confines of an open domain. Because the domain of one particular coordinate system does not cover the manifold, we have the freedom to choose the topological structure of the manifold. As we have seen in the Gödel universe and the Gott spacetime, choices were made which led to the existence of CTCs. In a similar prescription, our examples in flat-spacetime can exhibit non-causal structures. While general relativity does not prevent identifications of points separated by time-like intervals, it is ultimately up to us to decide whether or not paradoxical structures are allowed in our models of the physical universe.

Common sense tells us to adopt the causal structure which matches our intuition if we are to model the physical universe we live in. If we observe something that goes against our intuition, all avenues should be explored within our laws of physics. In the dark matter problem which basically arises from gravitational discrepancies in observations, attention should be focused upon models based on the premier law of gravitation, namely general relativity, and not just a limiting form of such, which we have seen to be more subtle than is generally believed.

The relativistic model of galaxies can resolve the flat-rotation curve problem with-
out the need for a vast massive halo of exotic dark matter. The derived total masses of galaxies were consistently lower than the Newtonian model by about 30%. The relativistic model has predictive ability in the threshold density for the onset of luminosity. The extension to the model beyond the HI region allows for modelling velocity-fall-off behaviours. It also provides the ability to compare directly with any observed velocity data in z-slices above the galactic plane. All of these allows for a complete global picture of the model of a galaxy.

The galactic model shows that we must re-examine how we think about general relativity. J. L. Synge wrote in his book[13] “Of all physicists, the general relativist has the least social commitment. He is the great specialist in gravitational theory and gravitation is socially significant, but he is not consulted in the building of a tower, a bridge, a ship, or an aeroplane, and even the astronauts can do without him until they start wondering which ether their signals travel in. Splitting hairs in an ivory tower is not to everyone’s taste, and no doubt many a relativist looks forward to the day when governments will seek his opinion on important questions.” Perhaps if relativists change their perspective on their underlying theory, features of relativity will be brought to light which may have important social significance and indeed governments may some day seek their opinions on important questions.
Bibliography


Appendix

With the metric,

$$ g_{ij} = \begin{bmatrix} -e^{\nu(r, z)} & 0 & 0 & 0 \\ 0 & N(r, z)^2 - r^2 & 0 & -N(r, z) \\ 0 & 0 & -e^{\nu(r, z)} & 0 \\ 0 & -N(r, z) & 0 & 1 \end{bmatrix} $$

the inverse metric is

$$ g^{ij} = \begin{bmatrix} -e^{-\nu(r, z)} & 0 & 0 & 0 \\ 0 & -r^{-2} & 0 & -N(r, z)/r^2 \\ 0 & 0 & -e^{-\nu(r, z)} & 0 \\ 0 & -N(r, z)/r^2 & 0 & 1 - r^{-2}N(r, z)^2 \end{bmatrix} $$

where $x^1 = r$, $x^2 = \phi$, $x^3 = z$ and $x^4 = t$. The (exact) Christoffel symbols of the second kind are

$$ \Gamma^l_{11} = \frac{1}{2} \nu_r $$

$$ \Gamma^l_{22} = \frac{(N N_r - r)}{e^\nu} $$

$$ \Gamma^l_{13} = \frac{1}{2} \nu_z $$

$$ \Gamma^l_{24} = -\frac{1}{2} \frac{N_r}{e^\nu} $$
\[ \Gamma_{33}^1 = -\frac{1}{2} \nu_r \]  
\[ \Gamma_{14}^2 = \frac{1}{2} \frac{N_r}{r^2} \]  
\[ \Gamma_{34}^2 = \frac{1}{2} \frac{N_z}{r^2} \]  
\[ \Gamma_{13}^3 = \frac{1}{2} \nu_r \]  
\[ \Gamma_{24}^3 = -\frac{1}{2} \frac{N_z}{e^{\nu}} \]  
\[ \Gamma_{12}^4 = -\frac{1}{2} \frac{N_r^2}{r^2} + \frac{N}{r} - \frac{1}{2} N_r \]  
\[ \Gamma_{23}^4 = -\frac{1}{2} \frac{N_z N^2}{r^2} - \frac{1}{2} N_z \]  
\[ \Gamma_{12}^5 = -\frac{1}{2} \frac{N N_r}{r^2} + \frac{1}{r} \]  
\[ \Gamma_{23}^5 = -\frac{1}{2} \frac{N N_z}{r^2} \]  

The (exact) Riemann curvature tensor is

\[ R_{1212} = -\frac{3}{4} N_r^2 - N N_{rr} + \frac{1}{2} \nu_r N N_r - \frac{1}{2} N_r \nu_r - \frac{1}{4} \frac{N^2 N_r^2}{r^2} + \frac{N N_r}{r} - \frac{1}{2} \nu_z N N_z \]  
\[ R_{1214} = \frac{1}{2} N_{rr} - \frac{1}{4} \nu_r N_r + \frac{N N_r^2}{4 r^2} - \frac{1}{2} \frac{N_r}{r} + \frac{1}{4} \nu_z N_z \]  
\[ R_{1223} = \frac{3}{4} N_r N_z + N N_{rz} - \frac{1}{2} \nu_z N N_r + \frac{1}{2} r \nu_z + \frac{N^2 N_r N_z}{4 r^2} - \frac{1}{2} \frac{N N_z}{r} - \frac{1}{2} \nu_r N N_z \]  
\[ R_{1234} = \frac{1}{2} N_{rz} - \frac{1}{4} \nu_z N_r + \frac{N_r N N_z}{4 r^2} - \frac{1}{4} \nu_r N_z \]  
\[ R_{1313} = \frac{1}{2} (\nu_{rr} + \nu_{zz}) e^{\nu} \]  
\[ R_{1324} = \frac{1}{2} \frac{N_z}{r} \]  
\[ R_{1414} = -\frac{1}{4} \frac{N_r^2}{r^2} \]  
\[ R_{1423} = -\frac{1}{2} N_{rz} + \frac{1}{4} \nu_z N_r - \frac{1}{4} \frac{N_r N N_z}{r^2} + \frac{N_z}{2 r} + \frac{1}{4} \nu_r N_z \]  
\[ R_{1434} = -\frac{1}{4} \frac{N_z N_r}{r^2} \]  
\[ R_{2323} = -\frac{3}{4} N_z^2 - N N_{zz} - \frac{1}{2} \nu_r N N_r + \frac{1}{2} r \nu_r + \frac{1}{2} \nu_z N N_z - \frac{1}{4} \frac{N^2 N_z^2}{r^2} \]
\[ R_{2334} = -\frac{1}{2} N_{zz} - \frac{1}{4} \nu_z N_r - \frac{1}{4} \frac{N N_z^2}{r^2} + \frac{1}{4} \nu_z N_z \]

\[ R_{2424} = -\frac{e^{-\nu}}{4} \left( N_r^2 + N_z^2 \right) \]

\[ R_{3434} = -\frac{1}{4} \frac{N_z^2}{r^2} \]

And the Ricci tensor is

\[ R_{11} = -\frac{N_r^2}{2r^2} + \frac{1}{2} \left( \nu_{zz} + \nu_{rr} - \frac{\nu_r}{r} \right) \]

\[ R_{13} = -\frac{N_z N_r}{2r^2} - \frac{\nu_z}{2r} \]

\[ R_{22} = -\left( \frac{1}{2} \left( 1 + \frac{N}{r^2} \right) \left( N_r^2 + N_z^2 \right) + N \left( N_{zz} + N_{rr} - \frac{N_r}{r} \right) \right) e^{-\nu} \]

\[ R_{24} = \left( \frac{1}{2} \left( N_{rr} + N_{zz} - \frac{N_r}{r} \right) + \frac{N}{2r^2} \left( N_r^2 + N_z^2 \right) \right) e^{-\nu} \]

\[ R_{33} = \frac{1}{2} \left( \nu_{rr} + \nu_{zz} + \frac{\nu_r}{r} \right) - \frac{N_z^2}{2r^2} \]

\[ R_{44} = -\frac{e^{-\nu}}{2r^2} \left( N_r^2 + N_z^2 \right) \]

The Ricci scalar is

\[ R = \frac{e^{-\nu}}{2r^2} \left( N_r^2 + N_z^2 \right) - \left( \nu_{rr} + \nu_{zz} \right) e^{-\nu} \]