Modern General Relativity Lecture Notes

Mike Guidry

This document summarizes all chapters of the first edition of *Modern General Relativity: Black Holes, Gravitational Waves, and Cosmology* by Mike Guidry (Cambridge University Press, 2019) in a format suitable for presentation. In the interest of conciseness, these presentation notes generally omit references and sources. These are documented fully in the book *Modern General Relativity*. This document is produced with LaTeX in amsmath style. For adopters who may wish to customize these lectures the LaTeX source files are available from the resources page for *Modern General Relativity* at the Cambridge University Press website.

Contents

Part I

General Relativity

Chapter 1

Introduction

General relativity is a theory of gravity that represents a radical new view of space and time.

- It supercedes Newtonian mechanics and Newtonian gravity.
- It reduces to those theories in the limit of velocities that are small with respect to the speed of light c and gravitational fields that are weak.
- It reduces to the theory of special relativity in the limit of weak gravitational fields, or for sufficiently local regions of spacetime in the presence of strong gravitational fields.

General relativity revises fundamentally the very meaning of space, time, and gravity:

- The effects of gravity *no longer appear as a force* but as the motion of *free particles* constrained to move in the straightest paths possible in a curved spacetime.
- That is, general relativity will identify the effects of gravity as arising from curvature in spacetime itself on *free particles*.

> John Wheeler: *mass tells space how to curve; curved space tells matter how to move.*

- Implied in the circularity of this statement is another basic feature ofgeneral relativity: it is a highly non-linear theory:

> Only when we know the curvature of space can we know the distribution and motion of matter, *but* the curvature of space is only understood when we know the distribution and motion of the matter.

As a result of the non-linear nature of general relativity and its formulation on a 4-dimensional spacetime manifold, it is difficult to find exact solutions and only a few of clear physical significance are known.

- In the general case one must solve the resulting non-linear equations numerically *(numerical relativity)*.
- However, we shall see that the simplest known solutions of general relativity may be formulated in remarkably transparent and elegant mathematical terms because of symmetries.
- These formulations may then be used to understand some of the most intriguing aspects of the theory:
  - *black holes*,
  - *quasars*,
  - *gamma-ray bursts*,
  - *dark matter*,
  - *dark energy*,
  - *the new cosmology*
  - *gravitational waves.*

These lecture notes are an attempt to come to grips with these ideas at a level appropriate for an advanced undergraduate physics major.

Chapter 2

Coordinate Systems and Transformations

A physical system has a symmetry under some operation if the system after the operation is observationally indistinguishable from the system before the operation.

Example: A perfectly uniform sphere has a symmetry under rotation about any axis because after the rotation the sphere looks the same as before the rotation.

The theory of relativity may be viewed as a symmetry under coordinate transformations.

- Two observers, referencing their measurements of the same physical phenomena to two different coordinate systems should deduce the *same laws of physics* from their observations.
- In special relativity one requires a symmetry under only a subset of possible coordinate transformations (those between systems that are not accelerated with respect to each other).
- General relativity requires that the laws of physics be invariant under the most general coordinate transformations.

To understand general relativity we must begin by examining the transformations that are possible between different coordinate systems.

2.1 Coordinate Systems in Euclidean Space

Our goal is to describe transformations between coordinates in a general curved space having

- three space-like coordinates and
- one timelike coordinate.

However, to introduce these concepts we shall begin with the simpler and more familiar case of vector fields defined in three-dimensional euclidean space.

- Assume a three-dimensional euclidean (flat) space having a cartesian coordinate system (x,y,z), and an associated set of mutually orthogonal unit vectors $(\boldsymbol{i},\boldsymbol{j},\boldsymbol{k})$.
- Assume that there is an alternative coordinate system (u,v,w), not necessarily cartesian, with the (x,y,z) coordinates related to the (u,v,w) coordinates by

$$x = x(u,v,w) \qquad y = y(u,v,w) \qquad z = z(u,v,w),$$

- Assume that the transformation is invertible so that we can solve for (u,v,w) in terms of (x,y,z).

Example 2.1

Take the (u,v,w) system to be the spherical coordinates (r,θ,φ), in which case

$$x = x(u,v,w) \qquad y = y(u,v,w) \qquad z = z(u,v,w),$$

takes the familiar form

$$x = r\sin\theta\cos\varphi \qquad y = r\sin\theta\sin\varphi \qquad z = r\cos\theta,$$

with the ranges of values $r \geq 0$ and $0 \leq \theta \leq \pi$ and $0 \leq \varphi \leq 2\pi$.

- The equations

$$x = x(u,v,w) \qquad y = y(u,v,w) \qquad z = z(u,v,w),$$

can be combined into a vector equation that gives a position vector $\boldsymbol{r}$ for a point in the space in terms of the (u,v,w) coordinates:

$$\boldsymbol{r} = x(u,v,w)\,\boldsymbol{i} + y(u,v,w)\,\boldsymbol{j} + z(u,v,w)\,\boldsymbol{k}.$$

- For example, in terms of the spherical coordinates (r,θ,φ),

$$\boldsymbol{r} = (r\sin\theta\cos\varphi)\,\boldsymbol{i} + (r\sin\theta\sin\varphi)\,\boldsymbol{j} + (r\cos\theta)\,\boldsymbol{k}.$$

- The second coordinate system in these examples generally is not cartesian but the space is still assumed to be euclidean.
- In transforming from the (x,y,z) coordinates to the (r,θ,φ) coordinates, we are just using a different scheme to label points in a flat space.
- This distinction is important because shortly we shall consider general coordinate transformations in spaces that may not obey euclidean geometry (curved spaces).

2.1.1 Basis Vectors

At any point $P(u_0, v_0, w_0)$ defined for specified coordinates (u_0, v_0, w_0), three surfaces pass. They are defined by $u = u_0$, $v = v_0$, and $w = w_0$, respectively.

- Any two of these surfaces meet in curves.
- From

$$\boldsymbol{r} = x(u,v,w)\,\boldsymbol{i} + y(u,v,w)\,\boldsymbol{j} + z(u,v,w)\,\boldsymbol{k}.$$

  we may obtain general parametric equations for coordinate surfaces or curves by setting one or two of the variables (u, v, w) equal to constants.
- For example, if we set v and w to constant values, $v = v_0$ and $w = w_0$, we obtain a parametric equation for a curve given by the intersection of $v = v_0$ and $w = w_0$,

$$\boldsymbol{r}(u) = x(u,v_0,w_0)\,\boldsymbol{i} + y(u,v_0,w_0)\,\boldsymbol{j} + z(u,v_0,w_0)\,\boldsymbol{k},$$

- This is a parametric equation in which u plays the role of a coordinate along the curve defined by the constraints $v = v_0$ and $w = w_0$.

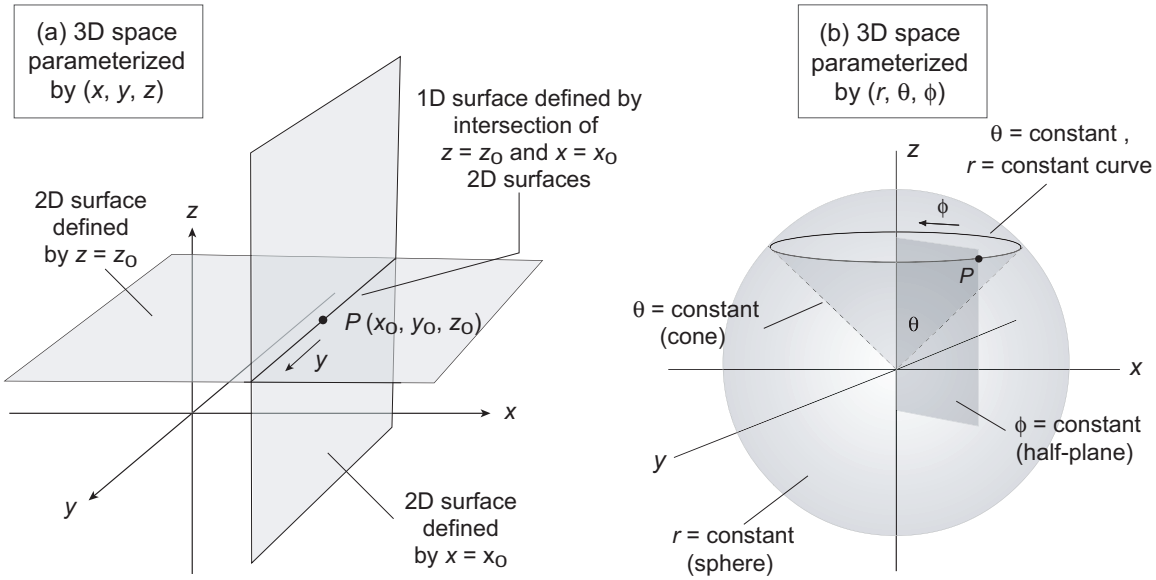


Figure 2.1: Examples of surfaces and curves arising from constraints. (a) In 3D euclidean space parameterized by cartesian coordinates (x, y, z), the constraints $x = x_0$ and $z = z_0$ define 2D planes and the intersection of these planes defines a 1D surface parameterized by the variable y. (b) In 3D space described in spherical coordinates (r, θ, φ), the constraint $r =$ constant defines a 2D sphere, the constraint $\theta =$ constant defines a cone, and the constraint $\varphi =$ constant defines a half-plane. The intersection of any two of these surfaces defines a curve parameterized by the variable not being held constant.

Fig. 2.1(b) illustrates for spherical coordinates (r, θ, φ):

- The surface corresponding to r = constant is a sphere parameterized by the variables θ and φ.
- The constraint θ = constant corresponds to a cone parameterized by the variables r and φ.
- Setting both r and θ to constants defines a curve that is the intersection of the sphere and the cone, which is parameterized by the variable φ.

- Partial differentiation of

$$\boldsymbol{r} = x(u,v,w)\,\boldsymbol{i} + y(u,v,w)\,\boldsymbol{j} + z(u,v,w)\,\boldsymbol{k},$$

  with respect to u, v, and w, respectively, gives tangents to the coordinate curves passing though the point P.

- These may be used to define a set of basis vectors $\boldsymbol{e}_i$ through

$$\boldsymbol{e}_u \equiv \frac{\partial \boldsymbol{r}}{\partial u} \qquad \boldsymbol{e}_v \equiv \frac{\partial \boldsymbol{r}}{\partial v} \qquad \boldsymbol{e}_w \equiv \frac{\partial \boldsymbol{r}}{\partial w},$$

  with all partial derivatives evaluated at the point $P = (u_0, v_0, w_0)$.

- This basis, generated by the tangents to the coordinate curves, is sometimes termed the *natural basis.* The following example illustrates for a spherical coordinate system.

Example 2.2

Consider the spherical coordinate system defined through

$$x = r\sin\theta\cos\varphi \qquad y = r\sin\theta\sin\varphi \qquad z = r\cos\theta.$$

The position vector $\boldsymbol{r}$ is

$$\boldsymbol{r} = (r\sin\theta\cos\varphi)\,\boldsymbol{i} + (r\sin\theta\sin\varphi)\,\boldsymbol{j} + (r\cos\theta)\,\boldsymbol{k},$$

and the natural basis is obtained from

$$\begin{aligned}
\boldsymbol{e}_1 \equiv \boldsymbol{e}_r &= \frac{\partial \boldsymbol{r}}{\partial r} = (\sin\theta\cos\varphi)\,\boldsymbol{i} + (\sin\theta\sin\varphi)\,\boldsymbol{j} + (\cos\theta)\,\boldsymbol{k} \\
\boldsymbol{e}_2 \equiv \boldsymbol{e}_\theta &= \frac{\partial \boldsymbol{r}}{\partial \theta} = (r\cos\theta\cos\varphi)\,\boldsymbol{i} + (r\cos\theta\sin\varphi)\,\boldsymbol{j} - (r\sin\theta)\,\boldsymbol{k} \\
\boldsymbol{e}_3 \equiv \boldsymbol{e}_\varphi &= \frac{\partial \boldsymbol{r}}{\partial \varphi} = -(r\sin\theta\sin\varphi)\,\boldsymbol{i} + (r\sin\theta\cos\varphi)\,\boldsymbol{j}.
\end{aligned}$$

These basis vectors are mutually orthogonal because

$$\boldsymbol{e}_1\cdot\boldsymbol{e}_2 = \boldsymbol{e}_2\cdot\boldsymbol{e}_3 = \boldsymbol{e}_3\cdot\boldsymbol{e}_1 = 0$$

For example,

$$\begin{aligned}
\boldsymbol{e}_1\cdot\boldsymbol{e}_2 &= r\sin\theta\cos\theta\cos^2\varphi + r\sin\theta\cos\theta\sin^2\varphi - r\cos\theta\sin\theta \\
&= r\sin\theta\cos\theta\underbrace{(\cos^2\varphi + \sin^2\varphi)}_{=1} - r\cos\theta\sin\theta = 0.
\end{aligned}$$

From the scalar products of the basis vectors with themselves, their lengths are

$$|\boldsymbol{e}_1| = 1 \qquad |\boldsymbol{e}_2| = r \qquad |\boldsymbol{e}_3| = r\sin\theta$$

and we can use these to define a normalized basis,

$$\begin{aligned}
\hat{\boldsymbol{e}}_1 &\equiv \frac{\boldsymbol{e}_1}{|\boldsymbol{e}_1|} = (\sin\theta\cos\varphi)\,\boldsymbol{i} + (\sin\theta\sin\varphi)\,\boldsymbol{j} + (\cos\theta)\,\boldsymbol{k} \\
\hat{\boldsymbol{e}}_2 &\equiv \frac{\boldsymbol{e}_2}{|\boldsymbol{e}_2|} = (\cos\theta\cos\varphi)\,\boldsymbol{i} + (\cos\theta\sin\varphi)\,\boldsymbol{j} - (\sin\theta)\,\boldsymbol{k} \\
\hat{\boldsymbol{e}}_3 &\equiv \frac{\boldsymbol{e}_3}{|\boldsymbol{e}_3|} = -(\sin\varphi)\,\boldsymbol{i} + (\cos\varphi)\,\boldsymbol{j}.
\end{aligned}$$

These basis vectors are now

- mutually orthogonal and
- of unit length.

They are illustrated in the following figure.

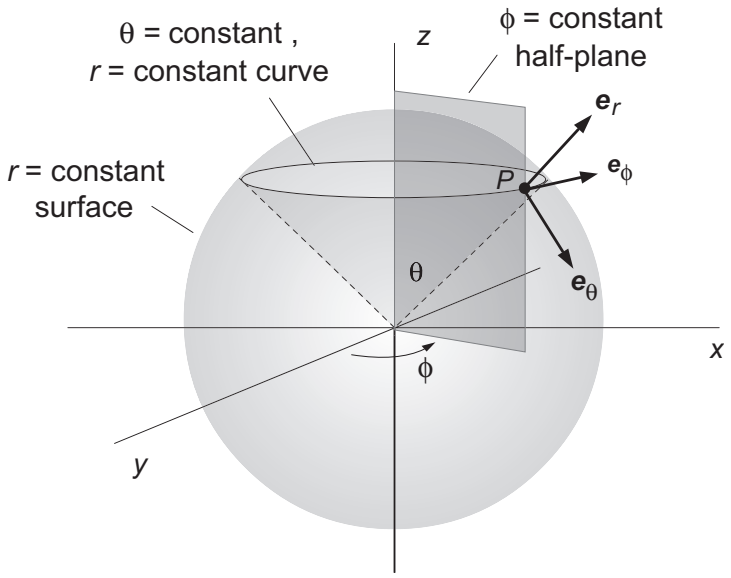


Figure 2.2: Basis vectors for the natural basis in spherical coordinates.

Figure 2.2 illustrates the geometry of the basis vectors derived in the preceding example.

- In many applications it is usual to assume that the coordinate system is orthogonal so that the basis vectors

$$\boldsymbol{e}_u \equiv \frac{\partial \boldsymbol{r}}{\partial u} \qquad \boldsymbol{e}_v \equiv \frac{\partial \boldsymbol{r}}{\partial v} \qquad \boldsymbol{e}_w \equiv \frac{\partial \boldsymbol{r}}{\partial w},$$

  are mutually orthogonal, and to normalize these basis vectors to unit length.

- In the more general applications that will interest us, the natural basis defined by the partial derivatives in the preceding equation need not be orthogonal or normalized to unit length

However, in the simple examples shown so far the natural basis is in fact orthogonal.

2.1.2 Dual Basis

It is also valid to construct a basis at P by using *normals rather than the tangents* to the coordinate surfaces.

- We assume that

$$x = x(u,v,w) \qquad y = y(u,v,w) \qquad z = z(u,v,w),$$

is invertible so we may solve for

$$u = u(x,y,z) \qquad v = v(x,y,z) \qquad w = w(x,y,z),$$

- The gradients

$$\nabla u = \frac{\partial u}{\partial x}\boldsymbol{i} + \frac{\partial u}{\partial y}\boldsymbol{j} + \frac{\partial u}{\partial z}\boldsymbol{k}$$
$$\nabla v = \frac{\partial v}{\partial x}\boldsymbol{i} + \frac{\partial v}{\partial y}\boldsymbol{j} + \frac{\partial v}{\partial z}\boldsymbol{k}$$
$$\nabla w = \frac{\partial w}{\partial x}\boldsymbol{i} + \frac{\partial w}{\partial y}\boldsymbol{j} + \frac{\partial w}{\partial z}\boldsymbol{k}$$

are normal to the three surfaces through P defined by $u = u_0$, $v = v_0$, and $w = w_0$, respectively.

- Therefore, we may choose as an alternative to the natural basis

$$\boldsymbol{e}_u \equiv \frac{\partial \boldsymbol{r}}{\partial u} \qquad \boldsymbol{e}_v \equiv \frac{\partial \boldsymbol{r}}{\partial v} \qquad \boldsymbol{e}_w \equiv \frac{\partial \boldsymbol{r}}{\partial w},$$

the basis

$$\boldsymbol{e}^u \equiv \nabla u \qquad \boldsymbol{e}^v \equiv \nabla v \qquad \boldsymbol{e}^w \equiv \nabla w.$$

- This basis $(\boldsymbol{e}^u, \boldsymbol{e}^v, \boldsymbol{e}^w)$, defined in terms of normals, is said to be the *dual* of the normal basis, defined in terms of tangents.
- Notice that we have chosen to distinguish the basis

$$\boldsymbol{e}^u \equiv \boldsymbol{\nabla} u \qquad \boldsymbol{e}^v \equiv \boldsymbol{\nabla} v \qquad \boldsymbol{e}^w \equiv \boldsymbol{\nabla} w.$$

  from the basis

$$\boldsymbol{e}_u \equiv \frac{\partial \boldsymbol{r}}{\partial u} \qquad \boldsymbol{e}_v \equiv \frac{\partial \boldsymbol{r}}{\partial v} \qquad \boldsymbol{e}_w \equiv \frac{\partial \boldsymbol{r}}{\partial w},$$

  by using *superscript indices* and *subscript indices*, respectively.

These two bases are equally valid.

- For orthogonal coordinate systems the set of normals to the planes corresponds to the set of tangents to the curves in orientation, differing possibly only in length.
- If the basis vectors are normalized, the normal basis and the dual basis for orthogonal coordinates are equivalent and our preceding distinction is not significant.
- However, for non-orthogonal coordinate systems the two bases generally are not equivalent and the distinction between upper and lower indices is relevant.

The following example illustrates.

Example 2.3

Define a coordinate system (u,v,w) in terms of cartesian coordinates (x,y,z) through

$$x = u+v \qquad y = u-v \qquad z = 2uv+w.$$

The position vector for a point $\boldsymbol{r}$ is then

$$\boldsymbol{r} = x\,\boldsymbol{i}+y\,\boldsymbol{j}+z\,\boldsymbol{k} = (u+v)\,\boldsymbol{i}+(u-v)\,\boldsymbol{j}+(2uv+w)\,\boldsymbol{k}$$

The natural basis is

$$\begin{aligned}\boldsymbol{e}_1 \equiv \boldsymbol{e}_u &= \frac{\partial \boldsymbol{r}}{\partial u} = \boldsymbol{i}+\boldsymbol{j}+2v\boldsymbol{k}\\ \boldsymbol{e}_2 \equiv \boldsymbol{e}_v &= \frac{\partial \boldsymbol{r}}{\partial v} = \boldsymbol{i}-\boldsymbol{j}+2u\boldsymbol{k}\\ \boldsymbol{e}_3 \equiv \boldsymbol{e}_w &= \frac{\partial \boldsymbol{r}}{\partial w} = \boldsymbol{k}.\end{aligned}$$

Solving the original equations for (u,v,w),

$$u = \tfrac{1}{2}(x+y) \qquad v = \tfrac{1}{2}(x-y) \qquad w = z-\tfrac{1}{2}(x^2-y^2),$$

and thus the dual basis is

$$\begin{aligned}\boldsymbol{e}^1 \equiv \boldsymbol{e}^u &= \boldsymbol{\nabla} u = \frac{\partial u}{\partial x}\boldsymbol{i}+\frac{\partial u}{\partial y}\boldsymbol{j}+\frac{\partial u}{\partial z}\boldsymbol{k} = \tfrac{1}{2}(\boldsymbol{i}+\boldsymbol{j})\\ \boldsymbol{e}^2 \equiv \boldsymbol{e}^v &= \boldsymbol{\nabla} v = \frac{\partial v}{\partial x}\boldsymbol{i}+\frac{\partial v}{\partial y}\boldsymbol{j}+\frac{\partial v}{\partial z}\boldsymbol{k} = \tfrac{1}{2}(\boldsymbol{i}-\boldsymbol{j})\\ \boldsymbol{e}^3 \equiv \boldsymbol{e}^w &= \boldsymbol{\nabla} w = \frac{\partial w}{\partial x}\boldsymbol{i}+\frac{\partial w}{\partial y}\boldsymbol{j}+\frac{\partial w}{\partial z}\boldsymbol{k} = -(u+v)\,\boldsymbol{i}+(u-v)\,\boldsymbol{j}+\boldsymbol{k}.\end{aligned}$$

- What about orthogonality? We can check by taking scalar products. For example,

$$\begin{aligned} \boldsymbol{e}_1 \cdot \boldsymbol{e}_2 &= (\boldsymbol{i} + \boldsymbol{j} + 2v\boldsymbol{k}) \cdot (\boldsymbol{i} - \boldsymbol{j} + 2u\boldsymbol{k}) \\ &= \boldsymbol{i}^2 - \boldsymbol{j}^2 + 4uv = 4uv, \end{aligned}$$

  where the orthonormality of the basis $(\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k})$ has been used. For the natural basis we find in general

$$\boldsymbol{e}_1 \cdot \boldsymbol{e}_2 = 4uv \qquad \boldsymbol{e}_2 \cdot \boldsymbol{e}_3 = 2u \qquad \boldsymbol{e}_3 \cdot \boldsymbol{e}_1 = 2v.$$

  Thus the normal basis is *non-orthogonal.*

- Taking the scalar products of the natural basis vectors with themselves gives

$$\boldsymbol{e}_1 \cdot \boldsymbol{e}_1 = 2 + 4v^2 \qquad \boldsymbol{e}_2 \cdot \boldsymbol{e}_2 = 2 + 4u^2 \qquad \boldsymbol{e}_3 \cdot \boldsymbol{e}_3 = 1,$$

  so the natural basis is also *not normalized to unit length.*

- It is also clear from the above expressions that generally $\boldsymbol{e}_i$ is not parallel to $\boldsymbol{e}^i$, so in this non-orthogonal case we see that *the normal basis and the dual basis are distinct.*

The preceding example illustrates that for the general case of coordinate systems that are not orthogonal,

$$\boldsymbol{e}^u \equiv \boldsymbol{\nabla} u \qquad \boldsymbol{e}^v \equiv \boldsymbol{\nabla} v \qquad \boldsymbol{e}^w \equiv \boldsymbol{\nabla} w \qquad \text{(dual basis)}$$

and

$$\boldsymbol{e}_u \equiv \frac{\partial \boldsymbol{r}}{\partial u} \qquad \boldsymbol{e}_v \equiv \frac{\partial \boldsymbol{r}}{\partial v} \qquad \boldsymbol{e}_w \equiv \frac{\partial \boldsymbol{r}}{\partial w} \qquad \text{(natural basis)}$$

define *different but equally valid bases*, and the placement of indices in upper or lower positions is important.

- In general relativity we shall generally be dealing with *non-orthogonal coordinate systems.*
- Henceforth the reader should assume that the *upper or lower placement of indices in equations is significant.*

2.1.3 Expansion of Vectors

An arbitrary vector $\boldsymbol{V}$ may be expanded in terms of the tangent basis $\{\boldsymbol{e}_i\}$ and an arbitrary dual vector $\boldsymbol{\omega}$ may be expanded in terms of the dual basis $\{\boldsymbol{e}^i\}$:

$$\boldsymbol{V} = V^1\boldsymbol{e}_1 + V^2\boldsymbol{e}_2 + V^3\boldsymbol{e}_3 = \sum_{i=1}^{3} V^i\boldsymbol{e}_i \equiv V^i\boldsymbol{e}_i \qquad \text{(natural basis)}$$

$$\boldsymbol{\omega} = \omega_1\boldsymbol{e}^1 + \omega_2\boldsymbol{e}^2 + \omega_3\boldsymbol{e}^3 = \sum_{i=1}^{3} \omega_i\boldsymbol{e}^i \equiv \omega_i\boldsymbol{e}^i \qquad \text{(dual basis)}$$

where we have introduced in the last step of each equation the *Einstein summation convention:*

- An index appearing twice on one side of an equation, once as a lower index and once as an upper index, implies a *summation on that index.*
- The summation index is termed a *dummy index;* summation on a dummy index on one side of an equation implies that it does not appear on the other side of the equation.
- If an index appears *more than twice* on the same side of an equation, it *probably indicates a mistake*.
- Since the dummy (repeated) index is summed over, it does not matter what the repeated index is, as long as it is not equivalent to another index in the equation.

> From this point onward, we shall usually assume the Einstein summation convention.

2.1.4 Scalar Product of Vectors and the Metric Tensor

- The upper-index coefficients V^i of the basis vectors $\boldsymbol{e}_i$ in

$$\boldsymbol{V}=V^1\boldsymbol{e}_1+V^2\boldsymbol{e}_2+V^3\boldsymbol{e}_3$$

  are termed the *components of the vector* in the basis $\boldsymbol{e}_i = \{\boldsymbol{e}_1,\boldsymbol{e}_2,\boldsymbol{e}_3\}$.

- The lower-index coefficients ω_i of the basis vectors $\boldsymbol{e}^i$ in

$$\boldsymbol{\omega}=\omega_1\boldsymbol{e}^1+\omega_2\boldsymbol{e}^2+\omega_3\boldsymbol{e}^3$$

  are termed the *components of the dual vector* in the basis $\boldsymbol{e}^i = \{\boldsymbol{e}^1,\boldsymbol{e}^2,\boldsymbol{e}^3\}$.

- Remember: *Components of vectors and dual vectors generally are distinct* for non-orthogonal coordinate systems.

However,

- Vector and dual vector spaces are *related fundamentally*.
- This permits vector components V^i and dual vector components ω_i to be treated as if they were different components of the same vector.

The first step in establishing this relationship is to introduce a *scalar product* and a *metric*.

2.1.5 Vector Scalar Product and the Metric Tensor

Utilizing the expansions in a vector basis:

- We can write the scalar product of two vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ as

$$\boldsymbol{A}\cdot\boldsymbol{B} = (A^i\boldsymbol{e}_i)\cdot(B^j\boldsymbol{e}_j) = \boldsymbol{e}_i\cdot\boldsymbol{e}_j A^i B^j = g_{ij}A^i B^j,$$

  where the *metric tensor component* g_{ij} is defined by

$$g_{ij} \equiv \boldsymbol{e}_i\cdot\boldsymbol{e}_j.$$

- Likewise, for the scalar product of dual vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$

$$\boldsymbol{\alpha}\cdot\boldsymbol{\beta} = \alpha_i\boldsymbol{e}^i\cdot\beta_j\boldsymbol{e}^j = g^{ij}\alpha_i\beta_j,$$

  where metric tensor components with upper indices are

$$g^{ij} \equiv \boldsymbol{e}^i\cdot\boldsymbol{e}^j,$$

  and the scalar product of dual vectors and vectors is

$$\boldsymbol{\alpha}\cdot\boldsymbol{B} = \alpha_i\boldsymbol{e}^i\cdot B^j\boldsymbol{e}_j = g^i_j\alpha_i B^j,$$

  where the metric tensor component with mixed indices is

$$g^i_j \equiv \boldsymbol{e}^i\cdot\boldsymbol{e}_j.$$

General properties of the metric tensor will be discussed below but first we use it to establish a relationship called *duality* between vector and dual vector spaces.

2.1.6 Relationship of Vectors and Dual vectors

There is little practical difference between vectors and dual vectors in euclidean space with cartesian coordinates.

- However, in a curved space the situation is more complex.
- The essential issue is *how to define a vector or dual vector in a curved space,* and what that implies.

The essential mathematics will be discussed in more depth later, but the salient points are that

1. Vectors are not specified directly in a curved space, but instead are defined in a euclidean vector space attached to the manifold at each point called the *tangent space.*
2. Likewise, dual vectors are defined in a euclidean vector space attached to the manifold at each spacetime point that is called the *cotangent space.*
3. The tangent space of vectors and the cotangent space of dual vectors at a point P are different but *dual to each other* in a manner that will be made precise below.
4. This duality allows objects in the two different spaces to be treated as effectively the same kinds of objects.

As will be discussed further later, vectors and dual vectors are special cases of *tensors,* and this permits an abstract definition in terms of mappings from vectors and dual vectors to real numbers.

To be specific,

- Dual vectors $\boldsymbol{\omega}$ are linear maps of vectors $\boldsymbol{V}$ to the real numbers: $\boldsymbol{\omega}(\boldsymbol{V}) = \omega_i V^i \in \mathbb{R}$.
- Vectors $\boldsymbol{V}$ are linear maps of dual vectors $\boldsymbol{\omega}$ to the real numbers: $\boldsymbol{V}(\boldsymbol{\omega}) = V^i \omega_i \in \mathbb{R}$.

Expressions like $\boldsymbol{\omega}(\boldsymbol{V}) = \omega_i V^i \in \mathbb{R}$ can be read as

- "Dual vectors $\boldsymbol{\omega}$ act linearly on vectors $\boldsymbol{V}$ to produce $\omega_i V^i \equiv \sum_i \omega_i V^i$, which are elements of the real numbers,"
- or "Dual vectors $\boldsymbol{\omega}$ are functions (maps) that take vectors $\boldsymbol{V}$ as arguments and yield $\omega_i V^i$, which are real numbers",

Linearity of the mapping means, for example,

$$\boldsymbol{\omega}(\alpha \boldsymbol{A} + \beta \boldsymbol{B}) = \alpha \boldsymbol{\omega}(\boldsymbol{A}) + \beta \boldsymbol{\omega}(\boldsymbol{B}),$$

where $\boldsymbol{\omega}$ is a dual vector, α and β are arbitrary real numbers, and $\boldsymbol{A}$ and $\boldsymbol{B}$ are arbitrary vectors.

- It is easy to show that the space of vectors and the space of dual vectors are both linear vector spaces.
- The vector space of vectors and corresponding vector space of dual vectors are said to be *dual to each other* because they are related by

$$\boldsymbol{\omega}(\boldsymbol{V}) = \boldsymbol{V}(\boldsymbol{\omega}) = V^i \omega_i \in \mathbb{R}.$$

Notice further that $\boldsymbol{A} \cdot \boldsymbol{B} = g_{ij} A^i B^j$

- Defines a linear map from the vectors to the real numbers, since it takes two vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ as arguments and returns the scalar product, which is a real number.
- Thus one may write

$$\boldsymbol{A}(\boldsymbol{B}) = \boldsymbol{A} \cdot \boldsymbol{B} \equiv A_i B^i = g_{ij} A^i B^j.$$

- But since in $A_i B^i = g_{ij} A^i B^j$ the vector B is arbitrary,

$$A_i = g_{ij} A^j,$$

- This specifies a correspondence between a vector with components A^i in the tangent space and a dual vector with components A_i in the cotangent space.
- Likewise, the above expression can be inverted using that the inverse of g_{ij} is g^{ij} to give

$$A^i = g^{ij} A_j.$$

- Hence, using the metric tensor to raise and lower indices by summing over a repeated index (*contraction*),
- we see that *vector and dual vector components are related through contraction with the metric tensor.*
- This is the precise sense in which the tangent and cotangent spaces are dual: *they are different, but closely related through the metric tensor.*

The duality of the vector and dual vector spaces may be incorporated concisely by

- Requiring that for the basis vectors $\{\boldsymbol{e}_i\}$ and basis dual vectors $\{\boldsymbol{e}^i\}$ satisfy

$$\boldsymbol{e}^i(\boldsymbol{e}_j) = \boldsymbol{e}^i \cdot \boldsymbol{e}_j = \delta^i_j,$$

  where the Kronecker delta is defined by

$$\delta^i_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}.$$

- This implies that the basis vectors can be used to project out the components of a vector $\boldsymbol{V}$ by taking the scalar product with the vector,

$$V^i = \boldsymbol{e}^i \cdot \boldsymbol{V} \qquad V_i = \boldsymbol{e}_i \cdot \boldsymbol{V}.$$

A lot of important mathematics has transpired in the last few equations, so let's take stock.

- For a space with metric tensor, vectors and dual vectors are in a one-to-one relationship that permits them to be manipulated effectively as if a dual vector were just a vector with a lower index.
- Indices on vectors can be raised or lowered as desired by contraction with the metric tensor.

Since all spaces of interest here have metrics, this reduces the practical implications of the distinction between vectors and dual vectors to keeping proper track of upper and lower positions for indices.

2.1.7 Properties of the Metric Tensor

- Because it may be defined through scalar products of basis vectors, the metric tensor must be symmetric in its indices:

$$g^{ij} = g^{ji} \qquad g_{ij} = g_{ji}.$$

- Since

$$g_{ij}a^i b^j = g_{ij}b^j a^i = a^i b_i \qquad g^{ij}a_i b_j = g^{ij}b_j a_i = a_i b^i$$

are *valid for arbitrary vector components,* it follows that

$$g_{ij}b^j = b_i \qquad g^{ij}b_j = b^i.$$

That is,

> Contraction with the metric tensor may be used to raise or lower an index on a vector.

- Thus the *scalar product of two vectors* may be written in any of the following equivalent ways,

$$\boldsymbol{a}\cdot\boldsymbol{b} \equiv a^i b_i \equiv a_i b^i = g_{ij}a^i b^j = g^{ij}a_i b_j = g^i_j a_i b^j.$$

- From the preceding expressions

$$b^i = g^{ij} b_j = g^{ij} \underbrace{g_{jk} b^k}_{b_j} \qquad b^i = \delta^i_k b^k,$$

and since this is valid for arbitrary components b^i,

$$g^{ij} g_{jk} = g_{kj} g^{ji} = \delta^i_k.$$

- Viewing g^{ij} as the elements of a matrix G and g_{ij} as the elements of a matrix $\tilde{G}$, the equations

$$g^{ij} = g^{ji} \qquad g_{ij} = g_{ji}.$$

are equivalent to the matrix equations

$$G = G^{\mathrm{T}} \qquad \tilde{G} = \tilde{G}^{\mathrm{T}},$$

where T denotes the transpose. The Kronecker delta is just the 3×3 unit matrix I, implying that

$$g^{ij} g_{jk} = g_{kj} g^{ji} = \delta^i_k$$

may be written as the matrix equations

$$\tilde{G} G = G \tilde{G} = I.$$

The matrix corresponding to the covariant components of the metric tensor is the inverse of the matrix corresponding to the contravariant components of the metric tensor: *one may be obtained from the other by matrix inversion*.

Box 2.1 The metric tensor for 3-dimensional euclidean space

$$\text{Components:} \quad g_{ij} \equiv \boldsymbol{e}_i \cdot \boldsymbol{e}_j \qquad g^{ij} \equiv \boldsymbol{e}^i \cdot \boldsymbol{e}^j \qquad g^i_j \equiv \boldsymbol{e}^i \cdot \boldsymbol{e}_j = \delta^i_j$$

$$\text{Scalar product:} \quad \boldsymbol{A} \cdot \boldsymbol{B} = g_{ij} A^i B^j = g^{ij} A_i B_j = g^i_j A_i B^j = A^i B_i = A_i B^i$$

$$\text{Symmetry:} \quad g^{ij} = g^{ji} \qquad g_{ij} = g_{ji}$$

$$\text{Contractions:} \quad g_{ij} A^j = A_i \qquad g^{ij} A_j = A^i$$

$$\text{Orthogonality:} \quad g^{ij} g_{jk} = g_{kj} g^{ji} = \delta^i_k$$

$$\text{Matrix properties :} \quad \tilde{G} G = G \tilde{G} = I \qquad G \equiv [g^{ij}] \qquad \tilde{G} \equiv [g_{ij}]$$

Some basic properties of the metric tensor are summarized in the box above.

2.1.8 Line Elements and Distances

- Coordinates $u^1(t)$, $u^2(t)$, and $u^3(t)$ parameterized by t.
- As the parameter t varies, the points characterized by the specific values of the coordinates

$$u^1 = u^1(t) \qquad u^2 = u^2(t) \qquad u^3 = u^3(t)$$

  will trace out a curve in the three-dimensional space.
- The position vector for these points as a function of t is

$$\begin{aligned}\boldsymbol{r}(t) = x(u^1(t),u^2(t),u^3(t))\,\boldsymbol{i} + y(u^1(t),u^2(t),u^3(t))\,\boldsymbol{j} \\ + z(u^1(t),u^2(t),u^3(t))\,\boldsymbol{k},\end{aligned}$$

- By the chain rule

$$\begin{aligned}\frac{d\boldsymbol{r}}{dt} &= \frac{\partial \boldsymbol{r}}{\partial u^1}\frac{du^1}{dt} + \frac{\partial \boldsymbol{r}}{\partial u^2}\frac{du^2}{dt} + \frac{\partial \boldsymbol{r}}{\partial u^3}\frac{du^3}{dt} \\ \dot{\boldsymbol{r}} &= \dot{u}^1\boldsymbol{e}_1 + \dot{u}^2\boldsymbol{e}_2 + \dot{u}^3\boldsymbol{e}_3,\end{aligned}$$

  where the definitions

$$\dot{\boldsymbol{r}} \equiv \frac{d\boldsymbol{r}}{dt} \qquad \boldsymbol{e}_i \equiv \frac{\partial \boldsymbol{r}}{\partial u^i} \qquad \dot{u}^i \equiv \frac{du^i}{dt}$$

  have been used.

- In summation convention the equation

$$\dot{\boldsymbol{r}} = \dot{u}^1 \boldsymbol{e}_1 + \dot{u}^2 \boldsymbol{e}_2 + \dot{u}^3 \boldsymbol{e}_3,$$

  is $\dot{\boldsymbol{r}} = \dot{u}^i \boldsymbol{e}_i$.

- This may be expressed in differential form as $d\boldsymbol{r} = du^i \boldsymbol{e}_i$.

- Thus the squared infinitesimal distance along the curve is

$$\begin{aligned} ds^2 = d\boldsymbol{r} \cdot d\boldsymbol{r} &= du^i \boldsymbol{e}_i \cdot du^j \boldsymbol{e}_j \\ &= \boldsymbol{e}_i \cdot \boldsymbol{e}_j du^i du^j \\ &= g_{ij} du^i du^j, \end{aligned}$$

  where $g_{ij} \equiv \boldsymbol{e}_i \cdot \boldsymbol{e}_j$ has been used.

- Notice that in expressing the line element

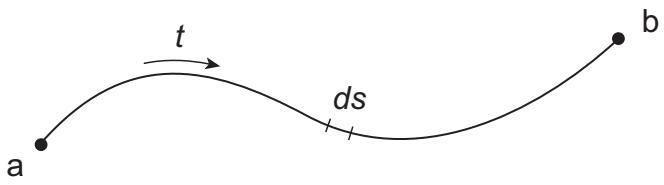
$$ds^2 = g_{ij}\,du^i du^j$$

we use the usual convention that $d\alpha^2 \equiv (d\alpha)^2$.

- That is, $d\alpha^2$ means the square of $d\alpha$, not the differential of α^2.
- Thus $ds^2 = g_{ij}\,du^i du^j$ is the *infinitesimal line element* for the space described by the metric g_{ij}.
- The length d of a finite segment between points a and b is obtained from the integral

$$d = \int_a^b ds = \int_a^b \left(g_{ij}\,du^i du^j\right)^{1/2} = \int_a^b \left(g_{ij}\frac{du^i}{dt}\frac{du^j}{dt}\right)^{1/2} dt,$$

where t parameterizes the position along the segment.



For example:

- The line element for two-dimensional euclidean space in cartesian coordinates (x, y) is given by

$$ds^2 = dx^2 + dy^2,$$

  which is just the Pythagorean theorem for right triangles having infinitesimal sides.

- The corresponding line element expressed in plane polar coordinates (r, φ) is then the familiar

$$ds^2 = dr^2 + r^2 d\varphi^2.$$

Example 2.4

For plane polar coordinates (r, φ) we have

$$x = r\cos\varphi \qquad y = r\sin\varphi,$$

so the position vector may be expressed as

$$\boldsymbol{r} = (r\cos\varphi)\,\boldsymbol{i} + (r\sin\varphi)\,\boldsymbol{j}.$$

Then the basis vectors in the natural basis are

$$\boldsymbol{e}_1 = \frac{\partial \boldsymbol{r}}{\partial r} = (\cos\varphi)\boldsymbol{i} + (\sin\varphi)\,\boldsymbol{j} \qquad \boldsymbol{e}_2 = \frac{\partial \boldsymbol{r}}{\partial \varphi} = -r(\sin\varphi)\,\boldsymbol{i} + r(\cos\varphi)\,\boldsymbol{j}.$$

The elements of the metric tensor then follow from $g_{ij} \equiv \boldsymbol{e}_i \cdot \boldsymbol{e}_j$:

$$g_{11} = \cos^2\varphi + \sin^2\varphi = 1 \qquad g_{22} = r^2(\cos^2\varphi + \sin^2\varphi) = r^2$$

and $g_{12} = g_{21} = 0$, or in matrix form

$$g_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix}.$$

Then the line element is

$$ds^2 = g_{ij}\,dx^i du^j = g_{11}(du^1)^2 + g_{22}(du^2)^2 = dr^2 + r^2 d\varphi^2,$$

where $u^1 = r$ and $u^2 = \varphi$.

This can be expressed as the matrix equation

$$\begin{aligned} ds^2 &= (dr \;\; d\varphi) \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix} \begin{pmatrix} dr \\ d\varphi \end{pmatrix} \\ &= (dr \;\; d\varphi) \begin{pmatrix} dr \\ r^2 d\varphi \end{pmatrix} = dr^2 + r^2 d\varphi^2. \end{aligned}$$

The line elements expressed in cartesian and polar coordinates in the preceding two examples

- Correspond to the *same space,* parameterized in terms of different coordinates.
- The form of the line element is different in the two parameterizations, but
- for any two nearby points the distance between them is given by ds, independent of the coordinate system.
- Thus, the line element ds is *invariant under coordinate transformations.*
- Since the distance between any two points that are not nearby can be obtained by integrating ds, we conclude that generally

> *The distance between any two points is invariant under coordinate transformations for metric spaces.*

The line element, which is specified in terms of the metric tensor, characterizes the geometry of the space because

- integrals of the line element define distances and
- angles can be defined in terms of ratios of distances.

Indeed, we could verify all the axioms of euclidean geometry starting from the line elements if we chose to do so.

2.2 Integration

Integration enters into physical theories in various ways, for example in the formulation of conservation laws.

- It is important to understand how the volume element for integrals behaves under change of coordinate systems.
- Trivial in euclidean space with orthonormal coordinates.
- Non-trivial in curved spaces, or even in flat spaces parameterized in non-cartesian coordinates.

We illustrate in flat 2D space with coordinates (x^1, x^2) and basis vectors $(\boldsymbol{e}_1, \boldsymbol{e}_2)$, assuming an angle θ between the basis vectors.

- As you are asked to demonstrate in a problem, the 2D volume (area) element is in this case

$$dA = \sqrt{\det g}\, dx^1 dx^2,$$

  where $\det g$ is the determinant of the metric tensor g_{ij}.
- For orthonormal coordinates g_{ij} is a unit matrix so $(\det g)^{1/2} = 1$.
- But in the general case the $(\det g)^{1/2}$ factor is not unity and its presence is essential to making integration invariant under change of coordinates.

As we will show later, this 2D example generalizes easily to define invariant integration in 4D spacetime.

2.3 Differentiation

Taking derivatives of vectors in spaces defined by *position-dependent metrics* will be crucial in general relativity.

- First consider the simpler case of taking the derivative of a vector in a euclidean space, but one *parameterized with a vector basis that may depend on the coordinates.*
- We may expand a vector $\boldsymbol{V}$ in a convenient basis $\boldsymbol{e}_i$,

$$\boldsymbol{V} = V^i \boldsymbol{e}_i.$$

- By the usual product rule, the partial derivative is given by a sum of two terms,

$$\frac{\partial \boldsymbol{V}}{\partial x^j} = \underbrace{\frac{\partial V^i}{\partial x^j}\boldsymbol{e}_i}_{\text{component}} + \underbrace{V^i \frac{\partial \boldsymbol{e}_i}{\partial x^j}}_{\text{basis}},$$

  1. The first term represents the *change in the component* V^i.
  2. The second term represents the *change in the basis vectors* $\boldsymbol{e}_i$.

- For the situation where we can choose a basis that is independent of coordinates, the second term is zero and we recover the expected formula.
- However, if the basis depends on the coordinates the second term will generally not be zero.

In the second term of

$$\frac{\partial \boldsymbol{V}}{\partial x^j} = \underbrace{\frac{\partial V^i}{\partial x^j}\boldsymbol{e}_i}_{\text{component}} + \underbrace{V^i \frac{\partial \boldsymbol{e}_i}{\partial x^j}}_{\text{basis}},$$

- The factor $\partial \boldsymbol{e}_i / \partial x^j$ resulting from the action of the derivative operator on the basis vectors is itself a vector and can be expanded in the vector basis,

$$\frac{\partial \boldsymbol{e}_i}{\partial x^j} = \Gamma^k_{ij} \boldsymbol{e}_k.$$

- The expansion coefficients Γ^k_{ij} may be interpreted as specifying the projection on the k axis of the rate of change in the j direction of a basis vector pointing in the i direction.

To give a concrete example, consider the 2D euclidean plane parameterized by polar $(r,\theta) = (x^1, x^2)$ coordinates.

- Taking basis vectors $(\boldsymbol{e}_r, \boldsymbol{e}_\theta)$, we have (for example)

$$\frac{\partial \boldsymbol{e}_r}{\partial \theta} = \Gamma^i_{r\theta} = \Gamma^r_{r\theta}\boldsymbol{e}_r + \Gamma^\theta_{r\theta}\boldsymbol{e}_\theta,$$

- Where the coefficient $\Gamma^r_{r\theta}$ is associated with the rate of change of the basis vector $\boldsymbol{e}_r$ with respect to θ in the direction $\boldsymbol{e}_r$
- The coefficient $\Gamma^\theta_{r\theta}$ is associated with the rate of change of the basis vector $\boldsymbol{e}_r$ with respect to θ in the direction $\boldsymbol{e}_\theta$.

- The coefficients Γ^k_{ij} are called *connection coefficients* or *Christoffel symbols.*
- Their generalization to 4-dimensional spacetime will be discussed more extensively later.
- There we shall see that the connection coefficients are central to
  1. The definition of derivatives
  2. A prescription for parallel transport of vectors in curved spacetime.

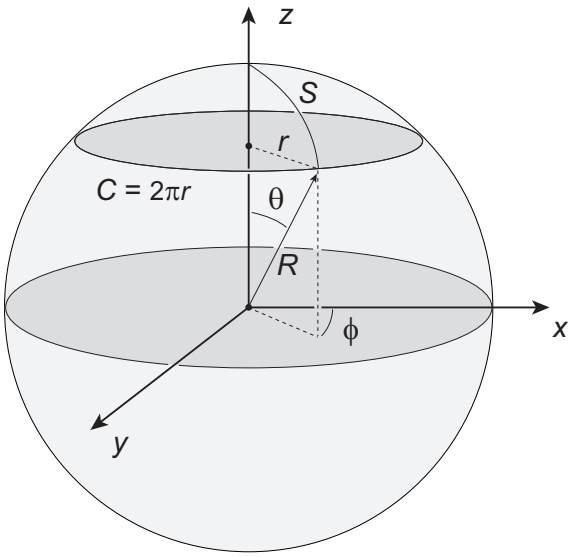


Figure 2.3: Measuring the circumference of a circle in curved space.

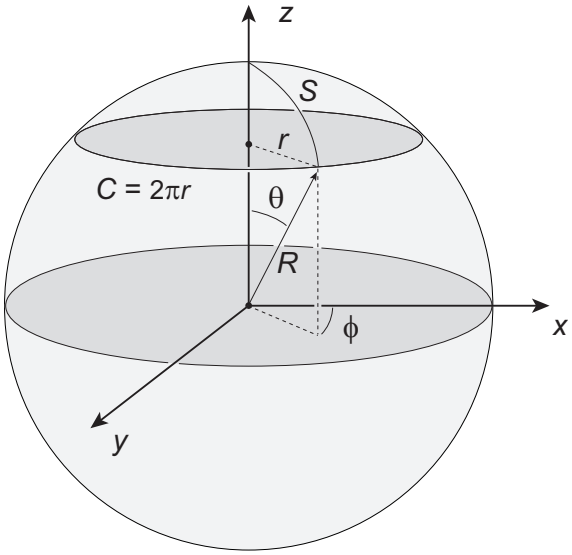
2.4 Non-Euclidean Geometry

Let us now consider *non-euclidean geometries.* A simple example is afforded by a sphere, as in Fig. 2.3.

- Imposing a standard polar coordinate system (θ, φ) on the surface of the sphere, the line element for a sphere is given by

$$ds^2 = R^2(d\theta^2 + \sin^2\theta d\varphi^2),$$

where R is the radius of the sphere.



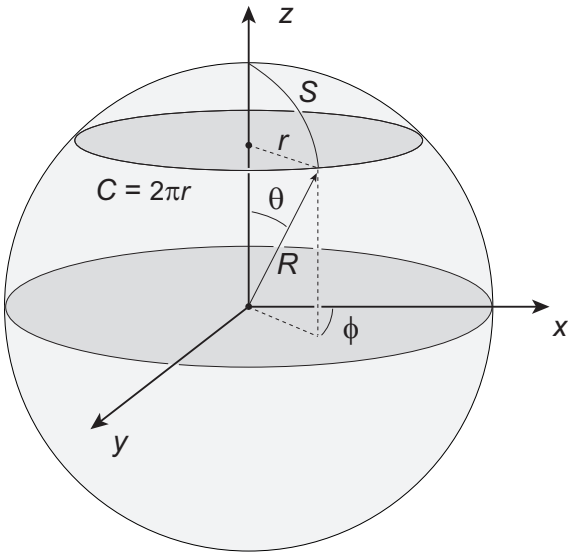
Let's calculate the ratio of the circumference of a circle to its radius for this non-euclidean space.

- We may define a circle in the two-dimensional space by marking a locus of points lying a constant distance S from a reference point (north pole in above figure).
- The θ angle subtended by S is S/R and $r = R\sin(S/R)$. Then from the geometry in the above figure, the circumference of the circle is

$$C = 2\pi r = 2\pi R \sin\frac{S}{R} = 2\pi S\left(1 - \frac{S^2}{6R^2} + \ldots\right).$$

- Alternatively, we may obtain the same result by integrating the line element $ds^2 = R^2(d\theta^2 + \sin^2\theta d\varphi^2)$,

$$C = \oint ds = \int_0^{2\pi} R\sin\frac{S}{R}d\varphi = 2\pi R\sin\frac{S}{R}.$$



- If the radius of the circle is much less than the radius of the sphere, the higher-order terms in the expansion of the sine may be ignored and we obtain the euclidean result $C \simeq 2\pi S$.
- But more generally the deviation of the circumference of small circles drawn on the sphere from $2\pi S$ is a measure of how much the sphere deviates from euclidean geometry.

Later, we will see how to use such considerations to define a quantitative measure of curvature for non-euclidean surfaces.

2.5 Transformations

It often proves useful to express physical quantities in more than one coordinate system.

- It therefore becomes necessary to understand how to transform between coordinate systems.
- This issue becomes particularly important in general relativity where it is essential to ensure that the laws of physics are not altered by the most general transformation between coordinate systems.

Let's consider two simple examples.

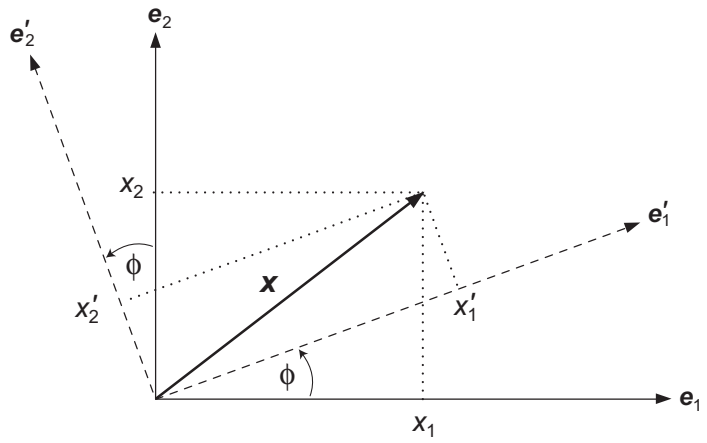


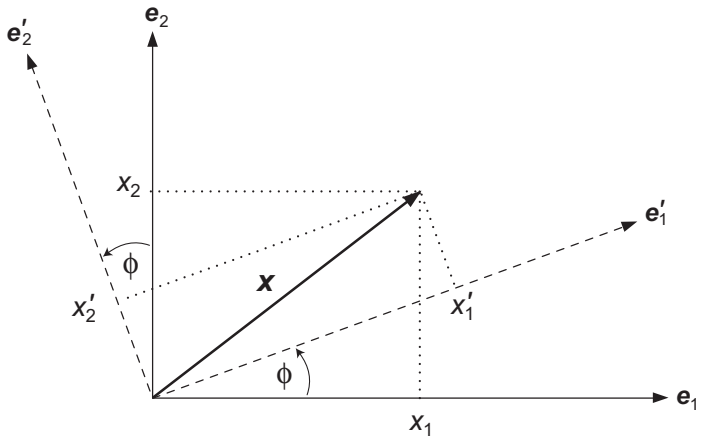
Figure 2.4: Rotation of coordinate system for a vector $\boldsymbol{x}$.

2.5.1 Rotational Symmetries

Consider the familiar example of the description of a vector under rotation of a coordinate system about the z axis by an angle φ, as illustrated in Fig. 2.4.

- In terms of the original basis vectors $\{\boldsymbol{e}_i\}$ the vector $\boldsymbol{x}$ has the components x_1 and x_2.
- After rotation of the coordinate system by the angle φ to give the new basis vectors $\{\boldsymbol{e}_i'\}$, the vector $\boldsymbol{x}$ has the components x_1' and x_2' in the new coordinate system.
- The vector $\boldsymbol{x}$ can be expanded in terms of the components for either of these bases:

$$\boldsymbol{x} = x^i \boldsymbol{e}_i = x'^i \boldsymbol{e}_i',$$



- We may use the geometry of the above figure to find that the components in the two bases are related by the transformation

$$\begin{pmatrix} x'^1 \\ x'^2 \\ x'^3 \end{pmatrix} = \begin{pmatrix} \cos\varphi & \sin\varphi & 0 \\ -\sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \\ x^3 \end{pmatrix},$$

  which may also be expressed as

$$x'^i = R^i_j x^j,$$

  where the R^i_j are the elements of the matrix in the preceding equation.

- This transformation law holds for any vector. (We may, in fact, *define* a vector in the x–y plane to be a quantity that obeys this transformation law.)

2.5.2 Galilean Transformations

Another simple example of a transformation is that between inertial frames in classical mechanics.

- Transformations between inertial frames with the same orientation are called *boosts*.
- In Newtonian physics time is considered an absolute quantity and boosts take the Galilean form

$$\boldsymbol{x}' = \boldsymbol{x}'(\boldsymbol{x},t) = \boldsymbol{x} - \boldsymbol{v}t \qquad t' = t'(\boldsymbol{x},t) = t.$$

- The Newtonian version of relativity asserts that the laws of physics are invariant under such Galilean transformations.
- Although the laws of mechanics at low velocity are invariant under Galilean transformations, *the laws of electromagnetism (Maxwell's equations) are not.*
- Indeed, the failure of Galilean invariance for the Maxwell equations was a large motivation in Einstein's eventual demonstration that *the laws of mechanics are not invariant with respect to Galilean transformations at high velocity.*

- As we shall discuss further later, in the absence of gravity the laws of both mechanics and electromagnetism are generally only invariant under Lorentz transformations.
- In the presence of a gravitational field, neither Galilean nor Lorentz invariance holds and we will be forced to seek a more general invariance to describe systems that are subject to gravitational forces.

Chapter 3

Tensors and Covariance

The term *covariance* implies a formalism in which *the laws of physics maintain the same form* under a specified set of transformations.

EXAMPLE: *Lorentz covariance* implies equations that are constructed in such a way that they *do not change their form under Lorentz transformations* (three boosts between inertial systems and three rotations).[a]

[a] An inertial system is a frame of reference in which Newton's first law of motion holds. Thus, for example, rotating frames and accelerated frames are not inertial. An inertial system is therefore in uniform translational motion with respect to any other inertial frame.

3.1 Spacetime Coordinates and Transformations

We will be concerned extensively with *spacetime,* which is an example of what mathematicians call a *manifold.*

- An *n-dimensional manifold* is
  - *a set* that can be parameterized continuously by
  - *n independent real coordinates* for each point (member of the set).
- We will assume the manifold to be differentiable at each point. Then we have a *differentiable manifold*.
- A *coordinate system*
  - associates n real parameters (labels) uniquely with each point of an n-dimensional manifold M
  - through a one-to-one mapping from $\mathbb{R}^n$ (cartesian product of n copies of the real numbers $\mathbb{R}$) to M.

> A *cartesian product* $X \times Y$ of two sets X and Y is the set of *all possible ordered pairs* (x, y) with x an element of X and y an element of Y.

- Generally, *more than one overlapping set of coordinates* is required to parameterize an entire manifold uniquely.
  - (See the discussion of *charts* and *atlases* in the book.)
  - For example, at least 2 overlapping sets of coordinates are required to parameterize a 2D sphere.

- Subsets of points within a manifold define
  - *curves* and
  - *surfaces*,

  which represent *submanifolds* of the full manifold.
- The manifolds that will interest us will be endowed with additional structure (in particular a geometry specified by a quadratic metric called *Riemannian geometry*).
- However, this will be sufficient definition for our initial purposes. We will get to Riemannian geometry and its central place in general relativity later.

We will sometimes use loose physics language and refer to manifolds simply as *spaces*.

Because relativity implies that space and time enter descriptions of nature on comparable footings,

- it will be useful to unify them into a 4-dimensional continuum termed *spacetime*.
- Spacetime is an example of a *differentiable manifold.*
- In spacetime points will be defined by *coordinates having four components*,
  - the first labeling the time multiplied by the speed of light c,
  - the other three labeling the spatial coordinates:

    $$x \equiv x^\mu = (x^0, x^1, x^2, x^3) = (ct, \boldsymbol{x}),$$

    where $\boldsymbol{x}$ denotes a vector with three components (x^1, x^2, x^3) labeling the spatial position.
- The first component x^0 is termed *timelike* and the last three components (x^1, x^2, x^3) are termed *spacelike.*
- As for the earlier discussion, the *placement of indices in upper or lower positions is meaningful*.
- Bold symbols will be used to denote (ordinary) vectors defined in the three spatial degrees of freedom,
- with 4-component vectors in spacetime denoted in non-bold symbols.
- For spacetime the modern convention is to number the indices beginning with zero rather than one.

The coordinate systems of interest will be assumed to be quite general, subject only to the requirement that

- they *assign a coordinate uniquely to every point* of spacetime, and that they be
- *differentiable to sufficient order* for the task at hand at every spacetime point.

3.2 Covariance and Tensor Notation

- We shall be concerned generically with a transformation between one set of spacetime coordinates, denoted by

$$x \equiv x^\mu = (x^0, x^1, x^2, x^3)$$

and a new set

$$x'^\mu = x'^\mu(x) \qquad \mu = 0,1,2,3$$

where $x = x^\mu$ denotes the original coordinates.

- This notation is an economical form of

$$x'^\mu = \xi^\mu(x^1, x^2, x^3, x^4) \qquad (\mu = 1,2,\ldots)$$

- where the single-valued, continuously differentiable functions ξ^μ
- assign a new (primed) coordinate (x'^1, x'^2, x'^3, x'^4) to a point of the manifold with old coordinates (x^1, x^2, x^3, x^4).

> This transformation may be abbreviated to $x'^\mu = \xi^\mu(x)$ and, even more tersely, to $x'^\mu = x'^\mu(x)$.

- Coordinates are just *labels,* so laws of physics cannot depend on them. Hence the system x'^μ is not privileged and this transformation should be *invertible.*
- Notice carefully that we are talking about the *same point* described in *two different coordinate systems.*

As introduced in Chapter 2, we shall generally use the Einstein summation convention for 4D spacetime:

- An index that is repeated, once as a superscript and once as a subscript, implies a summation over that index.
- Such an index is a dummy index that is removed by the summation and should not appear on the other side of the equation.
- A repeated (dummy) index may be replaced by any other index not already in use without altering equation: $A_\alpha B^\alpha = A_\beta B^\beta$.
- A superscript (subscript) in a denominator counts as a subscript (superscript) in a numerator.
- Greek indices $(\alpha, \beta, \ldots)$ denote the full set of spacetime indices running over 0, 1, 2, 3.
- Roman indices $(i, j, \ldots)$ denote the indices 1, 2, 3 running only over the spatial coordinates.
- Placement of indices matters: generally x^α and x_α will be different quantities.
- At all stages of manipulating equations, the indices on the two sides of an equation (including their up or down placement) must match.

As a minimum, we must consider the transformations of

- Fields
- Derivatives of fields
- Integrals of fields.

The first two are necessary to formulate *equations of motion,* and the latter enter into various *conservation laws.*

To facilitate this, we shall introduce a set of mathematical quantities called *tensors* that are a generalization of the idea of scalars and vectors to more components.

As a starting point, we must look more carefully at *how to define vectors in a curved space.*

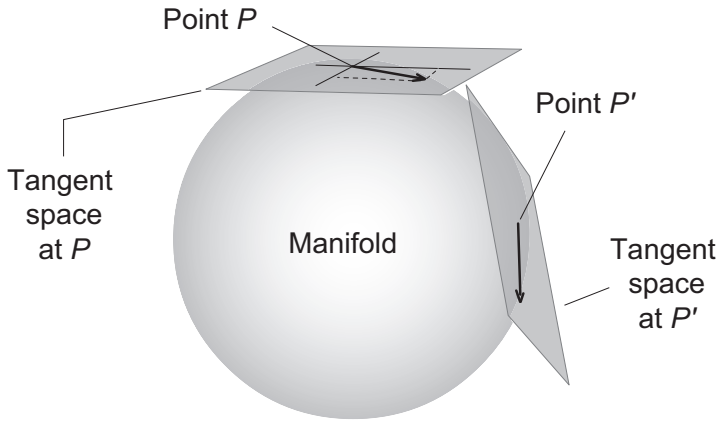
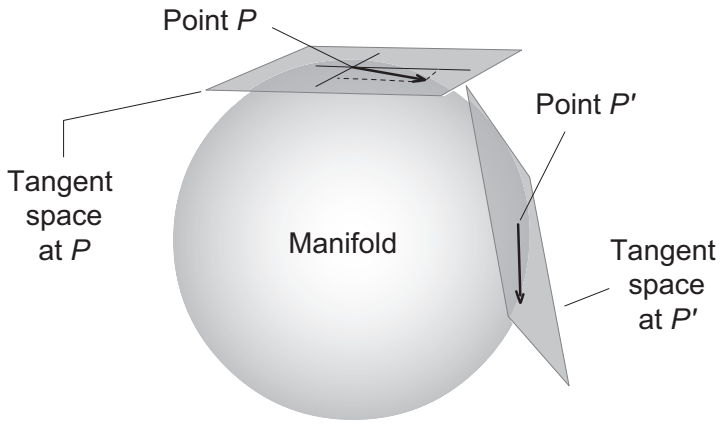


Figure 3.1: Tangent spaces and vectors in curved spaces, illustrated for the manifold S^2. Vectors (indicated by arrows) are defined in the tangent spaces at each point, not in the curved manifold. Embedding the 2D manifold in 3D euclidean space is for visualization purposes only; the tangent space has a specification that is intrinsic to the 2D manifold.

Spacetime is characterized by a manifold that is *not euclidean*.

- In euclidean space we are used to representing vectors as directed line segments of finite length.
- This picture won't do in curved spacetime, which is locally but not globally euclidean, so extended straight lines have no meaning.
- Thus in non-euclidean manifolds the first question that we need to address is how to define a vector at some spacetime point.
- *Answer:* vectors are not defined in the curved manifold itself but rather in a *tangent space* that may be visualized for a 2D manifold as a plane tangent to the point on the curved surface, as illustrated in Fig. 3.1 for a 2D sphere.



- The idea conveyed by the above figure in which planes tangent to a 2D surface are shown embedded in a 3D space is *useful conceptually* but it is *potentially misleading*.
- Defining the tangent space at each point is an *intrinsic process* with respect to a manifold and does not require embedding it in a higher-dimensional manifold, as will be shown later.

3.3 Tangent and Cotangent Bundles

An n-dimensional Riemannian manifold has at each point P

- An n-dimensional euclidean vector space T_P with a basis defined by the directional derivatives evaluated at P for coordinate curves passing through P.
- This is termed the *tangent space* and *vectors* at P are defined within that space.
- An intrinsically-defined n-dimensional euclidean vector space with a basis defined by viewing the coordinate curves as scalar fields and evaluating their gradients at P.
- This is termed the *cotangent space* T_P^* and *dual vectors* P are defined within this space.
- The tangent space T_P and the cotangent space T_P^* are *dual to each other*.

The definitions given above make clear that

- Vectors and dual vectors are *local to a point*.
- The tangent and cotangent spaces in which they are defined may be *constructed from the properties of the manifold alone*.

Thus the tangent space and cotangent space at each point of a manifold have an *intrinsic meaning*, independent of embedding in higher dimensions.

- The *tangent bundle* TM of a manifold M is a manifold consisting of all the tangent vectors defined in M, which is given by the disjoint union of all of its tangent spaces T_P.
- Likewise, a *cotangent bundle* T^*M of the manifold M is defined by the disjoint union of all the cotangent spaces T_P^* in M.
- Tangent or cotangent bundles are examples of a *fiber bundle,* which is a manifold E that is *locally* the *cartesian product* $E = F \times B$ of two spaces,
  - the *base space* B, and
  - the *fiber space* F (with the fiber at P corresponding to the tangent space T_P at P),

  but that *globally* may have a different topological structure.
- For a manifold of dimension n the tangent and cotangent fiber bundles are of dimension $2n$.

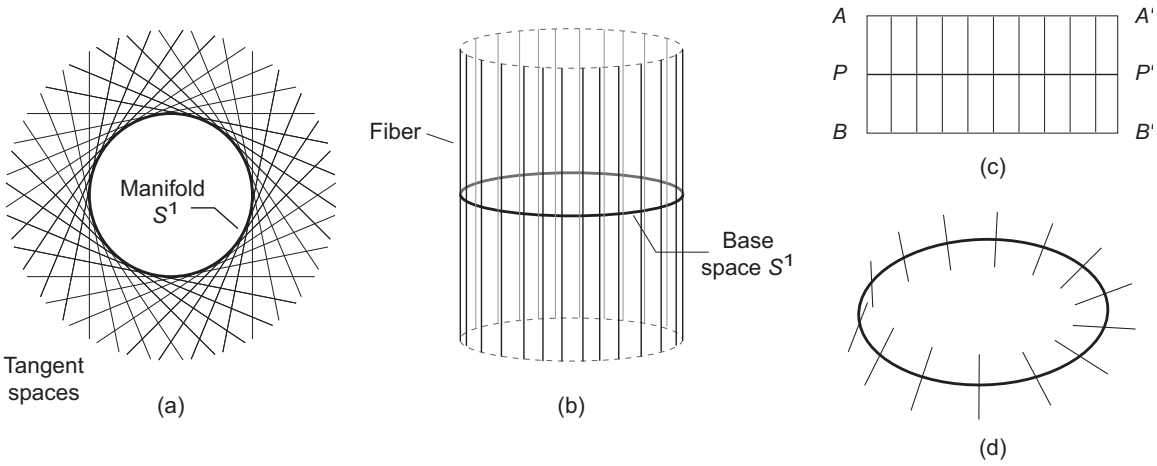


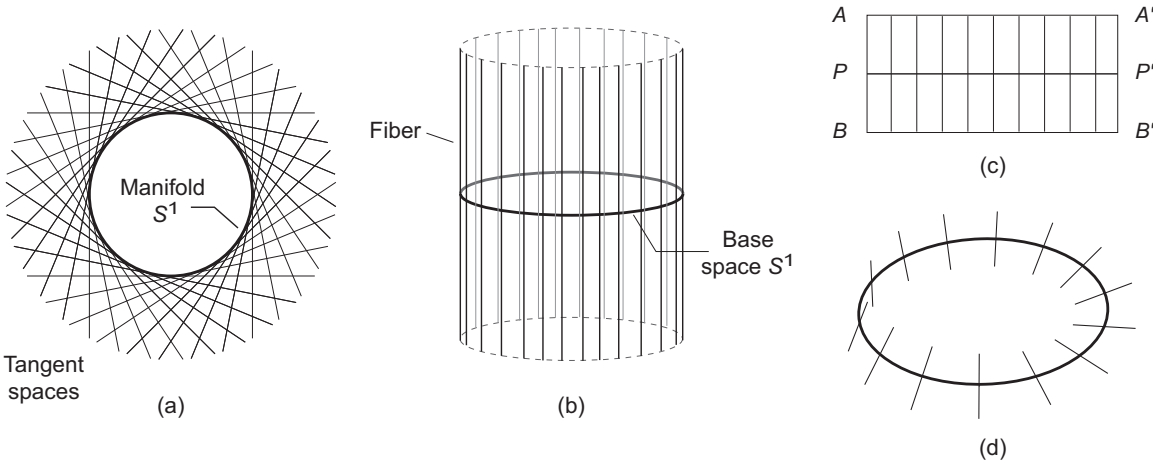
Figure 3.2: Tangent bundle TS^1 for the 1-dimensional manifold S^1.

Fig. 3.2 illustrates the basic idea of a *tangent bundle* for a manifold corresponding to a circle.

- (a) The *manifold* $M = S^1$ (the circle) and some of its *tangent spaces* (lines tangent to the circle).
- (b) The corresponding *tangent bundle* (locally and globally $R^1 \times S^1$).
- (c) Figure (b) cut vertically and rolled out flat. Figure (b) corresponds to identifying $A \leftrightarrow A'$ and $B \leftrightarrow B'$.
- (d) Tangent bundle with nontrivial topology *(Möbius band)* generated by identifying

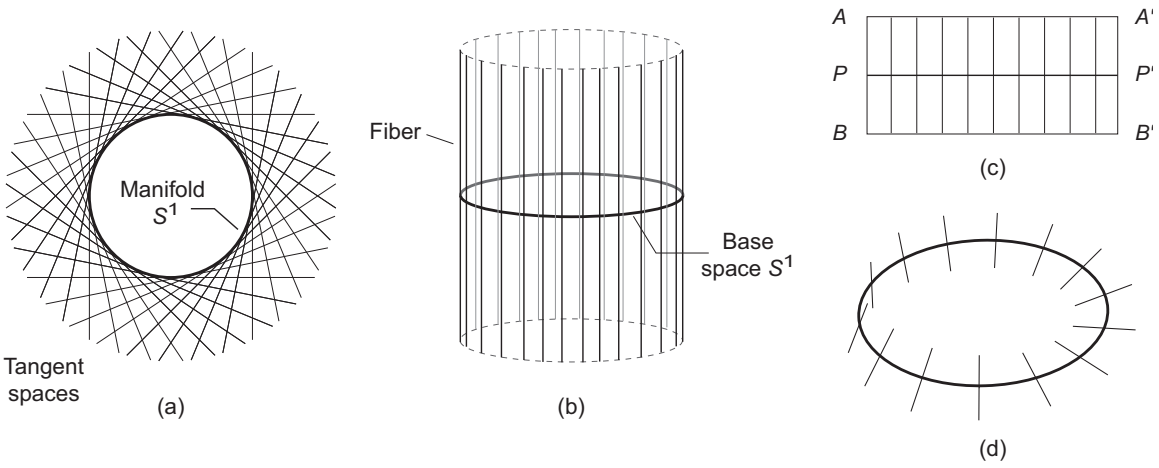
$$A \leftrightarrow B' \qquad B \leftrightarrow A'$$

  in (c).
- This is locally $R^1 \times S^1$, but not globally.



The lines corresponding to the tangent spaces (*fibers*) should be imagined extending to infinity to accommodate vectors of arbitrary length.

- The overlaps of these lines in (a) are meaningless since each tangent space is defined independently at a different point of the manifold.
- Therefore, in (b) the lines (fibers) corresponding to tangent spaces have been rotated and arrayed perpendicular to the base manifold so that they do not overlap.
- A location y on a fiber may be interpreted as existing at the point $P = x$ where the fiber intersects the base space, with y (vector length) given by the distance to the intersection of the fiber with the base space.
- Thus (x, y) identifies a point in the fiber bundle manifold uniquely.



- Cases (b) and (d) are *equivalent locally but distinct topologically:* the orientation of the fiber winds through π for once around the base space in (d), so that (d) cannot be deformed continuously into (b).

3.4 Coordinates in Spacetime

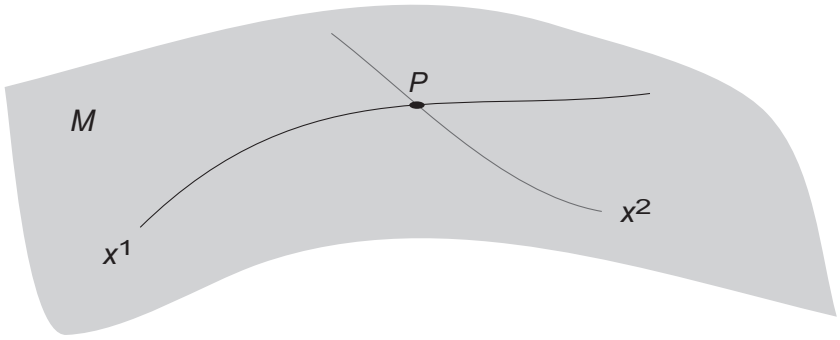
A universal coordinate system can be chosen in flat space,

- Basis vectors can be chosen that are mutually *orthogonal and constant*.
- Furthermore, these constant basis vectors can be *normalized to unit length* once and for all.
- Much of ordinary physics is conveniently described using such *orthonormal bases*.
- The situation is more complicated in curved manifolds (or in uncurved manifolds expressed in non-cartesian coordinates).
- Because of the position-dependent metric of curved spacetime it is most convenient in general relativity to choose basis vectors that
  - *depend on position* and that
  - *need not be orthogonal*.
- Since such basis vectors are position-dependent, it usually is *not useful to normalize them*.

3.4.1 Coordinate and Non-coordinate Bases

The standard conception of a vector as a directed line segment has ill-defined meaning in a curved manifold.

- The key to specifying vectors in curved space is to *separate the "directed" part from the "line segment" part* of the usual definition.
- This is because the *direction* for vectors of infinitesimal length can be defined consistently in curved or flat spaces using *directional derivatives.*



Consider a curve in a differentiable manifold along which one of the coordinates x^μ varies while all others $x^\nu (\nu \neq \mu)$ are held constant.

- This curve will be termed *the coordinate curve* x^μ.
- The figure above illustrates for a 2D manifold.
- Through any point P in spacetime 4 such curves will pass, corresponding to the coordinate curves x^μ with $\mu = (0,1,2,3)$.
- A convenient set of *position-dependent basis vectors* $e_\mu (\mu = 0,1,2,3)$ can be defined *at each point* P in the manifold by

$$e_\mu = \lim_{\delta x^\mu \to 0} \frac{\delta s}{\delta x^\mu},$$

where δs is the infinitesimal distance along the coordinate curve x^μ between the point P with coordinate x^μ and a nearby point with coordinate $x^\mu + \delta x^\mu$.

- For a parameterized curve $x^\mu(\lambda)$ having a tangent vector t with components $t^\mu = dx^\mu/d\lambda$ *(summation convention)*,

$$t = t^\mu e_\mu = \frac{dx^\mu}{d\lambda} e_\mu,$$

the *directional derivative* of an arbitrary scalar function $f(x^\mu)$ that is defined in the neighborhood of the curve is

$$\begin{aligned}\frac{df}{d\lambda} &\equiv \operatorname*{Lim}_{\varepsilon\to 0}\left[\frac{f(x^\mu(\lambda+\varepsilon)) - f(x^\mu(\lambda))}{\varepsilon}\right] \\ &= \frac{dx^\mu}{d\lambda}\frac{\partial f}{\partial x^\mu} = t^\mu \frac{\partial f}{\partial x^\mu},\end{aligned}$$

- Since $f(x)$ is arbitrary this implies the *operator relation*

$$\frac{d}{d\lambda} = \frac{dx^\mu}{d\lambda}\frac{\partial}{\partial x^\mu} = t^\mu \frac{\partial}{\partial x^\mu}.$$

- Hence we find that
  - the tangent components t^μ are associated with a unique directional derivative and
  - the *partial derivative operators* $\partial/\partial x^\mu$ *define the basis vectors* e_μ,

$$e_\mu = \frac{\partial}{\partial x^\mu} \equiv \partial_\mu.$$

- This permits an arbitrary vector to be expanded as

$$V = V^\mu e_\mu = V^\mu \frac{\partial}{\partial x^\mu} = V^\mu \partial_\mu.$$

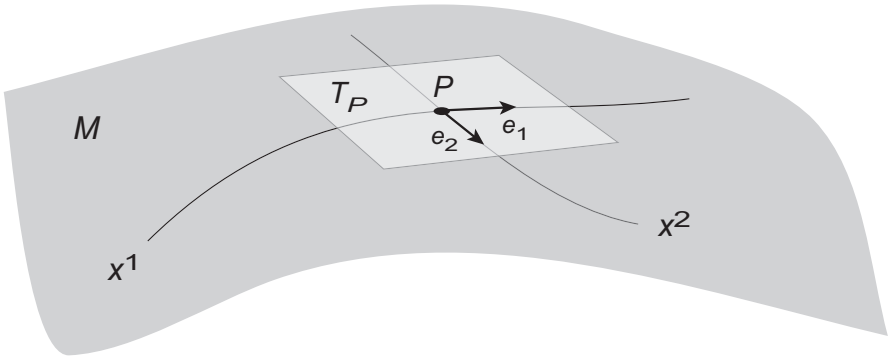


Figure 3.3: Tangent space T_P at a point P for a curved 2D manifold M. The vectors tangent to the coordinate curves at each point define a coordinate or holonomic basis. This figure is a generalization of Fig. 3.1 to an arbitrary curved 2D manifold with a position-dependent, non-orthogonal (coordinate) basis. This embedding of M in 3D euclidean space is for visualization purposes only; the basis vectors e_1 and e_2 of the tangent space are specified by directional derivatives of the coordinate curves evaluated entirely in M at the point P.

- Position-dependent basis vectors (that generally are neither orthogonal nor normalized) define a *coordinate basis* or *holonomic basis.*
- A basis using orthonormal coordinates is then termed a *non-coordinate basis* or an *anholonomic basis.*
- A coordinate basis is illustrated schematically in Fig. 3.3 for a generic curved 2D manifold.
- The definition of a vector in terms of directional derivatives is *valid in any curved or flat differentiable manifold.*
- It replaces the standard idea of a vector as the analog of a displacement vector between two points, which *does not generalize to curved manifolds*.

The separation between nearby points is

$$ds = e_\mu(x)dx^\mu,$$

from which

$$ds^2 = ds \cdot ds = (e_\mu \cdot e_\nu)dx^\mu dx^\nu = g_{\mu\nu}dx^\mu dx^\nu$$

with the *metric tensor components* $g_{\mu\nu}$ defined by,

$$e_\mu(x) \cdot e_\nu(x) \equiv g_{\mu\nu}(x),$$

which implies that in a coordinate basis the scalar product of vectors A and B is given by

$$A \cdot B = (A^\mu e_\mu) \cdot (B^\nu e_\nu) = g_{\mu\nu}A^\mu B^\nu.$$

These equations may be taken as a definition of a vector coordinate basis $\{e_\mu\}$.

The preceding discussion has been specifically for vectors and involves defining a basis for the tangent space T_P at each point P using the tangents $\partial/\partial x^\mu$ to coordinate curves passing through P.

- By analogy, a similar intrinsic procedure can be invoked to construct a basis for dual vectors in the cotangent space T_P^* at a point P using gradients to define basis vectors.
- This leads to equations analogous to those for the tangent space, but with the indices of the basis vectors in the upper position.
- A set of dual basis vectors e^μ may be used to expand dual vectors ω as
$$\omega = \omega_\mu e^\mu,$$
- This allows the metric tensor with upper indices to be defined through
$$e^\mu(x) \cdot e^\nu(x) \equiv g^{\mu\nu}(x),$$
with the scalar product of arbitrary dual vectors α and β given by
$$\alpha \cdot \beta = g^{\mu\nu}\alpha_\mu\beta_\nu.$$

These equations may be taken as a definition of a dual-vector coordinate basis $\{e^\mu\}$.

An *orthonormalized non-coordinate basis* is specified by requiring

$$e_{\hat{\mu}}(x) \cdot e_{\hat{\nu}}(x) = \eta_{\hat{\mu}\hat{\nu}},$$

where (See Ch. 4)

$$\eta_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \text{diag}\,(-1,1,1,1)$$

is the metric tensor of (flat) Minkowski spacetime.

- A common notational convention has been employed of using hats on indices to indicate explicitly that this is an *orthonormal and not coordinate basis.*
- Also, it is common to use η to denote specifically the metric for Minkowski space rather than the more general g.
- As elaborated further in the Problems,
  - the basis vectors of a coordinate basis have a vanishing *Lie bracket,* $[e_\mu, e_\nu] = 0$, while
  - for a non-coordinate basis $[e_{\hat{\mu}}, e_{\hat{\nu}}] \neq 0$,

  where the Lie bracket of two vector fields A and B is defined by the *commutator*

$$[A, B] \equiv AB - BA.$$

- This provides a formal way to identify a coordinate basis.

- The formulation of general relativity is most natural in a coordinate (holonomic) basis.
- However, we will see later that curved spacetime is *locally euclidean,* and that
- observers in a laboratory of small extent in spacetime can define a local coordinate system in a non-coordinate basis for interpreting measurements.
- Thus it is natural to
  - formulate general relativity in a coordinate basis, but
  - to use a non-coordinate basis for interpretation of some measurements.

We won't often need to display a basis explicitly in applications, but unless stated otherwise *a coordinate basis will be assumed.*

3.5 Tensors and Coordinate Transformations

In formulating general relativity we are interested in how quantities that enter the physical description of the Universe change when the spacetime coordinates are transformed.

- This requires understanding the transformations of
  - *fields*,
  - their *derivatives*, and
  - their *integrals*,
- To facilitate this task, it is useful to introduce a set of mathematical objects called *tensors.*
  - These have a *fundamental definition without reference to specific coordinate systems*.
  - However, for physical applications it often proves convenient to view tensors as *components expressed in a basis that transform in a precise way* if the coordinate system is changed.
- We shall develop and use both views of tensors.

- Let's note that our interest here almost always will be in *tensor fields,* which correspond to tensors of a given type defined at every point of the manifold.
- Since this is a rather trivial generalization of a tensor defined at a point, the discussion will for brevity often use shorthand like "vector" or "tensor" to mean "vector field" or "tensor field", respectively.
- This is unlikely to engender confusion, since the meaning should be clear from the context.

The *rank* of a tensor will be given a more fundamental definition below, but practically it is the *total number of indices* required to specify its components in some basis.

- Thus scalars are tensors of rank zero and vectors or dual vectors are tensors of rank one.
- This may be generalized to tensors carrying more than one index.
- As for vectors and dual vectors, the indices may either be upper (*contravariant*) or lower (*covariant*).
  - Tensors carrying only lower indices are termed *covariant tensors*.
  - Tensors carrying only upper indices are termed *contravariant tensors*.
  - Tensors carrying both lower and upper indices are termed *mixed tensors*.
- It is convenient to indicate the *type* of a tensor by the ordered pair (p,q), where when evaluated in a basis
  - p is the number of contravariant (upper) indices,
  - q is the number of covariant (lower) indices,

  and the rank of the tensor is $p+q$.

  Thus a dual vector is a rank-1 tensor of type $(0,1)$ having one covariant index.

Not all quantities with indices are tensor components; it is their *mathematical properties* that mark objects as tensors, not merely that they carry indices when evaluated in a basis.

There are two general views that we might take of tensors:

- Mathematicians prefer to define tensors in an elegant and abstract manner that often is more precise in defining some essential underlying mathematical concepts. This is sometimes called the *index-free formalism*.
- The characterization of tensors in terms of their *transformation properties* is particularly useful from a physical perspective.

We shall first discuss tensors from the mathematician's perspective, and then discuss them in terms of their transformation properties.

3.6 Tensors as Linear Maps to Real Numbers

From a fundamental perspective, a tensor of type (n,m)

- has input slots for n *vectors* and m *dual vectors*, and
- *acts linearly* on these inputs to *give a real number*.

If ω is a $(0,1)$ tensor (dual vector) and A and B are $(1,0)$ tensors (vectors), *linearity of the mapping* implies things like

$$\omega(aA+bB) = a\omega(A)+b\omega(B) \in \mathbb{R},$$

where a and b are arbitrary scalars and $\mathbb{R}$ denotes the set of real numbers.

- This definition makes *no reference to components* of the vectors or dual vectors.
- Hence the tensor map must give the same real number, *irrespective of any choice of coordinate system*.

In summary, *a tensor is*

- a *function of vectors and dual vectors themselves,* rather than of their components, or
- an *operator* that accepts vectors and dual vectors as input and produces a real number.

As a warmup exercise, consider a real-valued function of the coordinates $f(x)$.

- This function
  - takes no vectors or dual vectors as input and
  - yields a real number (the value of the function at x) as output.
- Thus it is a *tensor of rank zero* (a scalar).

Let's now give a few less-trivial examples of how this approach to tensors works, beginning with vectors and dual vectors.

Vector Space:

A *vector space* has a precise axiomatic definition but for our purposes it will be sufficient to view it more loosely as

- A set of objects (the vectors) that can be
  - multiplied by real numbers and
  - added in a linear way

  while exhibiting *closure:* any such operations on elements of the set give back a linear combination of elements.
- For arbitrary vectors A and B, and arbitrary scalars a and b, one expects then that expressions like

$$(a+b)(A+B) = aA + aB + bA + bB$$

  should be satisfied.
- A *basis* for a vector space is a set of vectors that
  - *span the space* (any vector is a linear combination of basis vectors) and
  - are *linearly independent* (no basis vector is a linear combination of other basis vectors).
- The number of basis vectors is the *dimension* of the space.

For spacetime, vector and dual vector spaces are *vector spaces of dimension 4* that are defined *at each point* of the manifold.

Confusion Alert

Be aware that "vector" is being used in our discussion in three senses:

- As an arbitrary element of an abstract vector space, according to the definition given above.
- As an element of an abstract vector space that *also* is a vector in the precise sense defined above [a tensor of type $(1, 0)$].
- Sometimes as a generic term for an element of an abstract vector space that *is either* a vector or a dual vector.

Which meaning is intended is usually clear from the context.

3.6.1 Vectors and Dual Vectors

Suppose a *vector field* to be defined on a manifold such that

- Each point P has associated with it a vector V that may be expanded in a (coordinate) vector basis e_μ,

  $$V = V^\mu e_\mu,$$

  where the basis vectors e_μ are defined in the tangent space T_P at each point of the manifold.

- There is a corresponding dual vector field ω defined at each point P that may be expanded in a (coordinate) dual-vector basis e^μ,

  $$\omega = \omega_\mu e^\mu,$$

  where the basis dual vectors e^μ are defined in the cotangent space T_P^* at each point of the manifold.

- Hence the e_μ are *basis vectors in the tangent bundle* of the manifold and

- the e^μ are *basis vectors in the cotangent bundle* of the manifold.

Duality: As introduced in Chapter 2 for euclidean spaces, the vector spaces for V and ω are said to be *dual:*

- The space of vectors (*tangent bundle*): *all linear maps of dual vectors to the real numbers*.
- The space of dual vectors (*cotangent bundle*): *all linear maps of vectors to the real numbers*.
- This *duality* of vector and dual vector spaces can be implemented systematically by requiring that

$$e^\mu(e_\nu) \equiv e^\mu \cdot e_\nu = \delta^\mu_\nu,$$

where the *Kronecker delta* is given by

$$\delta^\mu_\nu = \begin{cases} 1 & \mu = \nu \\ 0 & \mu \neq \nu \end{cases}.$$

- Alternative notation: $A(B) \leftrightarrow \langle A, B\rangle$, so we can also write

$$\langle e^\mu, e_\nu\rangle \equiv e^\mu \cdot e_\nu = \delta^\mu_\nu.$$

- *Example: A dual vector ω acts on a vector V* as

$$\begin{aligned} \omega(V) = \langle \omega, V\rangle &= \omega_\mu e^\mu(V^\nu e_\nu) && \textit{(expand in basis)} \\ &= \omega_\mu V^\nu e^\mu(e_\nu) && \textit{(linearity)} \\ &= \omega_\mu V^\nu \delta^\mu_\nu && \textit{(duality: } e^\mu(e_\nu) = \delta^\mu_\nu\textit{)} \\ &= \omega_\mu V^\mu \in \mathbb{R} && \textit{(scalar product} \in \mathbb{R}\textit{)}, \end{aligned}$$

where $\mathbb{R}$ denotes the real numbers.

Note: Basis vectors are defined in the tangent bundle and basis dual vectors are defined in the cotangent bundle of the manifold.

- Thus for applications in spacetime our concern is really with the action of *vector fields on dual vector fields* and vice versa.
- Thus what is returned is not a real number but rather a *scalar field of real numbers* defined over the manifold.
- This is just another example where for simplicity we have been careless about speaking of some thing when what is really meant is a field of those things.

> We trust that the reader is sophisticated enough by now to realize when *"thing"* really means *"field of things"*.

The preceding discussion illustrates clearly that

- A *dual vector* is an operator that accepts a vector as an argument and produces a real number: the *scalar product* $\omega_\mu V^\mu$, which is
  - unique and
  - independent of basis

  as output.
- A *vector* is an operator that accepts a dual vector as an argument and produces a real number equal to the scalar product, $V(\omega) = \omega_\mu V^\mu$, that is unique and independent of basis.
- These definitions involve no uncontracted indices, so the results are *independent of any basis choice*.

This suggests that vectors and dual vectors may be defined fundamentally in terms of *linear maps to the real numbers,* with no reference to a specific basis:

1. A *dual vector* is an operator that acts linearly on a vector to return a real number.
2. A *vector* is as an operator that acts linearly on a dual vector to return a real number.

For those conversant with linear algebra this may sound familiar, as suggested by the following example.

In the language of *linear algebra*,

- vectors may be represented as column vectors and
- dual vectors as row vectors,

and their matrix product is a number.

- For example,

$$A \equiv (a \;\; b) \quad B \equiv \begin{pmatrix} c \\ d \end{pmatrix}$$

$$AB = (a \;\; b) \begin{pmatrix} c \\ d \end{pmatrix} = ac + bd \in \mathbb{R}$$

  may be regarded as the dual vector A acting linearly on the vector B to produce the real number $ac + bd$:

$$A(B) \in \mathbb{R}.$$

- For readers familiar with Dirac notation for matrix elements in quantum mechanics,
  - a *ket* $|a\rangle$ is a *vector* and
  - a *bra* $\langle a|$ is a *dual vector*

  in the quantum linear vector space called *Hilbert space*.
- Mathematically the *vector space of bras is the dual of the vector space of kets*. Thus the overlap $\langle f|\, i\rangle$ is a number (a *c-number* in quantum lingo).

Components in a Basis: These definitions of vectors and dual vectors are *independent of any choice of basis* but practically *it often is convenient to work in a basis*.

- The components of a basis may be constructed from

$$V^\mu = V(e^\mu) = e^\mu \cdot V \qquad \omega_\mu = \omega(e_\mu) = e_\mu \cdot \omega,$$

which may be interpreted (for example) as

> A vector accepts a basis vector e^μ as input and acts linearly on it to return a real number that is the *component of the vector* evaluated in that basis.

- The validity of the equations above is easily checked:

$$e^\mu \cdot V = e^\mu \cdot (V^\alpha e_\alpha) = V^\alpha e^\mu(e_\alpha) = V^\alpha \delta^\mu_\alpha = V^\mu,$$
$$e_\mu \cdot \omega = e_\mu \cdot (\omega_\alpha e^\alpha) = \omega_\alpha e_\mu(e^\alpha) = \omega_\alpha \delta^\alpha_\mu = \omega_\mu,$$

where we have used

  - the basis expansions,

$$V = V^\alpha e_\alpha \qquad \omega = \omega_\alpha e^\alpha.$$

  - linearity,

$$\omega(aA + bB) = a\omega(A) + b\omega(B) \in \mathbb{R},$$

  - and duality,

$$e^\mu(e_\nu) = e^\mu \cdot e_\nu = \delta^\mu_\nu.$$

Transformations: Consider a coordinate transformation $x^\mu \to x'^\mu$ on a dual vector $\omega = \omega_\mu e^\mu$ and on a vector $V = V^\mu e_\mu$.

- Requiring that the distance ds be invariant under coordinate transformation means that *the basis vectors* e^μ *and* e_μ *must transform as* (see Problems)

$$e^\mu \to e'^\mu = \frac{\partial x'^\mu}{\partial x^\nu} e^\nu \qquad e_\mu \to e'_\mu = \frac{\partial x^\nu}{\partial x'^\mu} e_\nu .$$

- How then do the components ω_μ and V^μ transform?
- Consider the dual vector ω, which is a *geometrical object* existing independent of representation in a coordinate system, so it must be *invariant under change of coordinates.*
- This requires the components of ω to transform as

$$\omega_\nu \to \omega'_\nu = \frac{\partial x^\alpha}{\partial x'^\nu} \omega_\alpha ,$$

since then ω is invariant under $x^\mu \to x'^\mu$:

$$\begin{aligned}
\omega' =& \omega'_\mu e'^\mu \qquad \text{(expand in basis)} \\
=& \frac{\partial x^\alpha}{\partial x'^\mu} \omega_\alpha \frac{\partial x'^\mu}{\partial x^\nu} e^\nu \qquad \text{(transform basis and components)} \\
=& \frac{\partial x^\alpha}{\partial x'^\mu} \frac{\partial x'^\mu}{\partial x^\nu} \omega_\alpha e^\nu = \frac{\partial x^\alpha}{\partial x^\nu} \omega_\alpha e^\nu \qquad \text{(linearity + chain rule)} \\
=& \omega_\alpha e^\nu \delta^\alpha_\nu \qquad \text{(apply } \partial x^\alpha / \partial x^\nu \equiv \delta^\alpha_\nu \text{)} \\
=& \omega_\alpha e^\alpha = \omega \qquad \text{(scalar product} \in \mathbb{R}\text{)}.
\end{aligned}$$

This transformation law for the components ω_ν is the same one that will be used later to *define* a dual vector.

We have just shown that the definition of a dual vector as a geometrical object that maps vectors linearly to the real numbers

- *is independent of representation* in any particular basis, but
- *in a particular basis* the components of dual vectors must transform as
$$\omega_\nu \to \omega'_\nu = \frac{\partial x^\alpha}{\partial x'^\nu}\,\omega_\alpha .$$
- By a similar proof, vector components V^μ may be shown to have the transformation law
$$V^\nu \to V'^\mu = \frac{\partial x'^\mu}{\partial x^\nu}\,V^\nu .$$
- In the index-free picture currently under discussion
  - tensors are *defined as linear maps of vectors and dual vectors to the real numbers*, and
  - the transformation laws and associated tensor algebra for representation in a basis *follow from that definition*.

Later, we will see that we can turn things around and use such transformation laws as a *definition* of tensors.

3.6.2 Tensors of Higher Rank

Higher-rank tensors may be constructed by taking *tensor products* (denoted by $\otimes$; also termed the *direct product* or the *Kronecker product*) of lower-rank tensors.

> The *tensor product* of two vector spaces $\mathbb{U}$ and $\mathbb{V}$ produces a new vector space $\mathbb{U} \otimes \mathbb{V}$.
>
> - If $\mathbb{U}$ has a basis $\{u_1, u_2, \ldots\}$ and
> - $\mathbb{V}$ has a basis $\{v_1, v_2, \ldots\}$,
>
> then $\mathbb{U} \otimes \mathbb{V}$ is spanned by a basis consisting of all pairs (u_i, v_j).

Schematically, a mixed tensor T of rank (p,q) may be expressed as

$$T = T^{\mu_1 \mu_2 \cdots \mu_p}{}_{\nu_1 \nu_2 \cdots \nu_q} e_{\mu_1} \otimes e_{\mu_2} \otimes \cdots \otimes e_{\mu_p} \otimes e^{\nu_1} \otimes e^{\nu_2} \otimes \cdots \otimes e^{\nu_q},$$

where $\{e_\mu\}$ is a vector basis and $\{e^\nu\}$ is a dual vector basis.

Components of higher-rank tensors may be evaluated by *putting basis vectors in its input slots*.

- For example, consider a *rank-2 tensor* T. Its covariant, contravariant, and mixed components are given by

$$T(e_\mu, e_\nu) = T_{\mu\nu} \qquad T(e^\mu, e^\nu) = T^{\mu\nu}$$
$$T(e_\mu, e^\nu) = T_\mu{}^\nu \qquad T(e^\mu, e_\nu) = T^\mu{}_\nu,$$

  from which it follows that for vectors A and B,

$$\begin{aligned} T(A,B) &= T(A^\mu e_\mu, B^\nu e_\nu) \\ &= \underbrace{T(e_\mu, e_\nu)}_{T_{\mu\nu}} A^\mu B^\nu = T_{\mu\nu} A^\mu B^\nu. \end{aligned}$$

  Likewise for contravariant and mixed components.

- As another example, the tensor product $U = V \otimes T$ of a *vector* V and a *rank-2 tensor* T is a *rank-3 tensor* with one possible set of components

$$\begin{aligned} U^\mu{}_{\alpha\beta} = (V \otimes T)^\mu{}_{\alpha\beta} &= (V \otimes T)(e^\mu, e_\alpha, e_\beta) \\ &= V(e^\mu) T(e_\alpha, e_\beta) \\ &= V^\mu T_{\alpha\beta}. \end{aligned}$$

In the mixed rank-3 tensor we have offset the upper and lower indices horizontally. The reason is associated with the symmetry under permutation of indices and will be discussed later.

Finally consider the scalar product

$$A \cdot B = g_{\mu\nu} A^\mu B^\nu .$$

- The metric tensor g may be interpreted as an operator that
  - takes two vectors as input and
  - returns their scalar product,

  which is a real number. For example:

$$\begin{aligned} g(A,B) &= g(A^\mu e_\mu, B^\nu e_\nu) \\ &= g(e_\mu, e_\nu) A^\mu B^\nu \\ &= g_{\mu\nu} A^\mu B^\nu \\ &= A \cdot B \ \in \mathbb{R}, \end{aligned}$$

- since it takes two vectors as input and acts linearly on both of them to return a number (this is an example of a *multi-linear mapping*),
- Therefore, $g_{\mu\nu}$ represents components of a rank-2 tensor of type $(0,2)$.

For quantum mechanics in Dirac notation a ket $|i\rangle$ is a vector and a bra $\langle f|$ is a dual vector in *Hilbert space*.

- A *matrix element* of an operator $\hat{Q}$ is of the form $\langle f|\hat{Q}|i\rangle$ in Dirac notation.
- In quantum mechanics, a *a matrix element is a scalar*.
- Thus the operator $\hat{Q}$ is a rank-2 tensor of type $(1,1)$, since it takes one vector and one dual vector as input and produces a scalar.

(In quantum mechanics a matrix element may be a complex rather than real number, but that distinction is not important in the present discussion.)

3.6.3 Identification of Vectors and Dual Vectors

Let's now *greatly simplify* keeping track of the difference between vectors and dual vectors by demonstrating that

> At a point P the *metric tensor map* establishes a *one-to-one relationship* between a
>
> - *vector* in the tangent space at P and a
> - *dual vector* in the cotangent space at P.

- Consider the metric tensor, viewed as a rank-2 tensor that
  - accepts two vectors as inputs and
  - acts on them (multi-)linearly to give a real number.
- Schematically, this may be written as the operator $g(\cdot\,,\cdot)$, where *the dots indicate the input slots* for the two vectors.
- Suppose that a vector V is inserted into only *one* of the slots, giving $g(V,\cdot)$.
- What is the object $g(V,\cdot)$?
  - It has *one open slot* that can *accept a vector*,
  - on which it will *act linearly to return a real number*.

  But that is just the *definition of a dual vector!*
- Because it is associated directly with the vector V, let's call this dual vector $\tilde{V} \equiv g(V\,,\cdot)$.

- The components of this dual vector may be evaluated by *inserting a basis vector as argument* in the usual way,

$$\begin{aligned} V_\mu \equiv \tilde{V}(e_\mu) &= g(V, e_\mu) && \text{(definition)} \\ &= g(V^\nu e_\nu, e_\mu) && \text{(expand } V \text{ in basis)} \\ &= V^\nu g(e_\nu, e_\mu) && \text{(rearrange using linearity)} \\ &= g_{\mu\nu} V^\nu && \text{(definition of metric components).} \end{aligned}$$

- Likewise, using that $g_{\mu\nu}$ and $g^{\mu\nu}$ are matrix inverses,

$$V^\mu = g^{\mu\nu} V_\nu .$$

- Thus, the properties of the metric tensor
  - imply that vectors and dual vectors may be treated *as if they were both vectors*,
  - one with an upper index and one with a lower index,
  - with the two related by *contraction (summing over repeated indices)* with the metric tensor,

$$V_\mu = g_{\mu\nu} V^\nu \qquad V^\mu = g^{\mu\nu} V_\nu .$$

This is true *only for manifolds with a metric*, but that is *always the case for GR*.

- The relations

$$V_\mu = g_{\mu\nu}V^\nu \qquad V^\mu = g^{\mu\nu}V_\nu.$$

  are of great practical importance since

  - they allow the same symbol to be used for a vector and its corresponding dual vector, and
  - they reduce handling of vectors and dual vectors to
  - keeping proper track of the vertical position of indices in the summation convention.

- This identification works only for manifolds with metric tensors but that is no limitation for general relativity, which deals *only* with metric spaces.
- The scalar product between two vectors U and V can now be calculated as

  - the *complete contraction* $U_\alpha V^\alpha$ of one of the vectors
  - with the dual vector associated with the other vector:

$$g(U,V) = g_{\mu\nu}U^\mu V^\nu = U_\nu V^\nu.$$

The scalar product has no indices left after contraction and is said to be *fully contracted*.

Because tensors of higher rank are products of vectors and dual vectors, the preceding discussion is easily generalized:

- Contraction with the metric tensor can be used to *raise or lower any index* for a tensor of any rank.
- For example,

$$A^{\mu\nu} = g^{\mu\alpha} g^{\nu\beta} A_{\alpha\beta} \qquad A_{\mu\nu\lambda\sigma} = g_{\mu\rho} A^{\rho}{}_{\nu\lambda\sigma}.$$

- Since indices can be raised or lowered at will by a metric,

> Tensors may be thought of as *geometrical objects of a particular tensorial rank*, irrespective of their particular vertical arrangement of indices when evaluated in a specific basis.

- Of course this is true only in the abstract; index placement matters when tensors are evaluated in a specific basis.

3.7 Tensors Specified by Transformation Laws

In the preceding discussion tensors have been introduced at a fundamental level through *linear maps from vectors and dual vectors to the real numbers*.

- However, it was shown also that
  - when tensors are expressed in an arbitrary basis,
  - their components obey well-defined transformation laws under change of coordinates.
- This view of tensors as groups of quantities obeying particular transformation laws is often the most practical for physical applications because
  - It is less abstract and requires less new mathematics.
  - A physical interpretation often requires expression of the problem in a well-chosen basis anyway.
  - The component index formalism has a handy built-in error checking mechanism:

    Failure of indices to balance on the two sides of an equation is a sure sign of an error.
- The next sections will summarize the use of tensors to formulate invariant equations by exploiting the transformation properties of their components.

3.7.1 Scalar Transformation Law

Tensors may be viewed as generalizing the idea of scalars and vectors, so let's begin with these more familiar quantities.

Simplest possibility: A field has a single component (magnitude) at each point that is unchanged by the transformation

$$\varphi'(x') = \varphi(x).$$

Quantities such as $\varphi(x)$ that are unchanged under the coordinate transformation are called *scalars*.

EXAMPLE: Value of the temperature at different points on the surface of the Earth.

3.7.2 Vectors and Dual Vectors

Recall also that we can classify tensors by a notation (n,m), where n is the number of upper indices and m is the number of lower indices when evaluated in a basis.

- Thus a scalar is a tensor of type $(0,0)$, since it carries no indices.
- The sum of n and m is the rank of the tensor. A scalar is a tensor of rank zero.

There are *two kinds of rank-1 tensors,* having the index pattern $(0,1)$ and $(1,0)$, respectively. The first is called a *dual vector*, *covariant vector*, or *1-form:*

DUAL VECTOR:

The gradient of a scalar field $\varphi(x) = \partial\varphi(x)/\partial x$ transforms under change of coordinates as

$$\left(\frac{\partial\varphi(x)}{\partial x'^{\mu}}\right) = \frac{\partial x^{\nu}}{\partial x'^{\mu}}\left(\frac{\partial\varphi(x)}{\partial x^{\nu}}\right).$$

Remember in such expressions:

- the Einstein summation convention, and
- that all partial derivatives are *understood implicitly* to be evaluated at some point $P = x$.

A tensor having a transformation law that *mimics that of the scalar field gradient*,

$$A'_{\mu}(x') = \frac{\partial x^{\nu}}{\partial x'^{\mu}} A_{\nu}(x) \qquad \text{(dual vector)}$$

is of type $(0,1)$ and is termed a *dual vector* (also *1-form*, *covariant vector*, or *covector*).

ECONOMY OF NOTATION: The preceding equation

$$A'_\mu(x') = \frac{\partial x^\nu}{\partial x'^\mu} A_\nu(x) \qquad \text{(dual vector)}$$

really means four equations:

$$A'_\mu = \frac{\partial x^0}{\partial x'^\mu} A_0 + \frac{\partial x^1}{\partial x'^\mu} A_1 + \frac{\partial x^2}{\partial x'^\mu} A_2 + \frac{\partial x^3}{\partial x'^\mu} A_3 \quad (\mu = 0,1,2,3)$$

each containing four terms. It is equivalent to the matrix equation

$$\begin{pmatrix} A'_0 \\ A'_1 \\ A'_2 \\ A'_3 \end{pmatrix} = \begin{pmatrix} \frac{\partial x^0}{\partial x'^0} & \frac{\partial x^1}{\partial x'^0} & \frac{\partial x^2}{\partial x'^0} & \frac{\partial x^3}{\partial x'^0} \\ \frac{\partial x^0}{\partial x'^1} & \frac{\partial x^1}{\partial x'^1} & \frac{\partial x^2}{\partial x'^1} & \frac{\partial x^3}{\partial x'^1} \\ \frac{\partial x^0}{\partial x'^2} & \frac{\partial x^1}{\partial x'^2} & \frac{\partial x^2}{\partial x'^2} & \frac{\partial x^3}{\partial x'^2} \\ \frac{\partial x^0}{\partial x'^3} & \frac{\partial x^1}{\partial x'^3} & \frac{\partial x^2}{\partial x'^3} & \frac{\partial x^3}{\partial x'^3} \end{pmatrix} \begin{pmatrix} A_0 \\ A_1 \\ A_2 \\ A_3 \end{pmatrix}.$$

VECTORS: A differential dx transforms like

$$dx'^{\mu} = \frac{\partial x'^{\mu}}{\partial x^{\nu}} dx^{\nu},$$

which suggests a second rank-1 transformation rule

$$A'^{\mu}(x') = \frac{\partial x'^{\mu}}{\partial x^{\nu}} A^{\nu}(x) \qquad \text{(vector)}.$$

A tensor that behaves in this way is of type $(1,0)$ and is termed a *vector* or *contravariant vector*.

Notice carefully the difference between the transformation laws for a vector and a dual vector,

$$A'_{\mu}(x') = \frac{\partial x^{\nu}}{\partial x'^{\mu}} A_{\nu}(x) \qquad \text{(dual vector)},$$

$$A'^{\mu}(x') = \frac{\partial x'^{\mu}}{\partial x^{\nu}} A^{\nu}(x) \qquad \text{(vector)}.$$

The transformation rules are similar, differing only in

- the vertical placement of the old coordinates x and new coordinates x' in the partial derivatives, and
- the corresponding vertical placement of indices required for consistency in the summation convention.

The dual vector and vector transformation laws,

$$A'_\mu(x') = \frac{\partial x^\nu}{\partial x'^\mu} A_\nu(x) \qquad \text{(dual vector)}$$

$$A'^\mu(x') = \frac{\partial x'^\mu}{\partial x^\nu} A^\nu(x) \qquad \text{(vector).}$$

may be viewed as matrix equations,

$$A'_\mu(x') = \hat{U}^\nu_\mu A_\nu(x) \qquad A'^\mu(x') = U^\mu_\nu A^\nu(x),$$

with the matrices defined by

$$U \equiv \frac{\partial x'}{\partial x} \qquad \hat{U} \equiv \frac{\partial x}{\partial x'} \qquad \hat{U}U = I,$$

where I is the 4×4 unit matrix. In these transformations

- the matrix U is called the *Jacobian matrix* and
- the matrix $\hat{U}$ is called the *inverse Jacobian matrix*.

The historical use of *covariant* to refer to lower indices and *contravariant* to upper indices arises from tensor invariance.

- In terms of the Jacobian matrix U and the inverse Jacobian matrix $\hat{U}$, under the coordinate transformation $x \to x'$
  - a vector transforms as $V'^{\mu} = U^{\mu}_{\nu} V^{\nu}$ and
  - a dual vector transforms as $\omega'_{\mu} = \hat{U}^{\nu}_{\mu} \omega_{\nu}$,

  with a corresponding change of coordinate basis.
- A vector V is invariant, so the basis vectors must transform in just such a way to cancel the change in the components.
- Specifically, since invariance of V requires

$$V = V^{\mu} e_{\mu} = V'^{\mu} e'_{\mu},$$

  the basis vectors must transform as $e'_{\mu} = \hat{U}^{\nu}_{\mu} e_{\nu}$ so that

$$V'^{\mu} e'_{\mu} = U^{\mu}_{\nu} V^{\nu} \hat{U}^{\alpha}_{\mu} e_{\alpha} = U^{\mu}_{\nu} \hat{U}^{\alpha}_{\mu} V^{\nu} e_{\alpha} = V^{\mu} e_{\mu},$$

  where $U^{\mu}_{\nu} \hat{U}^{\alpha}_{\mu} = \delta^{\alpha}_{\nu}$ was used.
- By a similar proof the invariance of $\omega = \omega_{\mu} e^{\mu}$ requires that the basis dual vectors transform as $e'^{\mu} = U^{\mu}_{\nu} e^{\nu}$.
- Thus lower-index components transform with $\hat{U}$, just as the components of the vector basis e_{μ} transform. They *co-vary with the basis* and are termed *covariant components.*
- But upper-index components transform with U and thus "opposite" to transformation of the basis vectors e_{μ}; they *contra-vary* and are termed *contravariant components.*

Summarizing, we expect the possibility of two rank-1 tensors:

1. *Dual vectors* (also called *one-forms*, *covariant vectors*, or *covectors*), which carry a lower index and transform like the gradient of a scalar:

$$A'_{\mu}(x') = \frac{\partial x^{\nu}}{\partial x'^{\mu}} A_{\nu}(x) \qquad \text{(dual vector)}.$$

2. *Vectors* (also called *contravariant vectors*), which carry an upper index and transform like the coordinate differential:

$$A'^{\mu}(x') = \frac{\partial x'^{\mu}}{\partial x^{\nu}} A^{\nu}(x) \qquad \text{(vector)}.$$

In the general case they must be distinguished (*by placement—upper or lower—of their indices*).

3.7.3 Scalar Product

Covariant and contravariant indices on vectors permit a scalar product to be defined as

$$A\cdot B \equiv A_\mu B^\mu = A^\mu B_\mu .$$

This transforms as a scalar because from

$$A'_\mu(x') = \frac{\partial x^\nu}{\partial x'^\mu} A_\nu(x) \qquad A'^\mu(x') = \frac{\partial x'^\mu}{\partial x^\nu} A^\nu(x)$$

we have that

$$\begin{aligned} A'\cdot B' = A'_\mu B'^\mu &= \frac{\partial x^\nu}{\partial x'^\mu} A_\nu \frac{\partial x'^\mu}{\partial x^\alpha} B^\alpha = \frac{\partial x^\nu}{\partial x'^\mu}\frac{\partial x'^\mu}{\partial x^\alpha} A_\nu B^\alpha \\ &= \frac{\partial x^\nu}{\partial x^\alpha} A_\nu B^\alpha = \delta^\nu_\alpha A_\nu B^\alpha \\ &= A_\alpha B^\alpha = A\cdot B, \end{aligned}$$

where the *Kronecker delta* is given by

$$\delta^\mu_\nu = \frac{\partial x'^\mu}{\partial x'^\nu} = \frac{\partial x^\mu}{\partial x^\nu} = \begin{cases} 1 & (\mu = \nu) \\ 0 & (\mu \neq \nu) \end{cases} .$$

Thus $A'\cdot B' = A\cdot B$ and *the scalar product is invariant*.

> Eliminating indices by summing over repeated ones is called *contraction*. The scalar product has no tensor indices left so it is *fully contracted*.

3.7.4 Rank-2 Tensors

Three kinds of rank-2 tensors transform as

$$T'_{\mu\nu} = \frac{\partial x^\alpha}{\partial x'^\mu}\frac{\partial x^\beta}{\partial x'^\nu} T_{\alpha\beta},$$

$$T'^{\nu}_{\ \mu} = \frac{\partial x^\alpha}{\partial x'^\mu}\frac{\partial x'^\nu}{\partial x^\beta} T^{\beta}_{\ \alpha},$$

$$T'^{\mu\nu} = \frac{\partial x'^\mu}{\partial x^\alpha}\frac{\partial x'^\nu}{\partial x^\beta} T^{\alpha\beta}.$$

This pattern may be generalized to tensors of any rank.

- *Covariant Tensors:* carry only lower indices
- *Contravariant Tensors:* carry only upper indices
- *Mixed Tensors:* carry both upper and lower indices

EXAMPLE: the Kronecker delta δ^ν_μ is a rank-2 mixed tensor.

Handy to recall:

- Upper index μ on left side requires right-side "factor" $\partial x'^\mu / \partial x^\nu$ (prime in numerator).
- Lower index ν on left side requires right-side "factor" $\partial x^\mu / \partial x'^\nu$ (prime in denominator).
- "Vertical position of index on left = vertical position of primed coordinate on right"

Not all quantities carrying indices are tensors! It is

- the *transformation laws* for components in a basis, or
- that they provide *linear maps to the real numbers*

that define tensors.

NOTE: We often employ a standard shorthand by using

- "a tensor $T_{\mu\nu}$" to mean
- "a tensor with components $T_{\mu\nu}$" when evaluated in a basis.

3.7.5 Metric Tensor

A rank-2 tensor of particular importance is the *metric tensor* $g_{\mu\nu}$ because it is associated with the line element

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu$$

that defines distances in metric spaces. It is symmetric ($g_{\mu\nu} = g_{\nu\mu}$) and satisfies the usual rank-2 transformation rule

$$g'_{\mu\nu} = \frac{\partial x^\alpha}{\partial x'^\mu} \frac{\partial x^\beta}{\partial x'^\nu} g_{\alpha\beta}.$$

The contravariant components of the metric tensor $g^{\mu\nu}$ are defined by

$$g_{\mu\alpha} g^{\alpha\nu} = \delta^\nu_\mu.$$

(That is, $g_{\mu\nu}$ and $g^{\mu\nu}$ are *matrix inverses*.)

Contractions with the metric tensor may be used to raise and lower (any number of) tensor indices; for example,

$$A^\mu = g^{\mu\nu}A_\nu \qquad A_\mu = g_{\mu\nu}A^\nu$$

$$T^\mu{}_\nu = g_{\nu\alpha}T^{\mu\alpha} \qquad T^\alpha{}_{\beta\gamma} = g^{\alpha\mu}g_{\gamma\varepsilon}T_{\mu\beta}{}^\varepsilon.$$

Thus, the scalar product of vectors may also be expressed

$$A\cdot B = g_{\mu\nu}A^\mu B^\nu \equiv A_\nu B^\nu.$$

In mixed-tensor expressions as above the *relative horizontal order of upper and lower indices* can be important.

- For example, in

  $$T^\mu{}_\nu = g_{\nu\alpha}T^{\mu\alpha}$$

  the notation indicates that the mixed tensor on the left side of the equation was obtained by lowering the rightmost index of $T^{\mu\alpha}$ on the right side.

- This distinction is immaterial if the tensor is *symmetric under exchange of indices* (see following pages).

- However, which index is lowered or raised matters for tensors that are *antisymmetric under index exchange:*

  $$T^\mu{}_\nu = g_{\nu\alpha}T^{\mu\alpha} \qquad T_\nu{}^\mu = g_{\nu\alpha}T^{\alpha\mu}$$

  are equivalent if T is symmetric, but different if T is antisymmetric.

Vectors and dual vectors are distinct entities that are defined in different spaces.

- However, the preceding discussion make it clear that for the *special case of a manifold with metric,*
- indices on any tensor may be raised or lowered at will by contraction with the metric tensor.

> Defining a metric establishes a relationship that permits vectors and dual vectors to be treated as if they were (in effect) different representations of the same vector.

- Our discussion will usually proceed as if A^μ and A_μ are different forms of the same vector that are related by contraction with the metric tensor,
- But secretly we will remember that they really are different, and that it is only for metric spaces that this conflation is not likely to land us in trouble.

3.8 Symmetric and Antisymmetric Tensors

The symmetry of tensors under exchanging pairs of indices is often important.

- An arbitrary rank-2 tensor can always be decomposed into a *symmetric part* and an *antisymmetric part:*

$$T_{\alpha\beta} = \tfrac{1}{2}(T_{\alpha\beta} + T_{\beta\alpha}) + \tfrac{1}{2}(T_{\alpha\beta} - T_{\beta\alpha}),$$

  where the first term is clearly symmetric and the second term antisymmetric under exchange of indices.

- For completely symmetric and completely antisymmetric rank-2 tensors we have

$$T_{\alpha\beta} = \pm T_{\beta\alpha} \qquad T^{\alpha\beta} = \pm T^{\beta\alpha} \qquad T_{\alpha}{}^{\beta} = \pm T^{\beta}{}_{\alpha},$$

  where the plus sign holds if the tensor is symmetric and the minus sign if it is antisymmetric.

- More generally, we say that a tensor of rank two or higher is
  - *Symmetric* in any two of its indices if exchanging those indices leaves the tensor invariant and
  - *Antisymmetric* (sometimes termed *skew-symmetric*) in any two indices if it changes sign upon switching those indices.

- Symmetrizing and antisymmetrizing operations on tensor indices may be denoted by a bracket notation in which
  - $(\,)$ indicates symmetrization and
  - $[\,]$ indicates antisymmetrization

  over indices included in the brackets.
- For example, *symmetrization over all indices* for a rank-N covariant tensor $T_{\alpha,\beta,\ldots\omega}$ corresponds to

$$T_{(\alpha,\beta,\ldots\omega)} \equiv \frac{1}{N!}(\text{Sum permutations on indices } \alpha,\beta,\ldots\omega)$$

  and *antisymmetrization over all indices* of $T_{\alpha,\beta,\ldots\omega}$ corresponds to

$$T_{[\alpha,\beta,\ldots\omega]} \equiv \frac{1}{N!}(\pm\,\text{Sum permutations on indices } \alpha,\beta,\ldots\omega)\,,$$

  where the notation $\pm$ indicates that terms of the sum have
  - a plus sign if they correspond to an even number of index exchanges and
  - a negative sign if they correspond to an odd number of index exchanges.
- Thus, for rank-2 contravariant tensors we may write

$$T^{(\alpha\beta)} = \tfrac{1}{2}(T^{\alpha\beta} + T^{\beta\alpha}) \qquad T^{[\alpha\beta]} = \tfrac{1}{2}(T^{\alpha\beta} - T^{\beta\alpha})$$

- In the more general case one may be interested in symmetrizing or antisymmetrizing over only a *subset of indices for higher-rank tensors.*
- If the indices are contiguous the above notation suffices with only the indices to be symmetrized or antisymmetrized included in the brackets.
- In the event that indices to be symmetrized or antisymmetrized are not adjacent to each other, the preceding notation may be extended by using *vertical brackets to exclude indices from the symmetrization or antisymmetrization.*

Example: The expression

$$T_{\alpha[\beta|\gamma|\delta]} = \tfrac{1}{2}(T_{\alpha\beta\gamma\delta} - T_{\alpha\delta\gamma\beta})$$

corresponds to a rank-4 covariant tensor that has been antisymmetrized in its second and fourth indices only.

3.9 Summary of Algebraic Tensor Operations

Various algebraic operations are permitted for tensors:

- *Multiplication by a scalar:* For example,

$$aA^{\mu\nu} = B^{\mu\nu},$$

where a is a scalar and $A^{\mu\nu}$ and $B^{\mu\nu}$ are rank-2 tensors.

- *Addition or subtraction:* Two tensors of the same type may be added or subtracted (meaning that their components are added or subtracted) to produce a new tensor of the same type. For example,

$$A^{\mu} - B^{\mu} = C^{\mu},$$

where A^{μ}, B^{μ}, and C^{μ} are vectors.

- *Multiplication:* Tensors may be multiplied by forming products of components. The rank of the resultant tensor will be the sum of the ranks of the factors. Example:

$$A_{\mu\nu} = U_{\mu}V_{\nu},$$

- *Contraction:* For a tensor of type (n,m), a tensor of covariant rank $n-1$ and contravariant rank $m-1$ may be formed by setting one upper and one lower index equal and taking the implied sum. For example,

$$A = A^{\mu}{}_{\mu},$$

where A is a scalar and $A^{\mu}{}_{\nu}$ is a mixed rank-2 tensor.

3.10 Tensor Calculus on Curved Manifolds

To formulate physical theories in terms of tensors requires the ability to manipulate tensors mathematically.

- In addition to the algebraic rules for tensors described in preceding sections, we must formulate
  - a prescription to integrate tensor equations and
  - a prescription to differentiate them.
- *Tensor calculus* is mostly a straightforward generalization of normal calculus but additional complexity arises for two reasons:
  - It must be ensured that integration and differentiation *preserve any physical symmetries.*
  - It must be ensured that operations on tensor equations *preserve the tensor structure.*
- We will see that
  - *tensor integration* requires a simple modification of the standard integration rules, but
  - *derivatives of tensors* require a less-simple modification with far-reaching mathematical and physical implications.

3.10.1 Invariant Integration

Change of volume elements for spacetime integration:

$$d^4x = \det\left(\frac{\partial x}{\partial x'}\right) d^4x',$$

where $\det(\partial(x)/\partial(x'))$ is the *Jacobian determinant* of the transformation between the coordinates.

- The metric tensor transforms as

$$g'_{\mu\nu} = \frac{\partial x^\alpha}{\partial x'^\mu}\frac{\partial x^\beta}{\partial x'^\nu} g_{\alpha\beta} \qquad \text{(triple matrix product)}.$$

- Therefore, since:

$$\text{determinant of a product} = \text{product of determinants},$$

- the determinant of the metric tensor $g \equiv \det g_{\mu\nu}$ transforms as

$$g' = \det\left(\frac{\partial x}{\partial x'}\right)\det\left(\frac{\partial x}{\partial x'}\right) g \;\rightarrow\; \det\left(\frac{\partial x}{\partial x'}\right) = \frac{\sqrt{|g'|}}{\sqrt{|g|}},$$

which gives when inserted into the first equation

$$\sqrt{|g|}\, d^4x = \sqrt{|g'|}\, d^4x',$$

($|g|$ because g can be negative in 4-D spacetime).

Therefore, in integrals we shall employ

$$dV = \sqrt{|g|}\, d^4x,$$

as an *invariant volume element*.

Example: The metric for a *2-dimensional spherical manifold (2-sphere)* is specified by the line element

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

which is explicitly in spherical coordinates

$$d\ell^2 = R^2 d\theta^2 + R^2 \sin^2\theta d\varphi^2.$$

This may be written as the matrix equation

$$d\ell^2 = (d\theta \quad d\varphi) \underbrace{\begin{pmatrix} R^2 & 0 \\ 0 & R^2\sin^2\theta \end{pmatrix}}_{g_{ij}} \begin{pmatrix} d\theta \\ d\varphi \end{pmatrix}.$$

The area of the 2-sphere may then be expressed as the *"invariant volume integration"*

$$\begin{aligned} A &= \int \sqrt{|g|}\, d^2x = \int_0^{2\pi} d\varphi \int_0^{\pi} \sqrt{\det g_{ij}}\, d\theta \\ &= \int_0^{2\pi} d\varphi \int_0^{\pi} R^2 \sin\theta d\theta = 4\pi R^2. \end{aligned}$$

where the metric tensor g_{ij} is the 2×2 matrix in the preceding equation for the line element.

> In this 2-dimensional example the sign of the determinant is positive, so no absolute value is required under the radical.

3.10.2 Covariant Differentiation

Let's now consider the derivatives of tensor quantities. First introduce two common compact notations for partial derivatives

$$\partial_\mu \varphi = \varphi_{,\mu} \equiv \frac{\partial \varphi(x)}{\partial x^\mu} \qquad \partial^\mu \varphi = \varphi^{,\mu} \equiv \frac{\partial \varphi(x)}{\partial x_\mu}.$$

- The derivative of a scalar is a covariant vector and *scalars and their derivatives are well-defined tensors.*
- But, for the *derivative of a dual vector,* by the product rule

$$\begin{aligned} A'_{\mu,\nu} \equiv \frac{\partial A'_\mu}{\partial x'^\nu} &= \frac{\partial}{\partial x'^\nu} \underbrace{\left(A_\alpha \frac{\partial x^\alpha}{\partial x'^\mu} \right)}_{A'_\mu} \\ &= \frac{\partial A_\alpha}{\partial x'^\nu} \frac{\partial x^\alpha}{\partial x'^\mu} + A_\alpha \frac{\partial^2 x^\alpha}{\partial x'^\nu \partial x'^\mu} \quad (\partial/\partial x' \text{ of product}) \\ &= \underbrace{\frac{\partial A_\alpha}{\partial x^\beta} \frac{\partial x^\beta}{\partial x'^\nu}}_{\text{chain rule}} \frac{\partial x^\alpha}{\partial x'^\mu} + A_\alpha \frac{\partial^2 x^\alpha}{\partial x'^\nu \partial x'^\mu} \\ &= \underbrace{A_{\alpha,\beta} \frac{\partial x^\beta}{\partial x'^\nu} \frac{\partial x^\alpha}{\partial x'^\mu}}_{\textit{Tensor}} + \underbrace{A_\alpha \frac{\partial^2 x^\alpha}{\partial x'^\nu \partial x'^\mu}}_{\textit{Not a tensor!}} \end{aligned}$$

where $A_{\alpha,\beta} \equiv \partial A_\alpha / \partial x^\beta$. In curved spacetime *the second term can't be eliminated:*

> *Partial differentiation is NOT COVARIANT in curved spacetime* for tensors of rank 1 or higher!

Partial differentiation is not generally a covariant operation in curved manifolds.

- This will complicate the formalism immensely because
- the utility of the tensor framework rests on *preservation of tensor structure under transformations.*

It is desirable to define a new covariant derivative that ensures this automatically.

- The terms that violate tensor transformation laws for partial derivatives of tensors will involve *second derivatives*.
- The offending non-tensorial contributions can be eliminated systematically by introducing *additional fields on the manifold*.
- There is more than one way to do this, each leading to a different form of covariant differentiation. We will address three in this chapter:

  1. *covariant derivatives* and
  2. *absolute derivatives,* whichuse derivatives of the metric tensor field to cancel non-tensorial terms, and
  3. *Lie derivatives,* which use derivatives of an auxiliary vector field defined on the manifold to the same end.

We begin with the *covariant derivative*.

3.11 The Covariant Derivative

Let's define the *Christoffel symbols* $\Gamma^{\lambda}_{\alpha\beta}$ by requiring that they obey a transformation law

$$\Gamma'^{\lambda}_{\alpha\beta} = \Gamma^{\kappa}_{\mu\nu}\frac{\partial x^{\mu}}{\partial x'^{\alpha}}\frac{\partial x^{\nu}}{\partial x'^{\beta}}\frac{\partial x'^{\lambda}}{\partial x^{\kappa}} + \frac{\partial^2 x^{\mu}}{\partial x'^{\alpha}\partial x'^{\beta}}\frac{\partial x'^{\lambda}}{\partial x^{\mu}}.$$

($\Gamma^{\lambda}_{\alpha\beta}$ is *not a tensor—see the 2nd derivatives above!)* Then this transformation law imples that (Problem),

$$\underbrace{\left(A'_{\mu,\nu} - \Gamma'^{\lambda}_{\mu\nu}A'_{\lambda}\right)}_{\equiv B'_{\mu\nu}} = \underbrace{\left(A_{\alpha,\beta} - \Gamma^{\kappa}_{\alpha\beta}A_{\kappa}\right)}_{\equiv B_{\alpha\beta}}\frac{\partial x^{\alpha}}{\partial x'^{\mu}}\frac{\partial x^{\beta}}{\partial x'^{\nu}}.$$

For the quantity in parentheses this is the transformation law for a rank-2 covariant tensor:

$$B'_{\mu\nu} = B_{\alpha\beta}\frac{\partial x^{\alpha}}{\partial x'^{\mu}}\frac{\partial x^{\beta}}{\partial x'^{\nu}}.$$

This suggests that we define the *covariant derivative* of A_{μ} as

$$A_{\mu;\nu} \equiv \overbrace{\underbrace{A_{\mu,\nu}}_{\text{not tensor}} - \underbrace{\Gamma^{\lambda}_{\mu\nu}A_{\lambda}}_{\text{not tensor}}}^{\text{tensor}}$$

where in our notation

- a *subscript comma* denotes partial differentiation;
- a *subscript semicolon* denotes covariant differentiation

with respect to the variables following it.

- Then the covariant derivative of a dual vector

$$A_{\mu;\nu} \equiv \overbrace{\underbrace{A_{\mu,\nu}}_{\text{not tensor}} - \underbrace{\Gamma^{\lambda}_{\mu\nu}A_{\lambda}}_{\text{not tensor}}}^{\text{tensor}}$$

  transforms as a *covariant tensor of rank 2*, even though neither of its terms is a tensor.

- It will be useful to introduce also an alternative notation for the covariant derivative:

$$\nabla_{\nu}A_{\mu} = A_{\mu;\nu} \equiv \partial_{\nu}A_{\mu} - \Gamma^{\lambda}_{\mu\nu}A_{\lambda}\,.$$

  where ∇_{μ} denotes an operator that takes the covariant derivative with respect to x^{μ}:

Likewise, we can introduce the covariant derivative of a vector in either of the notations:

$$A^{\lambda}{}_{;\mu} = A^{\lambda}{}_{,\mu} + \Gamma^{\lambda}_{\alpha\mu} A^{\alpha} \qquad \nabla_{\mu} A^{\lambda} = \partial_{\mu} A^{\lambda} + \Gamma^{\lambda}_{\alpha\mu} A^{\alpha},$$

where the result is a mixed rank-2 tensor, and the covariant derivatives of the three possible rank-2 tensors through

$$\begin{aligned} A_{\mu\nu;\lambda} &= A_{\mu\nu,\lambda} - \Gamma^{\alpha}_{\mu\lambda} A_{\alpha\nu} - \Gamma^{\alpha}_{\nu\lambda} A_{\mu\alpha}, \\ A^{\mu}{}_{\nu;\lambda} &= A^{\mu}{}_{\nu,\lambda} + \Gamma^{\mu}_{\alpha\lambda} A^{\alpha}{}_{\nu} - \Gamma^{\alpha}_{\nu\lambda} A^{\mu}{}_{\alpha}, \\ A^{\mu\nu}{}_{;\lambda} &= A^{\mu\nu}{}_{,\lambda} + \Gamma^{\mu}_{\alpha\lambda} A^{\alpha\nu} + \Gamma^{\nu}_{\alpha\lambda} A^{\mu\alpha}, \end{aligned}$$

or in alternative notation

$$\begin{aligned} \nabla_{\lambda} A_{\mu\nu} &= \partial_{\lambda} A_{\mu\nu} - \Gamma^{\alpha}_{\mu\lambda} A_{\alpha\nu} - \Gamma^{\alpha}_{\nu\lambda} A_{\mu\alpha}, \\ \nabla_{\lambda} A^{\mu}{}_{\nu} &= \partial_{\lambda} A^{\mu}{}_{\nu} + \Gamma^{\mu}_{\alpha\lambda} A^{\alpha}{}_{\nu} - \Gamma^{\alpha}_{\nu\lambda} A^{\mu}{}_{\alpha}, \\ \nabla_{\lambda} A^{\mu\nu} &= \partial_{\lambda} A^{\mu\nu} + \Gamma^{\mu}_{\alpha\lambda} A^{\alpha\nu} + \Gamma^{\nu}_{\alpha\lambda} A^{\mu\alpha}, \end{aligned}$$

where the derivatives now define rank-3 tensors.

From the form of equations like

$$\begin{aligned}
\nabla_\lambda A_{\mu\nu} &= \partial_\lambda A_{\mu\nu} - \Gamma^\alpha_{\mu\lambda} A_{\alpha\nu} - \Gamma^\alpha_{\nu\lambda} A_{\mu\alpha}, \\
\nabla_\lambda A^\mu{}_\nu &= \partial_\lambda A^\mu{}_\nu + \Gamma^\mu_{\alpha\lambda} A^\alpha{}_\nu - \Gamma^\alpha_{\nu\lambda} A^\mu{}_\alpha, \\
\nabla_\lambda A^{\mu\nu} &= \partial_\lambda A^{\mu\nu} + \Gamma^\mu_{\alpha\lambda} A^{\alpha\nu} + \Gamma^\nu_{\alpha\lambda} A^{\mu\alpha},
\end{aligned}$$

we may formulate simple rules for constructing the covariant derivative of a tensor having any rank:

- Form the ordinary partial derivative of the tensor
- Add one Christoffel symbol term having the sign and form for a dual (covariant) vector for each lower index of the tensor
- Add one Christoffel symbol term having the sign and form for a (contravariant) vector for each upper index of the tensor.

Most rules for partial differentiation carry over with suitable generalization for covariant differentiation.

Example:
Covariant derivative of a product

$$(A_\mu B_\nu)_{;\lambda} = A_{\mu;\lambda} B_\nu + A_\mu B_{\nu;\lambda}\,.$$

which is the usual (Leibniz rule) result.

The most important exception concerns the properties of successive covariant differentiations.

- Although partial derivative operators normally commute, *covariant derivative operators generally do not commute* with each other.
- *The covariant derivative of the metric tensor vanishes* (Problem):

  $$\nabla_\alpha g_{\mu\nu} = g_{\mu\nu;\alpha} = 0,$$

  Some implications:
  - Raising and lowering index by contraction with $g_{\mu\nu}$ commutes with covariant differentiation.
  - This will allow in the Einstein field equations a *vacuum energy term* (accelerated expansion and dark energy) when we address cosmology.

3.12 Absolute Derivatives

Absolute derivatives (also termed *intrinsic derivatives*) are closely related to covariant derivatives.

- *Covariant derivatives* are defined over an entire manifold in terms of partial derivatives plus correction terms to cancel non-tensorial character.
- *Absolute derivatives* are defined only along constrained paths in terms of ordinary derivatives plus correction terms to cancel non-tensorial behavior.
- Using $D/D\sigma$ to denote the absolute derivative along a path parameterized by σ,

$$\frac{DA_\alpha}{D\sigma} \equiv \frac{dA_\alpha}{d\sigma} - \Gamma^\beta_{\alpha\gamma} A_\beta \frac{dx^\gamma}{d\sigma} \qquad \text{(Dual vectors)},$$

$$\frac{DA^\alpha}{D\sigma} \equiv \frac{dA^\alpha}{d\sigma} + \Gamma^\alpha_{\beta\gamma} A^\beta \frac{dx^\gamma}{d\sigma} \qquad \text{(Vectors)},$$

  with generalizations for higher-order tensors similar to that discussed earlier for covariant derivatives.
- The essential utility of both covariant and absolute derivatives is that
  - When they are *applied to tensor fields* they *produce tensor fields*.
  - They provide a *prescription for parallel transport* in curved spaces that will be discussed later.

3.13 Lie Derivatives

The covariant derivative was introduced as a modification of partial differentiation that respects tensor structure.

- An alternative way to modify partial differentiation so that when applied to tensors it yields tensors is through the *Lie derivative.*
- The covariant derivative uses derivatives of the metric tensor to cancel non-tensorial terms arising in partial differentiation.
- In contrast, the Lie derivative uses *derivatives of an auxiliary vector field* to cancel those same terms.
- We shall use primarily the covariant derivative to implement differentiation in curved spaces.
- However, it will be important in various contexts to have at least a conceptual understanding of the Lie derivative.

A primary reason is that symmetries important for a broad range of physical applications have an intimate connection to Lie derivatives.

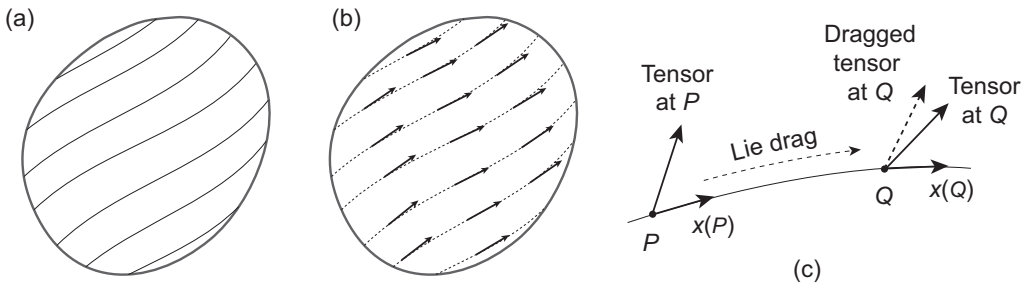


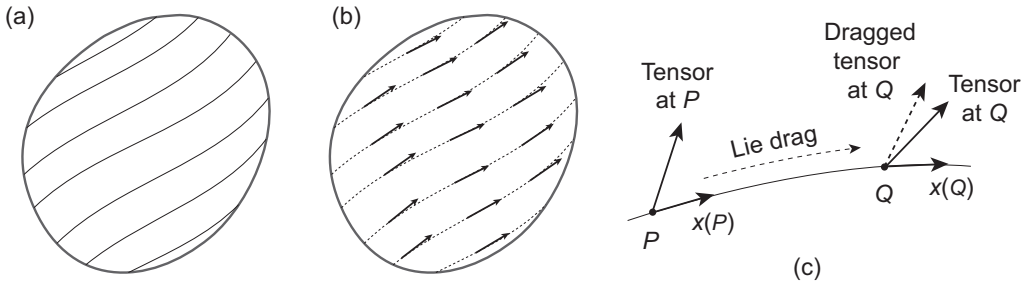
Figure 3.4: (a) A congruence for a manifold. (b) Tangent vector field for the curves of the congruence. (c) *Lie dragging* using the congruence to compare tensors at two nearby points.

The basic idea: The starting point is the idea of a *congruence* for a manifold, which is

- a set of non-intersecting curves that are
- space-filling in that for each point in the manifold exactly one curve of the congruence passes through it.
- As an example, for the sphere S^2 curves corresponding to lines of latitude define a congruence since for every point on S^2 exactly one curve of latitude passes through it.

Figure 3.4(a) illustrates a congruence for a general manifold.

- For each point in the manifold a basis vector may be defined by a tangent to the curve at that point [Fig. 3.4(b)].
- Thus a congruence defines a vector field on the manifold.
- The converse is also true. For any vector field X^μ defined on a manifold, a congruence can be generated by finding the curves for which the vectors of the vector field are tangent at each point of the curve.

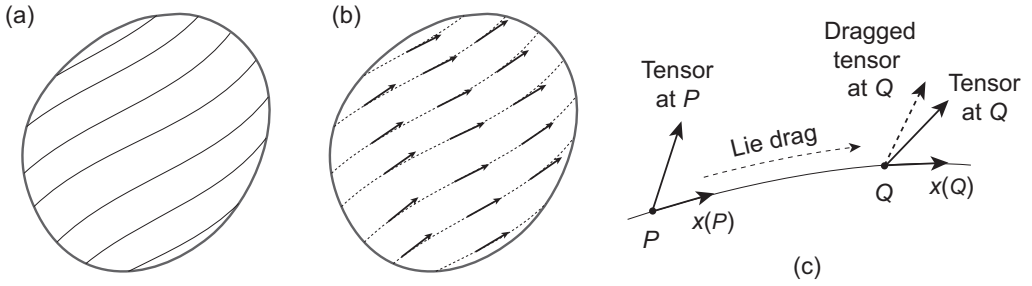


- Such congruences are familiar to the physicist in the guise of streamlines connecting magnetic field vectors or illustrating fluid velocity fields.
- In the present more general context,
  - a curve of the congruence for a vector field X^μ is termed an *orbit* and
  - the vectors X^μ are said to *generate the orbit*.

Figure (c) above illustrates the main idea of the Lie derivative.

- For a tensor field defined on the manifold, a tensor at the point P (here illustrated by a vector) is
- "dragged" (prescription given below) from the point P to the nearby point Q, along a curve of the congruence.
- This process is called *Lie dragging.*

Then the tensor defined naturally at Q, and the tensor dragged from P, may be subjected to the usual difference procedure to define a derivative since they are now located at the same point. The resulting derivative is termed the *Lie derivative*.



Constructing Lie derivatives: Suppose that we have a manifold with a vector field X^μ and a corresponding congruence.

- Let some tensor field be defined on the manifold.
- To be definite for this example it will be taken to be a contravariant rank-2 tensor, $T^{\mu\nu}(x)$.
- Now consider the transformation for small δu,

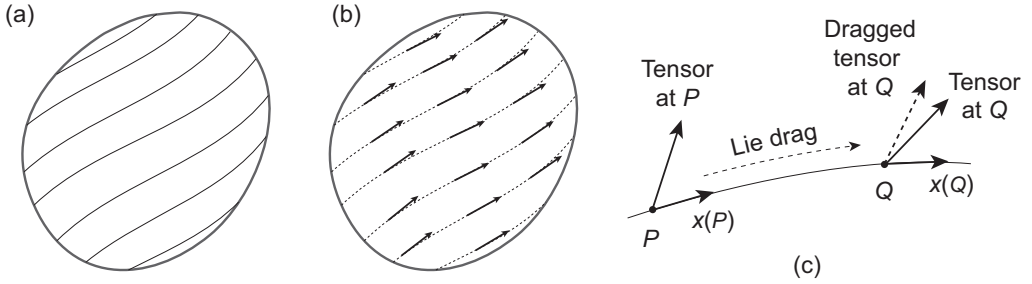
$$x'^\mu = x^\mu + X^\mu(x)\,\delta u.$$

- This will be regarded as an *active transformation* in which
  - the point P at x^μ
  - is sent to the point Q at

$$x^\mu + X^\mu(x)\delta u,$$

  with both points labeled in the *same coordinate system*.

By construction the point Q lies on the congruence through P that is generated by X^μ; see Fig. (c) above.



Now consider the tensor with components $T^{\mu\nu}(x)$ at P.

- It is mapped to the tensor with components $T'^{\mu\nu}(x')$ at Q by the transformation

$$x'^{\mu} = x^{\mu} + X^{\mu}(x)\,\delta u.$$

- That is, the tensor is *Lie-dragged* from P to Q by the transformation, as illustrated in Fig. (c) above.
- By the usual tensor transformation law,

$$\begin{aligned} T'^{\mu\nu}(x') &= \frac{\partial x'^{\mu}}{\partial x^{\alpha}}\frac{\partial x'^{\nu}}{\partial x^{\beta}}T^{\alpha\beta}(x) \\ &= T^{\mu\nu}(x) + \left[\partial_{\beta}X^{\nu}T^{\mu\beta} + \partial_{\alpha}X^{\mu}T^{\alpha\nu}\right]\delta u, \end{aligned}$$

  where we have again used

$$x'^{\mu} = x^{\mu} + X^{\mu}(x)\,\delta u.$$

- This gives the tensor $T^{\mu\nu}$ Lie-dragged from P to Q.

- The Lie derivative of $T^{\mu\nu}$ with respect to the congruence of the vector field X is defined by

$$\mathscr{L}_X T^{\mu\nu} \equiv \lim_{\delta u \to 0} \left(\frac{T^{\mu\nu}(x') - T'^{\mu\nu}(x')}{\delta u} \right).$$

- The tensor $T'^{\mu\nu}(x')$ was given above by

$$\begin{aligned} T'^{\mu\nu}(x') &= \frac{\partial x'^{\mu}}{\partial x^{\alpha}} \frac{\partial x'^{\nu}}{\partial x^{\beta}} T^{\alpha\beta}(x) \\ &= T^{\mu\nu}(x) + \left[\partial_\beta X^\nu T^{\mu\beta} + \partial_\alpha X^\mu T^{\alpha\nu} \right] \delta u, \end{aligned}$$

  and the tensor $T^{\mu\nu}(x')$ can be determined by expanding in a Taylor series around x,

$$T^{\mu\nu}(x') = T^{\mu\nu}(x) + (\delta u)\partial_\alpha T^{\mu\nu} X^\alpha + \mathscr{O}\left(\delta u^2\right).$$

- Then substituting gives for the Lie derivative of $T^{\mu\nu}$,

$$\mathscr{L}_X T^{\mu\nu} = X^\alpha \partial_\alpha T^{\mu\nu} - T^{\mu\alpha}\partial_\alpha X^\nu - T^{\alpha\nu}\partial_\alpha X^\mu.$$

By a similar procedure Lie derivatives for tensors of other ranks can be determined. For example,

$$\begin{aligned} \mathscr{L}_X \varphi &= X\varphi = X^\alpha \partial_\alpha \varphi, \\ \mathscr{L}_X A^\mu &= X^\alpha \partial_\alpha A^\mu - A^\alpha \partial_\alpha X^\mu, \\ \mathscr{L}_X A_\mu &= X^\alpha \partial_\alpha A_\mu + A_\alpha \partial_\mu X^\alpha, \\ \mathscr{L}_X A_{\mu\nu} &= X^\alpha \partial_\alpha A_{\mu\nu} + A_{\mu\alpha}\partial_\nu X^\alpha + A_{\alpha\nu}\partial_\mu X^\alpha, \\ \mathscr{L}_X A^{\mu\nu} &= X^\alpha \partial_\alpha A^{\mu\nu} - A^{\mu\alpha}\partial_\alpha X^\nu - A^{\alpha\nu}\partial_\alpha X^\mu. \end{aligned}$$

Mathematically the Lie derivative is a *more primitive concept* than the covariant derivative because it requires *less added structure on the manifold*.

- For example, in the preceding derivations no appeal was made to either a *metric* or to *connection (Christoffel) coefficients*.
- For manifolds of interest for general relativity (those having a *metric* and a *torsion-free connection*; see Section 7.8 of book), all partial derivatives ∂_μ may be replaced by covariant derivatives ∇_μ in the Lie derivative

  This is because all terms involving connection coefficients *cancel identically* if ∇_μ is substituted for ∂_μ in the preceding expressions for Lie derivatives.

For example,

$$\mathscr{L}_X A_{\mu\nu} = X^\alpha \nabla_\alpha A_{\mu\nu} + A_{\mu\alpha}\nabla_\nu X^\alpha + A_{\alpha\nu}\nabla_\mu X^\alpha$$

and

$$\mathscr{L}_X A_{\mu\nu} = X^\alpha \partial_\alpha A_{\mu\nu} + A_{\mu\alpha}\partial_\nu X^\alpha + A_{\alpha\nu}\partial_\mu X^\alpha,$$

are equally valid if the manifold has a torsion-free connection.

This demonstrates explicitly that even though only partial derivatives were used in its construction, *the Lie derivative of a tensor is a tensor*.

Lie transport and isometries:

A tensor T is said to be

- *Lie-transported* along a curve of the congruence associated with the vector field V if the Lie derivative vanishes,

$$\mathscr{L}_V T = 0.$$

- An interesting question for a manifold with metric is whether there exists a vector field K such that *the Lie derivative of the metric vanishes*,

$$\mathscr{L}_K g_{\mu\nu} = 0.$$

We shall take this question up later, where it will be shown that

- if $\mathscr{L}_K$ applied to the metric tensor gives zero,
- then K is a *Killing vector field* and
- the corresponding *Killing vectors* indicate directions in which *the metric is invariant*.

Such symmetries of the metric are called *isometries.*

3.14 Invariant Equations

The properties of tensors elaborated above ensure that

> *Any equation will be invariant under general coordinate transformations provided that it equates tensors of the same type (equates components having the same upper and lower indices when expressed in a basis).*

EXAMPLES:

- If $A^{\mu}{}_{\nu}$ and $B^{\mu}{}_{\nu}$ each transform as mixed rank-2 tensors and $A^{\mu}{}_{\nu} = B^{\mu}{}_{\nu}$ in the x coordinate system, then in the x' coordinate system $A'^{\mu}{}_{\nu} = B'^{\mu}{}_{\nu}$.
- Likewise, an equation that equates any tensor to zero (that is, sets all its components to zero) in some coordinate system is covariant under general coordinate transformations, implying that the tensor is equal to zero in all coordinate systems.
- However, equations such as $A^{\nu}{}_{\mu} = 10$ or $A^{\mu} = B_{\mu}$ might hold in particular coordinate systems but generally not in all coordinate systems because they equate tensors of different kinds:
  - a mixed rank-2 tensor with a scalar in the first case;
  - a dual vector with a vector in the second.

The preceding discussion suggests that invariance of a theory under general coordinate transformations will be guaranteed by carrying out the following steps.

1. Formulate all quantities in terms of tensors,
   - with tensor types matching on the two sides of any equation, and
   - with all algebraic manipulations corresponding to valid tensor operations (addition, multiplication, contraction, . . .).
2. Redefine any integration to be *invariant integration*.
3. Replace all partial derivatives with *covariant derivatives*.
4. Take care to remember that a *covariant differentiation generally does not commute* with a second covariant differentiation.

As will be demonstrated in subsequent chapters, this prescription in terms of tensors will provide a powerful formalism for dealing with mathematical relations that would be much more formidable in standard notation

Chapter 4

Lorentz Covariance and Special Relativity

To go beyond Newtonian gravitation we must consider, with Einstein, the intimate relationship between the curvature of space and the gravitational field.

- Mathematically, this extension is bound inextricably to the *geometry of spacetime*, and in particular to the aspect of geometry that permits quantitative measurement of distances.
- Let us first consider these ideas within the 4-dimensional spacetime termed *Minkowski space*.

As we shall see, requiring covariance within Minkowski space will lead us to the *special theory of relativity*.

4.0.1 The Indefinite Metric of Spacetime

A manifold equipped with a prescription for measuring distances is termed a *metric space* and the mathematical function that specifies distances is termed the *metric* for the space.

- Familiar examples of metrics were introduced earlier.
- In this section those ideas are applied to flat 4-dimensional spacetime, which is commonly termed *Minkowski space.*
- Although many concepts will be similar to those introduced earlier, fundamentally new features will enter.

Many of these new features are associated with the *indefinite metric* of Minkowski space.

- Minkowski space is *flat* but it is *not euclidean*, for it does not possess a euclidean metric.
- Many of the metrics employed in earlier chapters could be put into a diagonal form in which the *signs of the diagonal entries could all be chosen positive*.
- Such a metric is termed *positive definite.*
- In contrast, we will see that the Minkowski metric
  - can be put into diagonal form but
  - it is an essential property of Minkowski space that the *diagonal entries cannot all be chosen positive*.

Such a metric is termed *indefinite,* and it leads to properties differing fundamentally from those of euclidean spaces.

4.1 Minkowski Space

In a particular inertial frame, introduce unit vectors e_0, e_1, e_2, and e_3 that point along the t, x, y, and z axes. Any 4-vector A may be expressed in the form,

$$A = A^0 e_0 + A^1 e_1 + A^2 e_2 + A^3 e_3.$$

and the scalar product of 4-vectors is given by

$$A \cdot B = B \cdot A = (A^\mu e_\mu) \cdot (B^\nu e_\nu) = e_\mu \cdot e_\nu A^\mu B^\nu.$$

Note that generally we shall use

- non-bold symbols to denote 4-vectors
- bold symbols for 3-vectors.

We sometimes use a notation such as b^μ to stand generically for all components of a 4-vector.

Defining the metric tensor $\eta_{\mu\nu}$ in Minkowski space,

$$\eta_{\mu\nu} \equiv e_\mu \cdot e_\nu,$$

(it is conventional to denote the metric by $\eta_{\mu\nu}$ rather than $g_{\mu\nu}$ in Minkowski space) the scalar product may be expressed as

$$A \cdot B = \eta_{\mu\nu} A^\mu B^\nu,$$

and the Minkowski-space line element is

$$ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2 = \eta_{\mu\nu}dx^\mu dx^\nu,$$

where the *metric tensor of flat spacetime* may be expressed as

$$\eta_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \text{diag}\,(-1,1,1,1).$$

Thus $ds^2 = \eta_{\mu\nu}dx^\mu dx^\nu$ corresponds to the matrix equation

$$ds^2 = (cdt \;\; dx \;\; dy \;\; dz) \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} cdt \\ dx \\ dy \\ dz \end{pmatrix},$$

where ds^2 is the spacetime interval between x and $x+dx$ with

$$x = (x^0,x^1,x^2,x^3) = (ct,x^1,x^2,x^3) = (ct,\boldsymbol{x}).$$

- The Minkowski metric is indefinite and is sometimes termed *pseudo-euclidean.*
- As explained below, such metrics are also said to have a *Lorentzian signature.*

Example 4.1

Given a Minkowski vector with components

$$A^{\mu} = (A^0, A^1, A^2, A^3),$$

what are the components of the corresponding dual vector? Using $\eta_{\mu\nu}$ for the metric tensor, the indices may be lowered through the contraction

$$A_{\mu} = \eta_{\mu\nu} A^{\nu}.$$

Therefore, using the Minkowski metric tensor

$$\eta_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

the elements of the corresponding dual vector are

$$A_{\mu} = (-A^0, A^1, A^2, A^3).$$

This illustrates explicitly that

- vectors and dual vectors generally are not equivalent in non-euclidean manifolds, but that
- they are in one-to-one correspondence though contraction with the metric tensor.

The Minkowski-space metric

$$\eta_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

is diagonal, with relative sign of the diagonal terms $(-+++)$.

- This sign pattern is termed the *signature of the metric.*
- (Some authors instead define the signature to be an *integer* equal to the difference of the number of positive signs and number of negative signs.)
- It is also common in the literature to see the opposite signature, corresponding to the pattern $(+---)$ that results from multiplying the metric above by -1.
- This choice is *conventional* (no physics depends on it). Metrics with the signature $(-+++)$ or $(+---)$ are sometimes said to be *Lorentzian.*
- However, it is an *essential property of Minkowski space* that it is *not possible to have the same sign* for all terms in the signature of the metric.

The Minkowski metric is indefinite, in contrast to the positive definite metric of euclidean spaces.

4.1.1 Invariance of the Spacetime Interval

Special relativity follows from two assumptions:

- The speed of light is constant for all observers.
- The laws of physics can't depend on coordinates.

The postulate that the speed of light is a constant is equivalent to a statement that

> The spacetime interval ds^2 is an invariant that is *unchanged by transformations between inertial systems* (the Lorentz transformations; see below).

- This invariance does not hold for the euclidean spatial interval $dx^2 + dy^2 + dz^2$,
- nor does it hold for the time interval c^2dt^2.
- Only the particular combination of spatial and time intervals defined by

$$ds^2 = \underbrace{\overbrace{-c^2dt^2}^{\text{not invariant}} \overbrace{+dx^2+dy^2+dz^2}^{\text{not invariant}}}_{\text{invariant}}$$

  is (Lorentz) invariant.

> Because of this invariance, *Minkowski space is the natural manifold* for special relativity.

Example 4.2

Let's use the metric to determine the relationship between the time coordinate t and the proper time τ, with $\tau^2 \equiv -s^2/c^2$. From

$$ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2,$$

we may write

$$\begin{aligned} d\tau^2 &= \frac{-ds^2}{c^2} = \frac{1}{c^2}(c^2dt^2 - dx^2 - dy^2 - dz^2) \\ &= dt^2 \left\{ 1 - \frac{1}{c^2} \underbrace{\left[\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2 \right]}_{v^2} \right\} \\ &= \left(1 - \frac{v^2}{c^2}\right) dt^2. \end{aligned}$$

where v is the magnitude of the ordinary 3-velocity. Therefore, the *proper time* τ that elapses between *coordinate times* t_1 and t_2 is

$$\tau_{12} = \int_{t_1}^{t_2} \left(1 - \frac{v^2}{c^2}\right)^{1/2} dt.$$

The proper time interval τ_{12} *is shorter than the coordinate time interval* $t_2 - t_1$ because the square root is always less than one. If the velocity is constant, this reduces to

$$\Delta\tau = \left(1 - \frac{v^2}{c^2}\right)^{1/2} \Delta t,$$

which is the usual statement of *time dilation in special relativity*.

Table 4.1: Rank 0, 1, and 2 tensor transformation laws

Tensor	Transformation law
Scalar	$\varphi' = \varphi$
Covariant vector	$A'_\mu = \frac{\partial x^\nu}{\partial x'^\mu} A_\nu$
Contravariant vector	$A'^\mu = \frac{\partial x'^\mu}{\partial x^\nu} A^\nu$
Covariant rank-2	$T'_{\mu\nu} = \frac{\partial x^\alpha}{\partial x'^\mu}\frac{\partial x^\beta}{\partial x'^\nu} T_{\alpha\beta}$
Contravariant rank-2	$T'^{\mu\nu} = \frac{\partial x'^\mu}{\partial x^\alpha}\frac{\partial x'^\nu}{\partial x^\beta} T^{\alpha\beta}$
Mixed rank-2	$T'^\nu{}_\mu = \frac{\partial x^\alpha}{\partial x'^\mu}\frac{\partial x'^\nu}{\partial x^\beta} T^\beta{}_\alpha$

4.2 Tensors in Minkowski Space

In Minkowski space *transformations between coordinate systems are independent of spacetime.* Thus derivatives appearing in the general definitions of Table 4.1 for tensors are constants and for flat spacetime we have the simplified tensor transformation laws

$\varphi' = \varphi$	Scalar
$A'^\mu = \Lambda^\mu{}_\nu A^\nu$	Contravariant vector
$A'_\mu = \Lambda_\mu{}^\nu A_\nu$	Covariant vector
$T'^{\mu\nu} = \Lambda^\mu{}_\gamma \Lambda^\nu{}_\delta T^{\gamma\delta}$	Contravariant rank-2 tensor
$T'_{\mu\nu} = \Lambda_\mu{}^\gamma \Lambda_\nu{}^\delta T_{\gamma\delta}$	Covariant rank-2 tensor
$T'^\mu{}_\nu = \Lambda^\mu{}_\gamma \Lambda_\nu{}^\delta T^\gamma{}_\delta$	Mixed rank-2 tensor

where the $\Lambda^\mu{}_\nu$ *don't depend on the spacetime coordinates*.

In addition, for flat spacetime we may use a coordinate system for which the second term of

$$A'_{\mu,\nu} = \underbrace{A_{\alpha,\beta}\frac{\partial x^\beta}{\partial x'^\nu}\frac{\partial x^\alpha}{\partial x'^\mu}}_{\textit{Tensor}} + \underbrace{A_\alpha \frac{\partial^2 x^\alpha}{\partial x'^\nu \partial x'^\mu}}_{\textit{Not a tensor}}$$

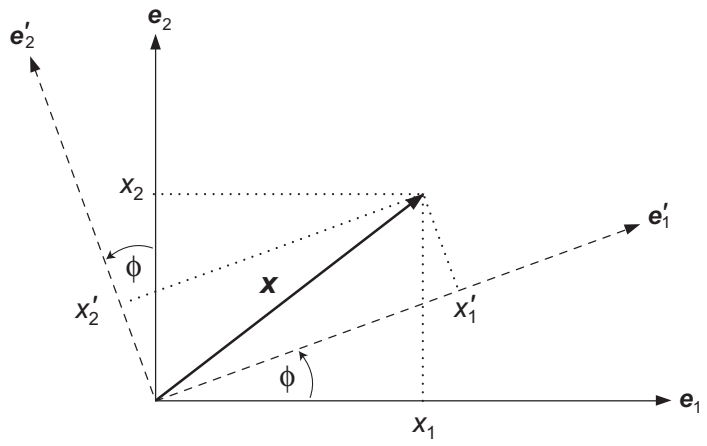
can be transformed away and *in flat spacetime covariant derivatives are equivalent to partial derivatives.*

In the Minkowski transformation laws the $\Lambda^\mu{}_\nu$ are elements of *Lorentz transformations,* to which we now turn our attention.

4.3 Lorentz Transformations

In 3-dimensional euclidean space, rotations are a particularly important class of transformations because they *change the direction for a 3-vector but preserve its length*.

- We wish to generalize this idea to investigate abstract rotations in the 4-dimensional Minkowski space that *change the direction but not the length of 4-vectors*.
- Such rotations in Minkowski space are termed *Lorentz transformations.*



Consider rotation of a 2D euclidean coordinate system (above)

- The length of an arbitrary vector $\boldsymbol{x}$ will be unchanged by this transformation if we require that $\boldsymbol{x}\cdot\boldsymbol{x} = \boldsymbol{x}'\cdot\boldsymbol{x}'$.
- Since $\boldsymbol{x}\cdot\boldsymbol{x} \equiv g_{ij}x^i x^j$, this requires that the transformation matrix R implementing the rotation $x'^i = R^i_{\ j}x^j$ act on the metric tensor g_{ij} in the following way

$$Rg_{ij}R^{\mathrm{T}} = g_{ij},$$

  where R^{T} is the transpose of R (switch rows and columns).
- For euclidean space the metric tensor is just the unit matrix so the above requirement reduces to $RR^{\mathrm{T}} = 1$, which is the condition that R be an *orthogonal matrix*.

Thus, we obtain by this somewhat pedantic route the well-known result that *rotations in euclidean space are implemented by orthogonal matrices*.

- But the requirement $Rg_{ij}R^{\mathrm{T}} = g_{ij}$ is *valid generally*, not just for euclidean spaces. Thus, let's use it as guidance to constructing *generalized rotations in Minkowski space*.
- By analogy with the above discussion of rotations in euclidean space, we seek a set of transformations that *leave the length of a 4-vector invariant* in the Minkowski space.
- We write the coordinate transformation in matrix form,

$$dx'^{\mu} = \Lambda^{\mu}{}_{\nu} dx^{\nu},$$

  where we expect the transformation matrix $\Lambda^{\mu}{}_{\nu}$ to satisfy the analog of $Rg_{ij}R^{\mathrm{T}} = g_{ij}$ for the Minkowski metric $\eta_{\mu\nu}$,

$$\Lambda\eta_{\mu\nu}\Lambda^{\mathrm{T}} = \eta_{\mu\nu},$$

  or explicitly in terms of matrix components,

$$\Lambda_{\mu}{}^{\rho}\Lambda^{\sigma}{}_{\nu}\eta_{\rho\sigma} = \eta_{\mu\nu}.$$

- Let us now use this property to construct the elements of the transformation matrix $\Lambda^{\mu}{}_{\nu}$. These will include
  - *rotations about the spatial axes* (corresponding to rotations within inertial systems) and
  - *transformations between inertial systems* moving at different constant velocities (*Lorentz boosts*).

  We consider first *rotations about the z axis*.

4.3.1 Rotations

For rotations about the z axis

$$\begin{pmatrix} x'^1 \\ x'^2 \end{pmatrix} = R \begin{pmatrix} x^1 \\ x^2 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \end{pmatrix},$$

where a, b, c, and d parameterize the transformation matrix.

- Rotations about a single axis are a 2D problem with euclidean metric, so the condition $Rg_{ij}R^{\mathrm{T}} = g_{ij}$ is

$$\underbrace{\begin{pmatrix} a & b \\ c & d \end{pmatrix}}_{R} \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}_{g_{ij}} \underbrace{\begin{pmatrix} a & c \\ b & d \end{pmatrix}}_{R^{\mathrm{T}}} = \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}_{g_{ij}},$$

- Carrying out the matrix multiplications gives

$$\begin{pmatrix} a^2+b^2 & ac+bd \\ ac+bd & c^2+d^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

  and comparing the two sides of the equation implies that

$$a^2+b^2 = 1 \qquad c^2+d^2 = 1 \qquad ac+bd = 0.$$

- These requirements are satisfied by the choices

$$a = \cos\varphi \qquad b = \sin\varphi \qquad c = -\sin\varphi \qquad d = \cos\varphi,$$

  and we obtain the expected result for an ordinary rotation,

$$\begin{pmatrix} x'^1 \\ x'^2 \end{pmatrix} = R \begin{pmatrix} x^1 \\ x^2 \end{pmatrix} = \begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \end{pmatrix}.$$

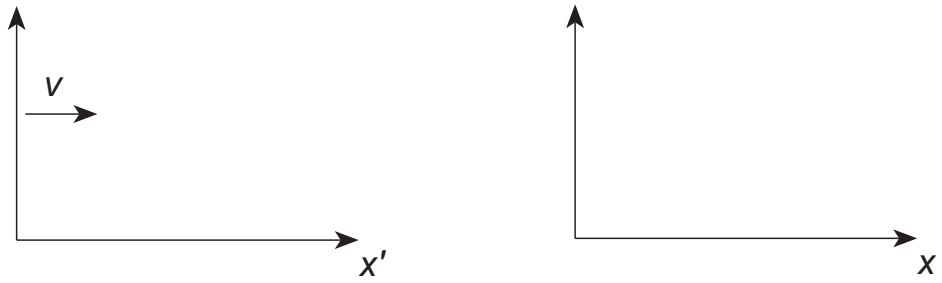


Figure 4.1: A Lorentz boost along the positive x axis.

Now, let's apply this same technique to determine the elements of a *Lorentz boost transformation*.

4.3.2 Lorentz Boosts

Consider a boost from one inertial system to a 2nd one moving at uniform velocity along the x axis (Fig. 4.1).

- The y and z coordinates are unaffected, so this also is effectively a 2-dimensional transformation on t and x,

$$\begin{pmatrix} cdt' \\ dx' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} cdt \\ dx \end{pmatrix}$$

- We can write the condition $\Lambda\eta_{\mu\nu}\Lambda^{\mathrm{T}} = \eta_{\mu\nu}$ out as

$$\underbrace{\begin{pmatrix} a & b \\ c & d \end{pmatrix}}_{\Lambda} \underbrace{\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}}_{\eta_{\mu\nu}} \underbrace{\begin{pmatrix} a & c \\ b & d \end{pmatrix}}_{\Lambda^{\mathrm{T}}} = \underbrace{\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}}_{\eta_{\mu\nu}},$$

(identical to rotations, except for the *indefinite metric*).

- Multiplying the matrices on the left side and comparing with the matrix on the right side in

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & c \\ b & d \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix},$$

gives the conditions

$$a^2 - b^2 = 1 \qquad -c^2 + d^2 = 1 \qquad -ac + bd = 0,$$

- These are satisfied if we choose

$$a = \cosh\xi \qquad b = \sinh\xi \qquad c = \sinh\xi \qquad d = \cosh\xi,$$

where ξ is a hyperbolic variable with $-\infty \leq \xi \leq \infty$.

- Therefore, the boost transformation may be written as

$$\begin{pmatrix} cdt' \\ dx' \end{pmatrix} = \begin{pmatrix} \cosh\xi & \sinh\xi \\ \sinh\xi & \cosh\xi \end{pmatrix} \begin{pmatrix} cdt \\ dx \end{pmatrix} \qquad \text{(Lorentz boost).}$$

Which may be compared with the rotational result

$$\begin{pmatrix} x'^1 \\ x'^2 \end{pmatrix} = \begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \end{pmatrix} \qquad \text{(Spatial rotation).}$$

The respective derivations make clear that

- the appearance of hyperbolic functions in the boosts, rather than trigonometric functions as in rotations, traces to the role of the indefinite metric

$$g_{\mu\nu} = \text{diag}\,(-1,1)$$

  in the boosts.
- The hyperbolic functions suggest that the boost transformations are *"rotations" in Minkowski space*.
- But these rotations
  - *mix space and time*, and
  - will have unusual properties since they correspond to *rotations through imaginary angles*.
- These unusual properties follow from the metric:
  - The conserved invariant interval is not the length of spatial vectors or time intervals separately.
  - Rather it is the specific *mixture of time and space intervals* implied by the Minkowski line element with *indefinite metric*:

$$ds^2 = (cdt \;\; dx \;\; dy \;\; dz) \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} cdt \\ dx \\ dy \\ dz \end{pmatrix}.$$

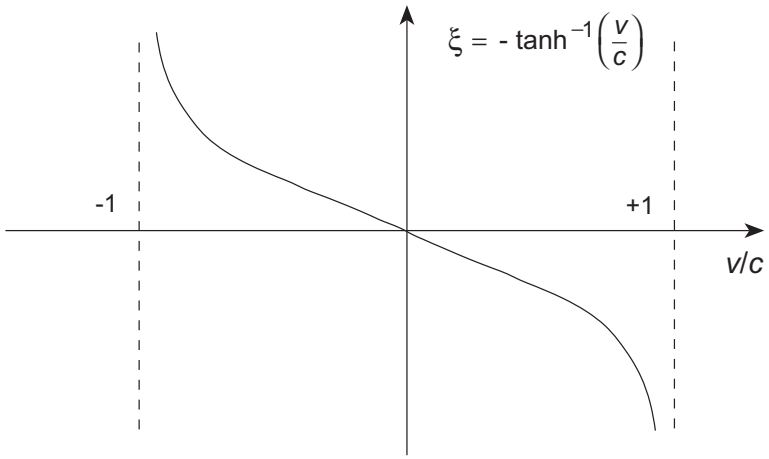


Figure 4.2: Dependence of the Lorentz parameter ξ on $\beta = v/c$.

We can put the Lorentz boost transformation into a more familiar form by relating the boost parameter ξ to the boost velocity.

- Let's work with finite space and time intervals by replacing $dt \to t$ and $dx \to x$ in the preceding equations.
- The velocity of the boosted system is $v = x/t$. From

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \begin{pmatrix} \cosh\xi & \sinh\xi \\ \sinh\xi & \cosh\xi \end{pmatrix} \begin{pmatrix} ct \\ x \end{pmatrix},$$

  the origin ($x' = 0$) of the boosted system is

$$x' = ct\sinh\xi + x\cosh\xi = 0 \quad \to \quad x\cosh\xi = -ct\sinh\xi.$$

- Therefore, $x/t = -c\sinh\xi/\cosh\xi$, so

$$\beta \equiv \frac{v}{c} = \frac{x}{ct} = -\frac{\sinh\xi}{\cosh\xi} = -\tanh\xi.$$

- This relationship between ξ and β is plotted in Fig. 4.2.

- Using the identity $1 = \cosh^2\xi - \sinh^2\xi$, and the definition

$$\gamma \equiv \left(1 - \frac{v^2}{c^2}\right)^{-1/2} = \frac{1}{\sqrt{1 - v^2/c^2}}$$

of the *Lorentz γ factor,* we may write

$$\begin{aligned}\cosh\xi &= \sqrt{\frac{\cosh^2\xi}{1}} = \sqrt{\frac{\cosh^2\xi}{\cosh^2\xi - \sinh^2\xi}} \\ &= \sqrt{\frac{1}{1 - \sinh^2\xi/\cosh^2\xi}} = \frac{1}{\sqrt{1-\beta^2}} \\ &= \frac{1}{\sqrt{1 - v^2/c^2}} = \gamma,\end{aligned}$$

- From this result and

$$\beta = -\frac{\sinh\xi}{\cosh\xi}$$

we obtain

$$\sinh\xi = -\beta\cosh\xi = -\beta\gamma.$$

- Thus, inserting $\cosh\xi = \gamma$ and $\sinh\xi = -\beta\gamma$ in the Lorentz transformation for finite intervals gives

$$\begin{aligned}\begin{pmatrix} ct' \\ x' \end{pmatrix} &= \begin{pmatrix} \cosh\xi & \sinh\xi \\ \sinh\xi & \cosh\xi \end{pmatrix}\begin{pmatrix} ct \\ x \end{pmatrix} \\ &= \begin{pmatrix} \gamma & -\gamma\beta \\ -\gamma\beta & \gamma \end{pmatrix}\begin{pmatrix} ct \\ x \end{pmatrix} = \gamma\begin{pmatrix} 1 & -\beta \\ -\beta & 1 \end{pmatrix}\begin{pmatrix} ct \\ x \end{pmatrix}.\end{aligned}$$

- Writing the matrix expression

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \gamma \begin{pmatrix} 1 & -\beta \\ -\beta & 1 \end{pmatrix} \begin{pmatrix} ct \\ x \end{pmatrix}.$$

  out explicitly for finite intervals gives the Lorentz boost equations (for the specific case of a positive boost along the x axis) in standard textbook form,

$$\begin{aligned} t' &= \gamma\left(t - \frac{vx}{c^2}\right) \\ x' &= \gamma(x - vt) \\ y' &= y \qquad z' = z. \end{aligned}$$

- The inverse transformation corresponds to the replacement $v \to -v$.

- Clearly these reduce to the Galilean boost equations

$$\boldsymbol{x}' = \boldsymbol{x}'(\boldsymbol{x}, t) = \boldsymbol{x} - \boldsymbol{v}t \qquad t' = t'(\boldsymbol{x}, t) = t.$$

  in the limit that v/c vanishes (so $\gamma \to 1$), as we would expect.

It is easily verified (Problem) that *Lorentz transformations leave invariant the spacetime interval* ds^2.

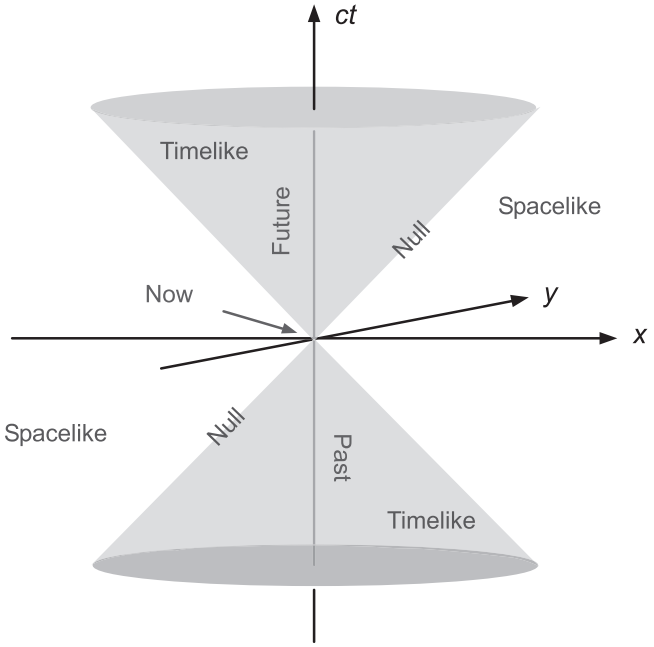


Figure 4.3: The lightcone diagram for two space and one time dimensions.

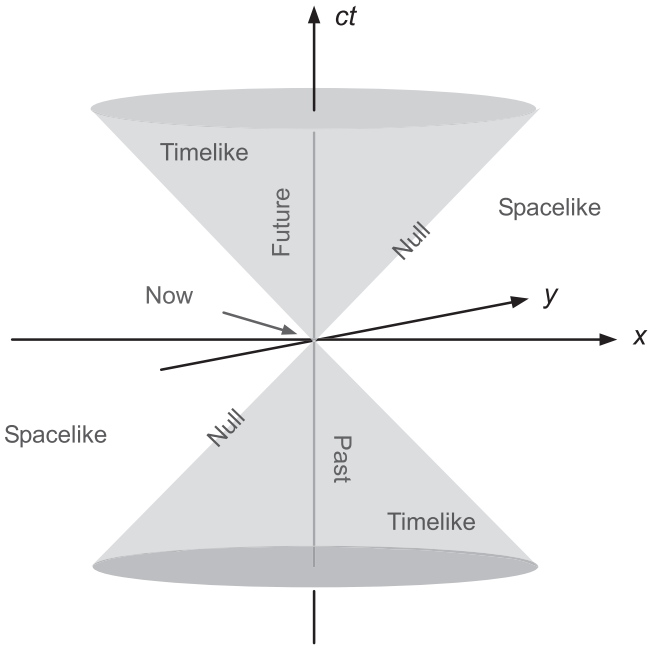
4.4 Lightcone Diagrams

By virtue of the line element (which defines a cone)

$$ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2,$$

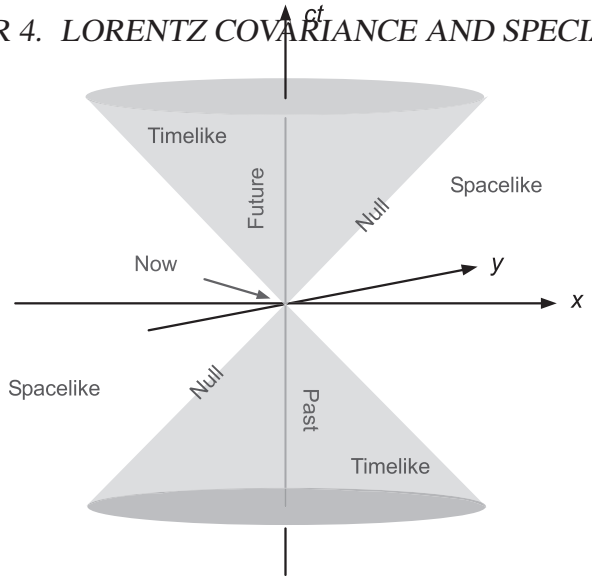
the Minkowski spacetime may be classified according to the lightcone diagram exhibited in Fig. 4.3.

> The lightcone is a 3D surface in 4D spacetime and events and intervals in spacetime may be characterized according to whether they lie inside of, outside of, or on the lightcone.



The standard terminology [assuming our $(-1,1,1,1)$ metric signature]:

- If $ds^2 < 0$ the interval is termed *timelike*.
- If $ds^2 > 0$ the interval is termed *spacelike*.
- If $ds^2 = 0$ the interval is called *lightlike* (or sometimes *null*).



The lightcone clarifies the distinction between Minkowski spacetime and a 4D euclidean space:

- Two points in Minkowski spacetime are separated by the interval ds defined through

$$ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2.$$

This interval could be

- positive,
- negative, or
- zero,

which embodies *impossibilities for a euclidean space*.

In particular, lightlike particles have worldlines confined to the lightcone and *the square of the separation of any two points on a lightlike worldline is ZERO.*

Example 4.3

The Minkowski line element in one space and one time dimension [often termed $(1+1)$ dimensions] is $ds^2 = -c^2dt^2 + dx^2$. Thus, if $ds^2 = 0$

$$-c^2dt^2 + dx^2 = 0 \quad \longrightarrow \quad \underbrace{\left(\frac{dx}{dt}\right)^2}_{v^2} = c^2 \quad \longrightarrow \quad v = \pm c.$$

We can generalize this result easily to the full 4D spacetime and we conclude that

- Events in Minkowski space separated by a null interval ($ds^2 = 0$) are connected by signals moving at light velocity, $v = c$.
- If the time (ct) and space axes have the same scales, the world-line of a freely propagating photon (or any massless particle) always make $\pm 45^\circ$ angles in the lightcone diagram.
- Events at timelike separations (inside the lightcone) are connected by signals with $v < c$, and
- Those with spacelike separations (outside the lightcone) could be connected only by signals with $v > c$ (which would violate causality).

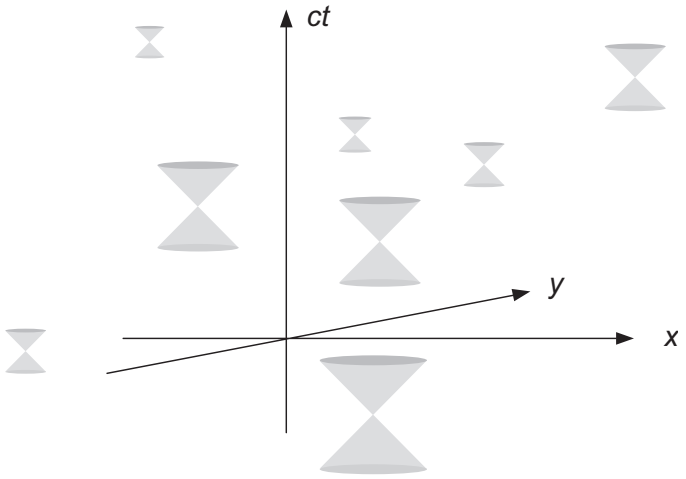


Figure 4.4: Each point of spacetime has its own lightcone.

We have placed the lightcone in the earlier illustration at the origin of our coordinate system, but in general we may imagine a lightcone attached to every point in the spacetime, as illustrated in Fig. 4.4.

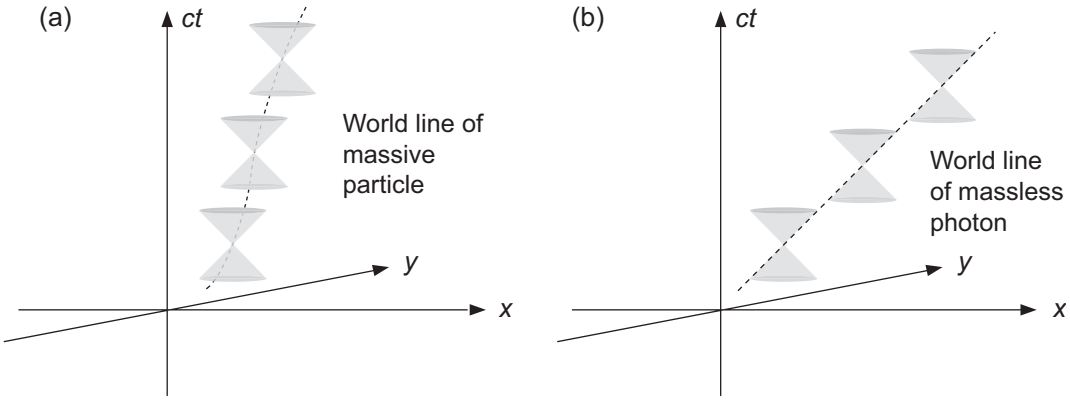


Figure 4.5: Worldlines for massive particles and for massless particles such as photons.

A tangent to the worldline of any particle defines the local velocity of the particle and constant velocity implies straight worldlines. Therefore, as illustrated in Fig. 4.5,

- Light must always travel in straight lines (in Minkowski space; *not* in curved space), and always on the lightcone, since $v = c =$ constant.
- Thus photons have *constant local velocities* equal to c.
- Worldlines for any massive particle lie inside the local lightcone since $v \leq c$ (*timelike trajectory*, since always within the lightcone).
- The worldline for the massive particle in this particular example is curved *(acceleration)*.
- For non-accelerated massive particles the worldline would be straight, but always *within the lightcone.*

Invariance and Simultaneity

- In Galilean relativity, an event picks out a *hyperplane of simultaneity* in the spacetime diagram consisting of all events occurring at the same time as the event.
- All observers agree on what constitutes this set of simultaneous events because *Galilean relativity of simultaneity is independent of the observer*.
- *In Einstein's relativity, simultaneity depends on the observer* and *hyperplanes of constant coordinate time have no invariant meaning*.
- However, all observers agree on the classification of events relative to local lightcones, because *the speed of light is invariant for all observers*.

As we shall now discuss, *The local lightcones define an invariant spacetime structure that may be used to classify events.*

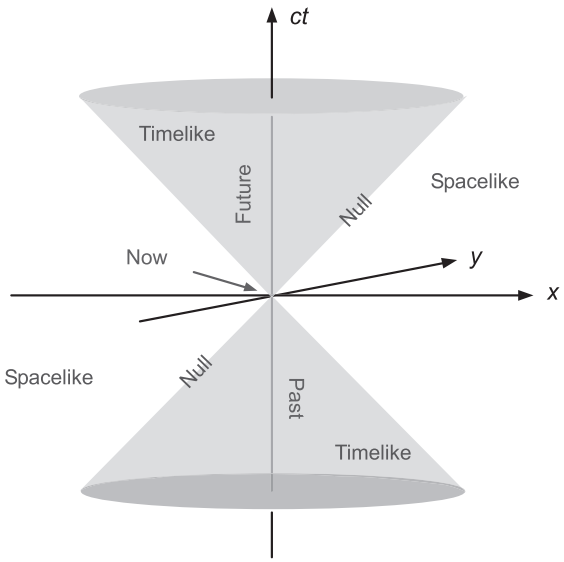


Figure 4.6: The lightcone diagram for two space and one time dimensions.

4.5 Causal Structure of Spacetime

The causal properties of Minkowski spacetime are encoded in its light cone structure, which requires that $v \leq c$ for all signals.

- Each point in spacetime may be viewed as lying at the apex of a lightcone (*"Now"*).
- An event at the origin of a lightcone may influence any event in its forward lightcone (*the "Future"*).
- The event at the origin of the lightcone may be influenced by events in its backward lightcone (*the "Past"*).
- Events at spacelike separations are causally disconnected from the event at the origin.
- Events on the lightcone are connected by $v = c$ signals.

The lightcone is a *surface separating the knowable from the unknowable* for an observer at the apex of the lightcone.

This lightcone structure of spacetime ensures that all velocities obey locally the constraint $v \leq c$.

- Velocities are *defined and measured locally*.
- Hence covariant field theories in either flat or curved space are *guaranteed to respect the speed limit* $v \leq c$.
- This is true irrespective of whether globally velocities appear to exceed c.

EXAMPLE: In the Hubble expansion of the Universe,

- Galaxies beyond a certain distance (the horizon) would *appear* to recede from us at velocities in excess of c.
- However, all *local measurements* in that expanding, possibly curved, space would determine the velocity of light to be c.

4.5.1 Time Machines and Causality Paradoxes

When time travel comes up it is usually about going *backward* in time.

- Traveling *forward* in time requires no special talent.
- It is easy to arrange various scenarios consistent with relativity where a person could travel into a future time even faster than normal.
- For example, in the *twin paradox* discussed later it is possible to arrange for a traveler to arrive back at Earth centuries in the future relative to clocks that remain on Earth.
- Similar options exist using the *gravitational time dilation* to be described later chapters.

However, the real question is, could you go *back in time* to explore your earlier history?

- No! Not according to current understanding.
- To bend a forward-going timelike worldline continuously into a backward-going one requires *going outside the local lightcone*, requiring that $v > c$.
- If *closed timelike loops* were permitted, travel to earlier times might be possible.
- However, they are forbidden if energy densities are never negative and the Universe has the topology in evidence.

Thus, the determined time traveler has two options:

- Find some negative energy, or
- Find structures with an exotic spacetime topology allowing closed timelike loops.

Unfortunately for the aspiring time traveler,

- negative energy is probably forbidden in classical gravity and
- there is no evidence at present for exotic spacetime topologies with closed timelike loops.

  - These statements are based entirely on *classical gravity considerations;*
  - it is unknown at present whether they could be modified by some future understanding of *quantum gravity*.

From the preceding discussion we may conclude that

- the axioms of special relativity are fundamentally at odds with the Newtonian *concept of absolute simultaneity*, since
- the demand that light have the *same speed for all observers* necessarily means that
- the apparent temporal order of two events *depends upon the observer*.

However, the abolishment of absolute simultaneity

- introduces *no causal ambiguity* because
- all observers agree on the *lightcone structure of spacetime*.
- Thus, for example, all observers will agree that

> Event A can cause event B *only* if A lies in the *past lightcone* of B.

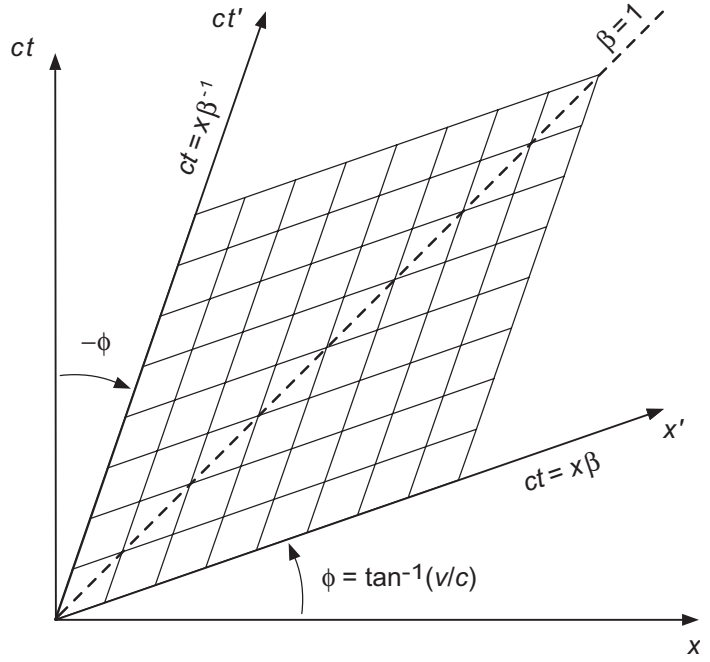
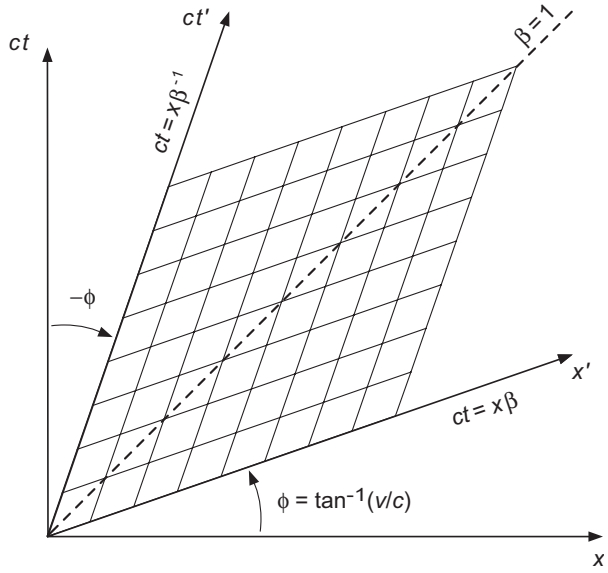


Figure 4.7: Lorentz boost transformation in a spacetime diagram.

4.6 Lorentz Transformations in Spacetime Diagrams

It is instructive to look at the action of Lorentz transformations in the spacetime (lightcone) diagram. If we consider boosts only in the x direction, the relevant part of the spacetime diagram in some inertial frame corresponds to a plot with axes ct and x, as in the figure above.



Let's ask what happens to these axes under the Lorentz boost

$$ct' = c\gamma\left(t - \frac{vx}{c^2}\right) \qquad x' = \gamma(x - vt).$$

- The t' axis corresponds to $x' = 0$. From the 2nd equation

$$x = vt \longrightarrow x/c = (v/c)t = \beta t,$$

  so the equation for the y' axis is $ct = x\beta^{-1}$, with $\beta = v/c$.

- Likewise, the x' axis corresponds to $t' = 0$, which implies from the 1st equation that $ct = (v/c)x = x\beta$.

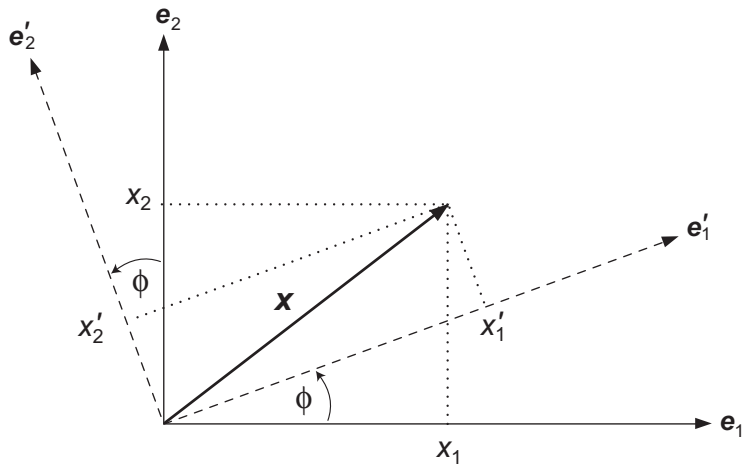
- Thus, the equations of the x' and t' axes [in the (x, ct) coordinate system] are $ct = x\beta$ and $ct = x\beta^{-1}$, respectively.

- The x' and t' axes for the boosted system are also shown in the figure for a positive value of β.

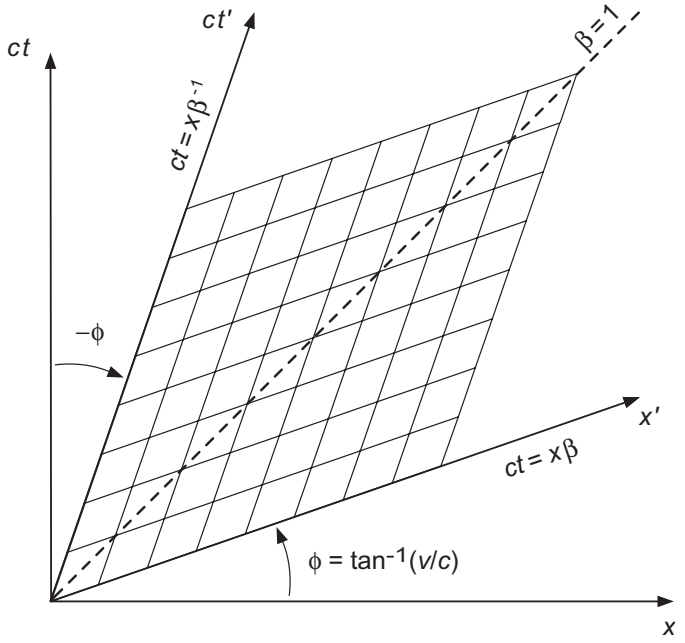
- Time and space axes rotated by same angle, but in *opposite directions* (Cause: the *indefinite Minkowski metric*).

- The rotation angle is given by $\tan\varphi = v/c$.

Ordinary rotations: the two axes rotate by the same angle in the *same direction*



Lorentz boost "rotations": the two axes rotate by the same angle but in *opposite directions*



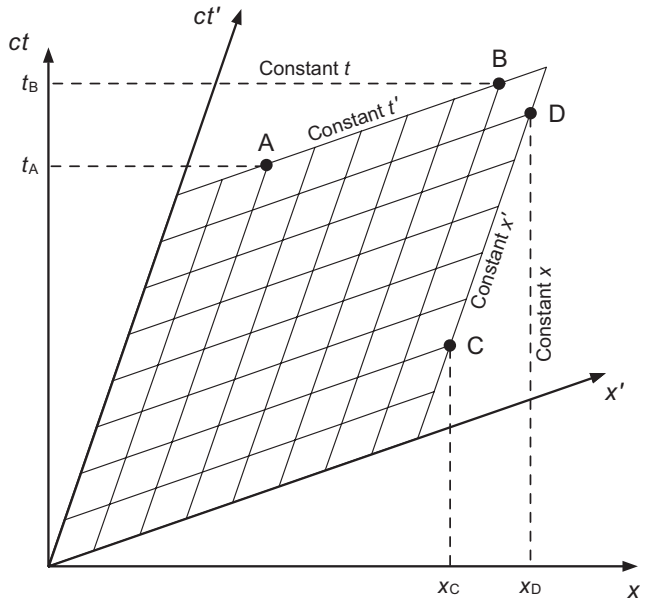


Figure 4.8: Comparison of events in boosted and unboosted frames.

Most of special relativity follows from this figure.

- For example, relativity of simultaneity is illustrated above.
- Points A and B lie on the same t' line, so they are *simultaneous in the boosted frame*.
- But from the dashed projections on the ct axis, *event A occurs before event B in the unboosted frame*.
- Likewise, points C and D lie at the same value of x' in the boosted frame and so are *spatially congruent*, but in the unboosted frame $x_C \neq x_D$.

Relativistic time dilation and space contraction follow rather directly from these observations.

Example 4.4

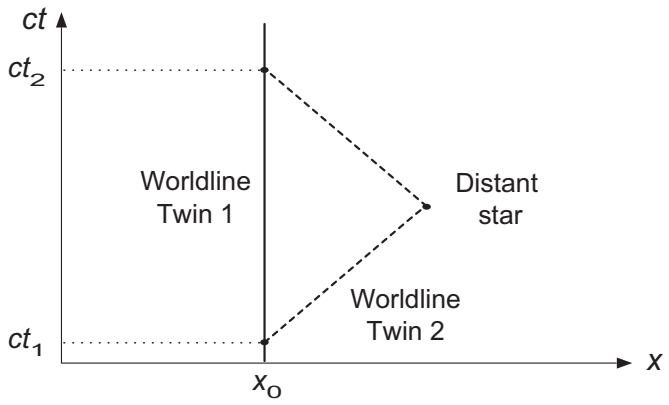
The time registered by a clock moving between two points depends on the path followed, as suggested by the time-dilation formula.

$$d\tau^2 = \left(1 - \frac{v^2}{c^2}\right) dt^2.$$

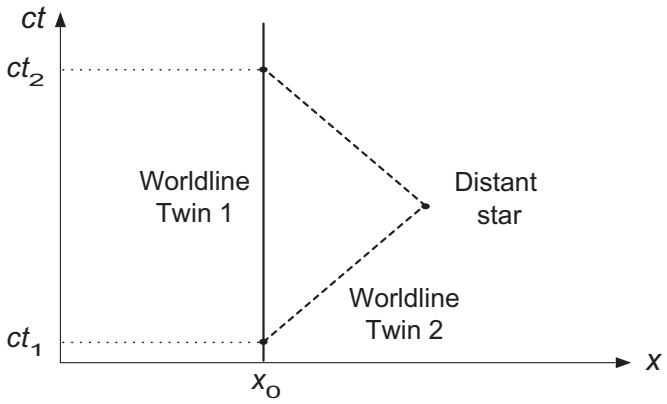
> The proper time τ is the time registered by a clock carried by an observer on a spacetime path.

That this is true even if the path returns to the initial spatial position is the source of the *twin paradox* of special relativity.

- Twins are initially at rest in the same inertial frame.
- Twin 2 travels at $v \sim c$ to a distant star and then returns at the same speed to the starting point.
- Twin 1 remains at the starting point.
- The relevant spacetime paths are:



- The elapsed time on the clock carried by Twin 2 is *always smaller* than that for the clock carried by Twin 1 (see above equation).



- The (seeming) paradox arises if one describes things from the point of view of Twin 2, who sees Twin 1 move away and then back.
- This seems to be symmetric with the case of Twin 1 watching Twin 2 move away and then back.

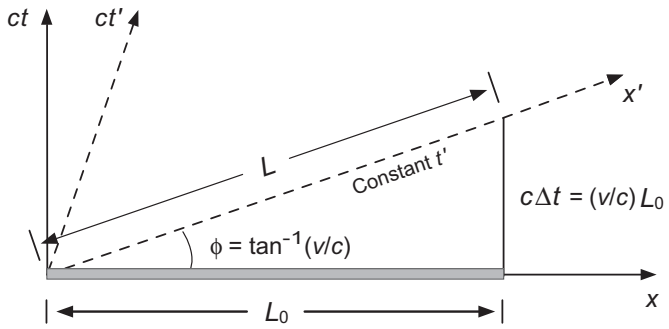
> But it isn't: *the twins travel different worldlines*, and *different distances along these worldlines.*

- For example, *Twin 2 experiences accelerations but Twin 1 does not*, so their worldlines cannot be equivalent.
- Their clocks record the proper time on their respective worldlines and thus differ when they are rejoined.
- This indicates unambiguously that Twin 2 is younger at the end of the journey.

> When properly analyzed, there is *no paradox* in the "twin paradox".

Space Contraction

Consider a rod of *proper length* L_0, as measured in its own rest frame (ct, x), that is oriented along the x axis.

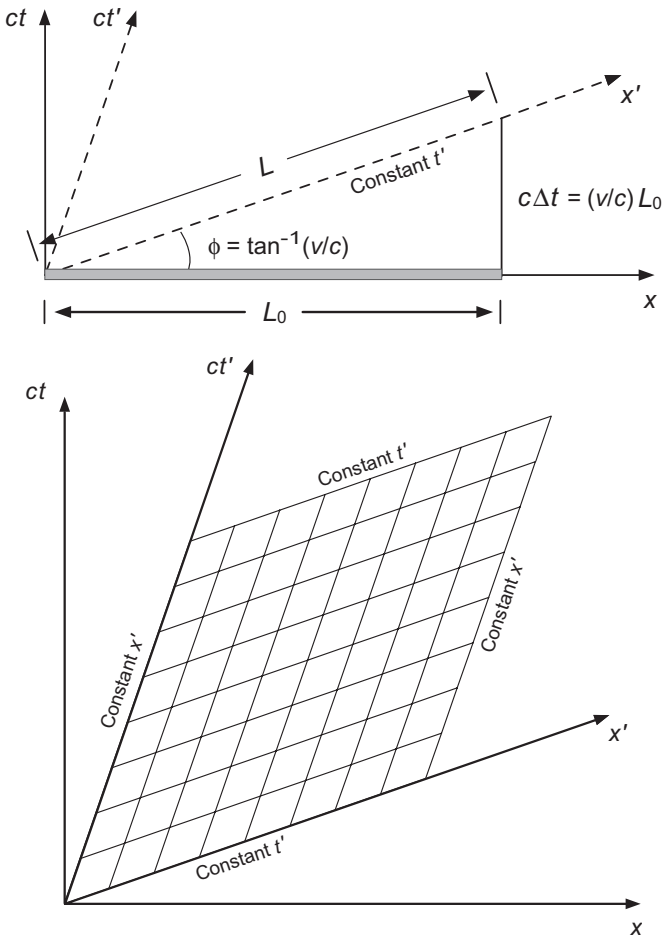


Note: *"Proper"* in relativity denotes a quantity *measured in the rest frame of an object* (proper time, proper length, ...).

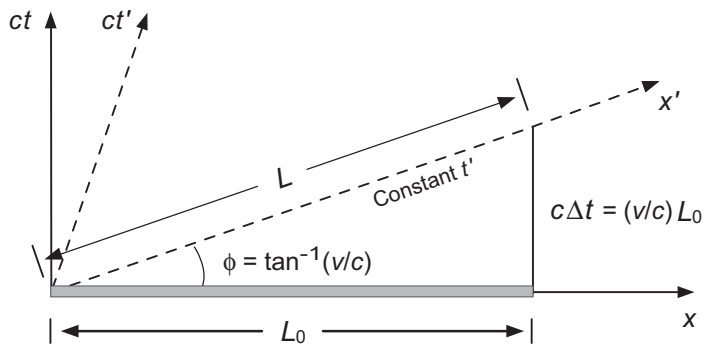
What is the length of the rod L as observed in the *boosted* (ct', x') frame? *Fundamental measurement issues:*

- *Distances* must be measured between spacetime points *at the same time.*
- *Elapsed times* must be measured at spacetime points *at the same place.*

Example: Length of an arrow in flight is not given by the difference between the location of its tip at one time and its tail at a different time. The measurements must be made *at the same time*.



- The frame (ct', x') is boosted by a velocity v along the positive x axis relative to the (ct, x) frame.
- Therefore, in the primed frame the rod will have a velocity v in the negative x' direction.
- Determining the length L observed in the primed frame requires that the positions of the ends of the rod be measured *simultaneously in that frame.*
- The axis labeled x' has constant t' (bottom figure), so the distance L (top figure) is the *length in the primed frame*.

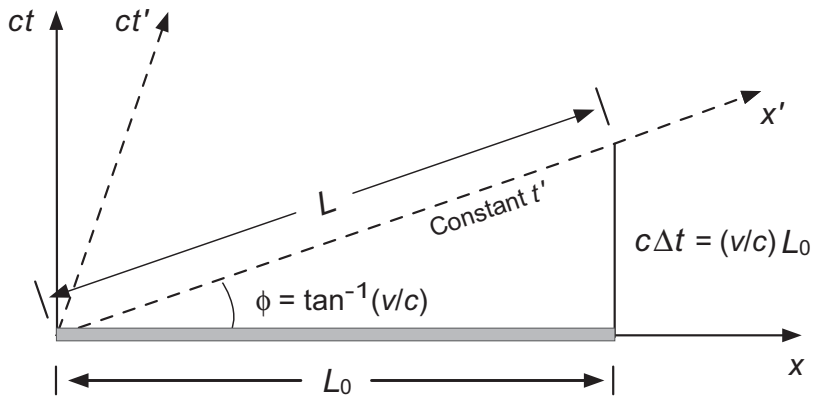


- The distance L *seems* longer than L_0, but this is deceiving because *the diagram is in Minkowski spacetime* but our brain has a euclidean-space bias.
- We are familiar with perceived distances being different from actual distances from flat map projections.

Map Projections

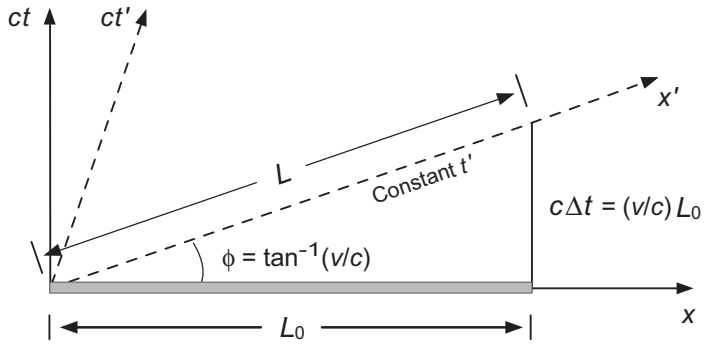
Mercator
(preserves angles, distorts sizes)

Source: *http://www.culturaldetective.com/worldmaps.html*



- A Mercator projection of the globe onto a euclidean sheet of paper gives misleading distance information—Greenland isn't really larger than Brazil, for example,

We must *always trust the metric* to determine the correct distance in any space.



- From the Minkowski indefinite-metric line element

$$ds^2 = -c^2dt^2 + dx^2.$$

and from the triangle *(Minkowski Pythagorean theorem),*

$$L^2 = L_0^2 - (c\Delta t)^2.$$

- But $c\Delta t = (v/c)L_0$ (because from the diagram $\tan\varphi = c\Delta t/L_0$ and $\tan\varphi = v/c$). Therefore,

$$\begin{aligned} L &= (L_0^2 - (c\Delta t)^2)^{1/2} \\ &= \left(L_0^2 - \left(\frac{v}{c}L_0\right)^2\right)^{1/2} \\ &= L_0\left(1 - \frac{v^2}{c^2}\right)^{1/2} \end{aligned}$$

- L is *shorter than the proper length* L_0, even though it seems longer in the figure. *TRUST THE METRIC!*

> But this is just *special relativistic length-contraction,* which is seen to be nothing more than the *Pythagorean theorem for Minkowski space*.

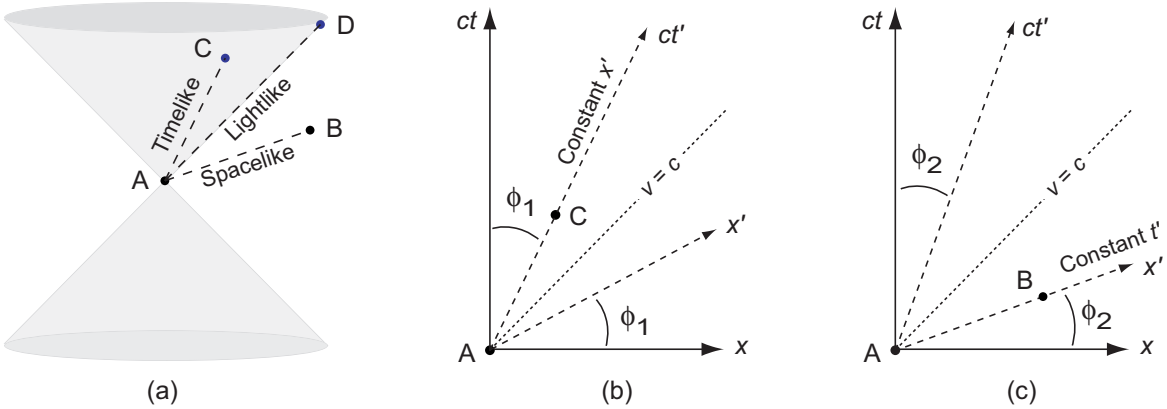
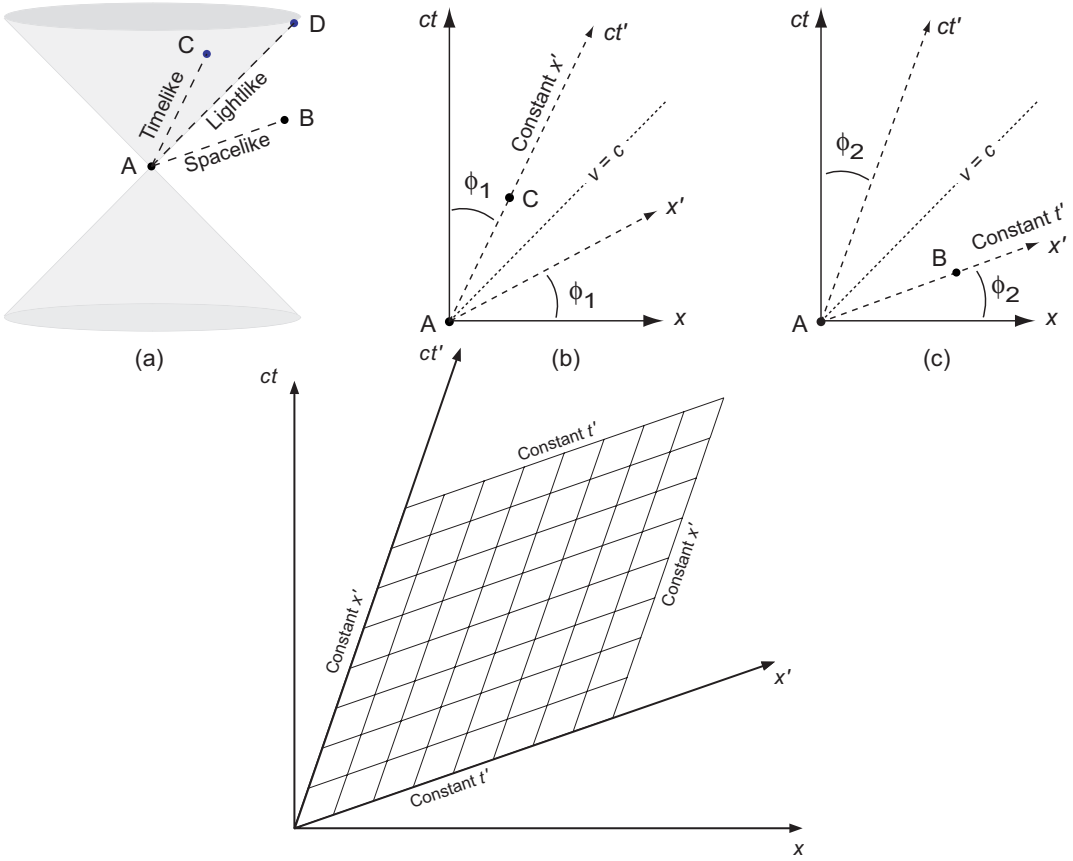


Figure 4.9: (a) Timelike, lightlike (null), and spacelike separations. (b) Lorentz transformation that brings the timelike separated points A and C of (a) into spatial congruence (they lie along a line of constant x' in the primed system). (c) Lorentz transformation that brings the spacelike separated points A and B of (a) into coincidence in time (they lie along a line of constant t' in the primed system.

As we have seen, the spacetime separation between any two events (*spacetime interval*) may be classified in a relativistically invariant way as

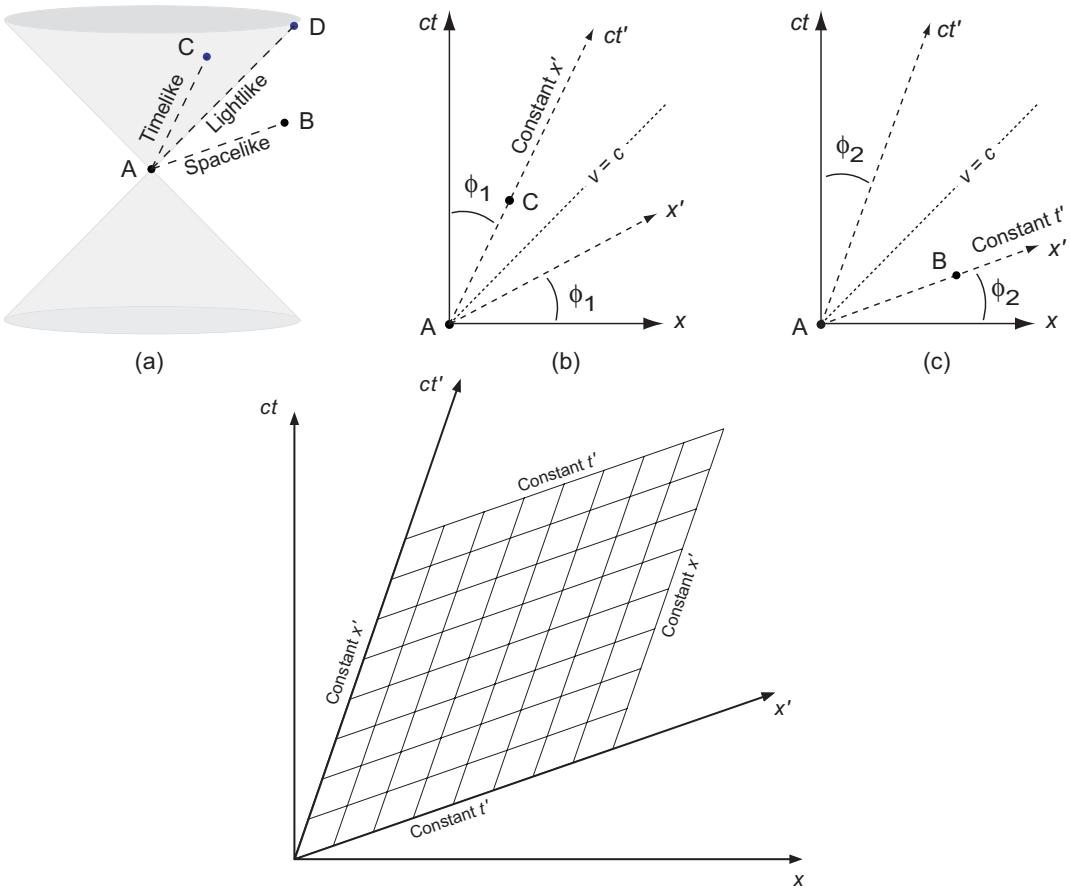
1. *timelike*,
2. *lightlike*, or
3. *spacelike*

by constructing the lightcone at one of the points, as illustrated in Fig. 4.9(a).



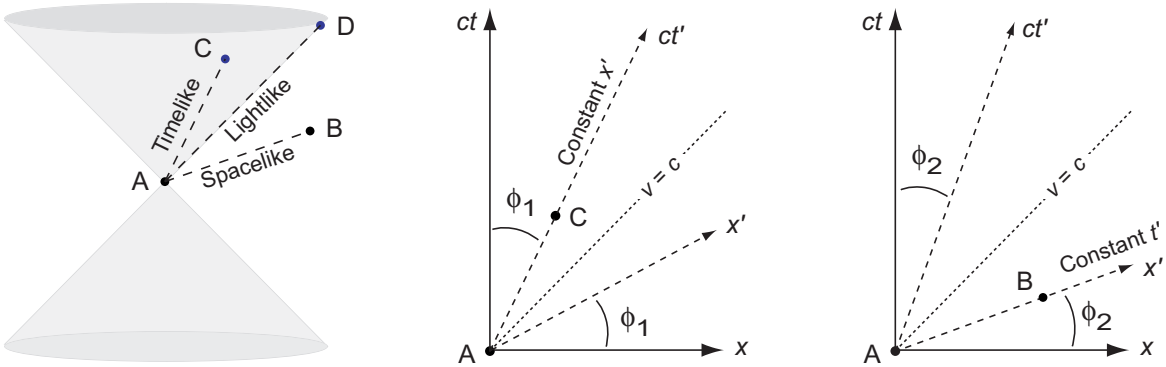
The geometry of the above figures suggests another important distinction between points at spacelike separations [the line AB in Fig. (a)] and timelike separations [the line AC in Fig. (a)]:

- If two events have *timelike separation*, a Lorentz transformation can bring them into spatial congruence.
- Figure (b) illustrates a coordinate system (ct', x').
  - It is related to the original system by an x-axis Lorentz boost by $v/c = \tan\varphi_1$,
  - in which *A and C have the same coordinate* x'.



On the other hand

- If two events have a *spacelike separation*, a Lorentz transformation exists that can synchronize the two points.
- Figure (c) illustrates an x-axis Lorentz boost by $v/c = \tan\varphi_2$ to a system in which *A and B have the same time* t'.



Notice that the maximum values of φ_1 and φ_2 are *limited by the* $v = c$ *line*.

- Thus, the Lorentz transformation to bring point A into spatial congruence with point C
  - exists only if point C lies to the *left of the* $v = c$ *line*
  - and thus is separated by a *timelike interval* from point A.
- Likewise, the Lorentz transformation to synchronize point A with point B
  - exists only if B lies to the *right of the* $v = c$ *line*,
  - meaning that it is separated by a *spacelike interval from A.*

4.7 Lorentz Covariance of Maxwell's Equations

We conclude this chapter by examining the Lorentz invariance of the Maxwell equations that describe classical electromagnetism. There are several motivations.

- It provides a nice example of how useful Lorentz invariance and Lorentz tensors can be.
- The properties of the Maxwell equations influenced Einstein strongly in his development of the special theory of relativity.
- There are many useful parallels between general relativity and the Maxwell theory, particularly for weak gravity where the Einstein field equations may be linearized.

Understanding covariance of the Maxwell equations will prove particularly important when gravitational waves are discussed in later chapters.

4.7.1 Maxwell Equations in Non-covariant Form

In free space, using Heaviside–Lorentz, $c = 1$ units, the Maxwell equations may be written as

$$\begin{aligned} \boldsymbol{\nabla}\cdot\boldsymbol{E} &= \rho, \\ \frac{\partial \boldsymbol{B}}{\partial t} + \boldsymbol{\nabla}\times\boldsymbol{E} &= 0, \\ \boldsymbol{\nabla}\cdot\boldsymbol{B} &= 0, \\ \boldsymbol{\nabla}\times\boldsymbol{B} - \frac{\partial \boldsymbol{E}}{\partial t} &= \boldsymbol{j}, \end{aligned}$$

where $\boldsymbol{E}$ is the electric field, $\boldsymbol{B}$ is the magnetic field, with the charge density ρ and current vector $\boldsymbol{j}$ required to satisfy the equation of continuity

$$\frac{\partial \rho}{\partial t} + \boldsymbol{\nabla}\cdot\boldsymbol{j} = 0.$$

Maxwell's equations are consistent with special relativity.

- However, in the above form this covariance is *not manifest,* since these equations are formulated in terms of 3-vectors and separate derivatives with respect to space and time, not Minkowski tensors.
- It proves useful to reformulate the Maxwell equations in a manner that is manifestly covariant with respect to Lorentz transformations.

The usual route to accomplishing this begins by replacing the electric and magnetic fields by new variables.

4.7.2 Scalar and Vector Potentials

The electric and magnetic fields may be eliminated in favor of a *vector potential* $\boldsymbol{A}$ and a *scalar potential* φ through the definitions

$$\boldsymbol{B} \equiv \boldsymbol{\nabla} \times \boldsymbol{A} \qquad \boldsymbol{E} \equiv -\boldsymbol{\nabla}\varphi - \frac{\partial \boldsymbol{A}}{\partial t}.$$

The vector identities

$$\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \boldsymbol{B}) = 0 \qquad \boldsymbol{\nabla} \times \boldsymbol{\nabla}\varphi = 0,$$

may then be used to show that the second and third Maxwell equations are satisfied identically, and the identity

$$\boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \boldsymbol{A}) = \boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \boldsymbol{A}) - \boldsymbol{\nabla}^2 \boldsymbol{A},$$

may be used to write the remaining two Maxwell equations as the coupled second-order equations

$$\begin{aligned} \boldsymbol{\nabla}^2 \varphi + \frac{\partial}{\partial t} \boldsymbol{\nabla} \cdot \boldsymbol{A} &= -\rho \\ \boldsymbol{\nabla}^2 \boldsymbol{A} - \frac{\partial^2 \boldsymbol{A}}{\partial t^2} - \boldsymbol{\nabla}\left(\boldsymbol{\nabla} \cdot \boldsymbol{A} + \frac{\partial \varphi}{\partial t}\right) &= -\boldsymbol{j}. \end{aligned}$$

These equations may then be *decoupled* by exploiting a fundamental symmetry of electromagnetism termed *gauge invariance*.

4.7.3 Gauge Transformations

Because of the identity $\boldsymbol{\nabla}\times\boldsymbol{\nabla}\varphi = 0$, the simultaneous transformations

$$\boldsymbol{A} \to \boldsymbol{A} + \boldsymbol{\nabla}\chi \qquad \varphi \to \varphi - \frac{\partial \chi}{\partial t}$$

for an arbitrary scalar function χ do not change the $\boldsymbol{E}$ and $\boldsymbol{B}$ fields; thus, *they leave the Maxwell equations invariant*.

- These are termed (classical) *gauge transformations.*
- This freedom of gauge transformation may be used to decouple the Maxwell equations.
- For example, if a set of potentials $(\boldsymbol{A}, \varphi)$ that satisfy

  $$\boldsymbol{\nabla}\cdot\boldsymbol{A} + \frac{\partial \varphi}{\partial t} = 0,$$

  is chosen, the equations decouple to yield

  $$\boldsymbol{\nabla}^2\varphi - \frac{\partial^2 \varphi}{\partial t^2} = -\rho \qquad \boldsymbol{\nabla}^2\boldsymbol{A} - \frac{\partial^2 \boldsymbol{A}}{\partial t^2} = -\boldsymbol{j},$$

  which may be solved independently for $\boldsymbol{A}$ and φ.
- Such a constraint is called a *gauge condition* and imposing the constraint is termed *fixing the gauge.*

> The particular choice of gauge in the above example is termed the *Lorenz gauge*.

Another common gauge is the *Coulomb gauge*, with a gauge-fixing condition

$$\boldsymbol{\nabla}\cdot\boldsymbol{A} = 0,$$

which leads to the decoupled Maxwell equations

$$\boldsymbol{\nabla}^2\varphi = -\rho \qquad \boldsymbol{\nabla}^2\boldsymbol{A} - \frac{\partial^2\boldsymbol{A}}{\partial^2 t} = \boldsymbol{\nabla}\frac{\partial\varphi}{\partial t} - \boldsymbol{j}.$$

Let's utilize the shorthand for derivatives introduced earlier:

$$\partial^\mu \equiv \frac{\partial}{\partial x_\mu} = (\partial^0, \partial^1, \partial^2, \partial^3) = \left(-\frac{\partial}{\partial x^0}, \boldsymbol{\nabla}\right),$$
$$\partial_\mu \equiv \frac{\partial}{\partial x^\mu} = (\partial_0, \partial_1, \partial_2, \partial_3) = \left(\frac{\partial}{\partial x^0}, \boldsymbol{\nabla}\right),$$

where, for example, $\partial^1 = \partial/\partial x_1$ and the 3-divergence is

$$\boldsymbol{\nabla} \equiv (\partial^1, \partial^2, \partial^3).$$

A *covariant formalism* then results from introducing

- the *4-vector potential* A^μ,
- the *4-current* j^μ, and
- the *d'Alembertian operator* $\Box$

through the definitions

$$A^\mu \equiv (\varphi, \boldsymbol{A}) = (A^0, \boldsymbol{A}) \qquad j^\mu \equiv (\rho, \boldsymbol{j}) \qquad \Box \equiv \partial_\mu \partial^\mu.$$

Then a gauge transformation takes the form

$$A^\mu \to A^\mu - \partial^\mu \chi \equiv A'^\mu$$

and the preceding examples of gauge-fixing constraints become

$$\partial_\mu A^\mu = 0 \ \text{(Lorenz gauge)} \qquad \boldsymbol{\nabla}\cdot\boldsymbol{A} = 0 \ \text{(Coulomb gauge)}.$$

> The *Lorenz condition is covariant* (formulated in terms of 4-vectors); the *Coulomb gauge condition is not covariant* (formulated in terms of 3-vectors).

The operator $\Box$ is *Lorentz invariant* since

$$\Box' = \partial'_\mu \partial'^\mu = \Lambda^\nu{}_\mu \Lambda_\lambda{}^\mu \partial_\nu \partial^\lambda = \partial_\mu \partial^\mu = \Box.$$

Thus, the Lorenz-gauge wave equation may be expressed in the *manifestly covariant form*

$$\Box A^\mu = j^\mu$$

and the continuity equation becomes

$$\partial_\mu j^\mu = 0.$$

The Maxwell wave equations in Lorenz gauge are *manifestly covariant*.

- This, coupled with the *gauge invariance of electromagnetism*, ensures that the Maxwell equations are covariant in all gauges.
- However—as was seen in the example of the Coulomb gauge—the covariance may not be manifest for a particular choice of gauge.

Let's now see how to formulate the Maxwell equations in a *manifestly covariant form*.

4.7.4 Maxwell Equations in Manifestly Covariant Form

The Maxwell equations may be cast in a manifestly covariant form by constructing the components of the electric and magnetic fields in terms of the potentials (Problems).

- Proceeding in this manner, we find that the six independent components of the 3-vectors $\boldsymbol{E}$ and $\boldsymbol{B}$ are elements of an antisymmetric rank-2 *electromagnetic field tensor*

$$F^{\mu\nu} = -F^{\nu\mu} = \partial^\mu A^\nu - \partial^\nu A^\mu,$$

  which may be expressed in matrix form as

$$F^{\mu\nu} = \begin{pmatrix} 0 & -E^1 & -E^2 & -E^3 \\ E^1 & 0 & -B^3 & B^2 \\ E^2 & B^3 & 0 & -B^1 \\ E^3 & -B^2 & B^1 & 0 \end{pmatrix}.$$

- That is, the electric field $\boldsymbol{E}$ and the magnetic field $\boldsymbol{B}$
  - are *vectors* in 3D euclidean space but
  - their six components together form an antisymmetric *rank-2 tensor* in Minkowski space.

Now let's employ the *Levi–Civita symbol* $\varepsilon_{\alpha\beta\gamma\delta}$, where

- $\varepsilon_{\alpha\beta\gamma\delta}$ has the value $+1$ for $\alpha\beta\gamma\delta = 0123$ and cyclic permutations,
- -1 for odd permutations, and
- zero if any two indices are equal,
- and use it to define the *dual field tensor* $\mathscr{F}^{\mu\nu}$ by

$$\mathscr{F}^{\mu\nu} \equiv \tfrac{1}{2}\varepsilon^{\mu\nu\gamma\delta}F_{\gamma\delta} = \begin{pmatrix} 0 & -B^1 & -B^2 & -B^3 \\ B^1 & 0 & E^3 & -E^2 \\ B^2 & -E^3 & 0 & E^1 \\ B^3 & E^2 & -E^1 & 0 \end{pmatrix}.$$

- Then *two of the four Maxwell equations* may be written

$$\partial_\mu F^{\mu\nu} = j^\nu,$$

- and the *other two Maxwell equations* may be written as

$$\partial_\mu \mathscr{F}^{\mu\nu} = 0.$$

The Maxwell equations in this form are *manifestly covariant* (under Lorentz transformations) because they are formulated *exclusively in terms of Lorentz tensors*.

Chapter 5

Lorentz-Invariant Dynamics

In the preceding chapter we introduced the Minkowski metric and covariance with respect to Lorentz transformations between inertial systems. This was shown to lead to the basic properties of special relativity:

- relativity of simultaneity,
- time dilation, and
- space contraction.

In this chapter we continue that discussion for flat Minkowski space and consider general properties of trajectories for particles and for light in Minkowski spacetime.

5.1 Geometrized Units

It is convenient to introduce a new set of units in which c and/or G can be set to unit value so that they do not appear explicitly in equations.

- These are called *geometrized units* or $c = G = 1$ units.
- Such units are also sometimes called *natural units,* because they are suggested by the physics of the problem.

Geometrized units, and how to convert between standard units and geometrized units, are explained in examples below and in an Appendix of the book.

Assuming $c = G = 1$ and setting

$$1 = c = 2.9979 \times 10^{10}\,\text{cm s}^{-1}$$
$$1 = G = 6.6720 \times 10^{-8}\,\text{cm}^3\ \text{g}^{-1}\ \text{s}^{-2},$$

we may solve for standard units in terms of these new units.

- From the first equation $1 = 2.9979 \times 10^{10}\,\text{cm s}^{-1}$, implying that

  $$1\,\text{s} = 2.9979 \times 10^{10}\,\text{cm} \qquad 1\,\text{cm} = (2.9979 \times 10^{10})^{-1}\,\text{s},$$

  and from the second $1 = 6.6720 \times 10^{-8}\,\text{cm}^3\ \text{g}^{-1}\ \text{s}^{-2}$, so

  $$\begin{aligned} 1\,\text{g} &= 6.6720 \times 10^{-8}\,\text{cm}^3\ \text{s}^{-2} \\ &= 6.6720 \times 10^{-8}\,\text{cm}^3 \left(\frac{1}{2.9979 \times 10^{10}\,\text{cm}}\right)^2 \\ &= 7.4237 \times 10^{-29}\,\text{cm}. \end{aligned}$$

  Thus *distance, time, and mass all have the dimension of length* in geometrized units.

- Likewise, we may derive from the above relations

  $$1\,\text{erg} = 1\,\text{g cm}^2\ \text{s}^{-2} = 8.2601 \times 10^{-50}\,\text{cm},$$
  $$1\,\text{g cm}^{-3} = 17.4237 \times 10^{-29}\,\text{cm}^{-2},$$
  $$1M_\odot = 1.4766\,\text{km},$$

  and so on.

- Velocity is dimensionless in these units since

$$1\,\mathrm{cm\ s^{-1}} = 2.9979 \times 10^{-10}$$

(that is, v is measured in units of v/c).

From this point onward we shall commonly work in $c = G = 1$ (or $c = 1$ units if gravity isn't involved), unless the explicit restoration of c or G factors is desirable

- for clarity,
- to make a particular point, or
- to compute numbers in "engineering units".

In geometrized units

- all occurences of G and c are *omitted from equations*.
- Calculating quantities in standard units then requires *reinserting appropriate combinations of* c *and* G to give the right physical dimensions for each term.

Example: it will be shown later that the Schwarzschild radius defining the event horizon for a spherical black hole in geometrized units is $r_S = 2M$, where M is the mass.

- Both sides of this equation have dimensions of length in geometrized units.
- What is the Sun's Schwarzschild radius in standard units?

The result may be obtained by inspection since

- In geometrized units $1\,M_\odot = 1.4766\,\text{km}$.
- Thus, for the Sun

$$r_S = 2M_\odot = 2 \times 1.4766\,\text{km} = 2.95\,\text{km}.$$

Alternatively, to convert this equation to CGS units note that

- $r_S = 2M$ implies that the right side must be multiplied by a combination of G and c having the units of cm g^{-1} to make it dimensionally correct in the CGS system.
- Clearly this requires the combination G/c^2, so in CGS units the Schwarzschild radius is $r_S = 2GM/c^2$.

Quantity	Symbol	Geometrized unit	Standard unit	Conversion
Mass	M	$\mathscr{L}$	$\mathscr{M}$	GM/c^2
Length	L	$\mathscr{L}$	$\mathscr{L}$	L
Time	t	$\mathscr{L}$	$\mathscr{T}$	ct
Spacetime distance	s	$\mathscr{L}$	$\mathscr{L}$	s
Proper time	τ	$\mathscr{L}$	$\mathscr{T}$	$c\tau$
Energy	E	$\mathscr{L}$	$\mathscr{M}(\mathscr{L}/\mathscr{T})^2$	GE/c^4
Momentum	p	$\mathscr{L}$	$\mathscr{M}(\mathscr{L}/\mathscr{T})$	Gp/c^3
Angular momentum	J	$\mathscr{L}^2$	$\mathscr{M}(\mathscr{L}^2/\mathscr{T})$	GJ/c^3
Luminosity (power)	L	dimensionless	$\mathscr{M}(\mathscr{L}^2/\mathscr{T}^3)$	GL/c^5
Energy density	ε	$\mathscr{L}^{-2}$	$\mathscr{M}/(\mathscr{L}\mathscr{T}^2)$	$G\varepsilon/c^4$
Momentum density	π_i	$\mathscr{L}^{-2}$	$\mathscr{M}/(\mathscr{L}^2\mathscr{T})$	$G\pi_i/c^3$
Pressure	P	$\mathscr{L}^{-2}$	$\mathscr{M}/(\mathscr{L}\mathscr{T}^2)$	GP/c^4
Energy / unit mass	ε	dimensionless	$(\mathscr{L}/\mathscr{T})^2$	ε/c^2
Ang. mom. / unit mass	ℓ	$\mathscr{L}$	$\mathscr{L}^2/\mathscr{T}$	ℓ/c
Planck constant	$\hbar$	$\mathscr{L}^2$	$\mathscr{M}(\mathscr{L}^2/\mathscr{T})$	$G\hbar/c^3$

The standard unit of length is $\mathscr{L}$, the standard unit of mass is $\mathscr{M}$, and the standard unit of time is $\mathscr{T}$. To convert equations to standard units from geometrized units, replace quantities in column 2 with quantities in the last column. To convert from standard to geometrized units, multiply by the factor of G and c appearing in the last column.

The Table above (from Appendix of book) also may be used to read off the same result:

- From the table, conversion of $R_{\rm S} = 2M$ from natural to standard units requires the replacements

$$r_{\rm S} \to r_{\rm S} \qquad M \to GM/c^2.$$

- This gives $r_{\rm S} = 2GM/c^2$.

Finally as a check, if the problem is worked directly in CGS units:

$$\begin{aligned} r_{\mathrm{S}}^{\odot} &= \frac{2(6.674\times10^{-8}\,\mathrm{cm^3\ g^{-1}\ s^{-2}})(1.989\times10^{33}\,\mathrm{g})}{(3\times10^{10}\,\mathrm{cm\ s^{-1}})^2} \\ &= 2.95\times10^{5}\,\mathrm{cm} \\ &= \ 2.95\,\mathrm{km}, \end{aligned}$$

which is the same result as obtained above.

5.2 Velocity and Momentum for Massive Particles

Particles with finite mass follow *timelike worldlines*.

- The worldline for a particle is conveniently parameterized in terms of a variable that changes continuously along the worldline.
- For timelike trajectories the natural choice for this parameter is the *proper time* τ.

The equation of the worldline may then be expressed as

$$x^\mu = x^\mu(\tau)$$

and we may define a velocity 4-vector (the *4-velocity*) by

$$u^\mu = (u^0, u^1, u^2, u^3) = \left(\frac{dx^0}{d\tau}, \frac{dx^1}{d\tau}, \frac{dx^2}{d\tau}, \frac{dx^3}{d\tau}\right).$$

The *proper time interval* $d\tau$ and *spacetime interval* ds are related by

$$d\tau^2 = -ds^2,$$

and the *coordinate time interval* dt and the *proper time interval* $d\tau$ are related through special-relativistic time dilation:

$$d\tau = dt\left(1 - \boldsymbol{v}^2\right)^{1/2} = \frac{1}{\gamma}dt \qquad \gamma \equiv \left(1 - \boldsymbol{v}^2\right)^{-1/2}$$

where $\boldsymbol{v}$ is the *3-velocity*, $v^i = dx^i/dt$. (Note: $c = 1$ units!)

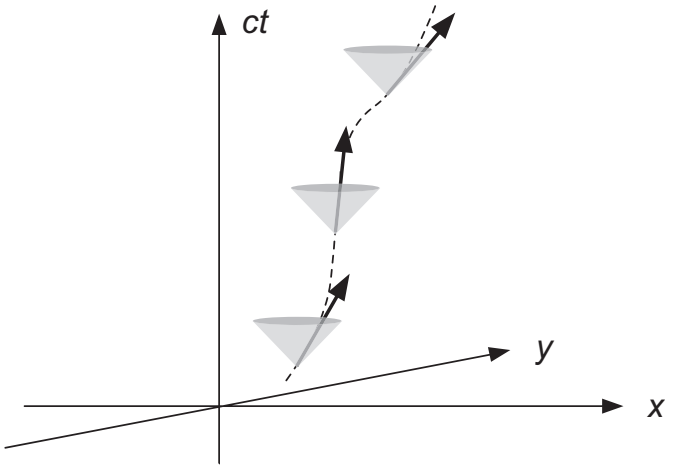


Figure 5.1: The 4-velocity along a timelike worldline.

The 4-velocity for a *massive (timelike) particle* is

- *tangent to the worldline* of the particle at any point and
- *lies within the forward light cone* (Fig. 5.1).

Since $dt = \gamma d\tau$,

$$u^0 = \frac{dx^0}{d\tau} = \frac{dt}{d\tau} = \frac{\gamma d\tau}{d\tau} = \gamma = \left(1 - \boldsymbol{v}^2\right)^{-1/2}$$

$$u^i = \frac{dx^i}{d\tau} = \underbrace{\frac{dx^i}{dt}}_{v_i}\underbrace{\frac{dt}{d\tau}}_{\gamma} = v_i\gamma = v_i\left(1 - \boldsymbol{v}^2\right)^{-1/2}$$

so that we may write for the components of the 4-velocity

$$u^\mu = (\gamma, \gamma\boldsymbol{v}) \qquad \gamma = \left(1 - \boldsymbol{v}^2\right)^{-1/2}.$$

(*Remember:* we are using units where $c = 1$.)

Since we have

$$ds^2 = -d\tau^2 = \eta_{\mu\nu} dx^\mu dx^\nu,$$

which gives, upon dividing by $d\tau^2$,

$$-1 = \eta_{\mu\nu} \frac{dx^\mu}{d\tau} \frac{dx^\nu}{d\tau} = \eta_{\mu\nu} u^\mu u^\nu = u \cdot u,$$

the scalar product of u with itself gives the normalization

$$u \cdot u = -1,$$

$$u^\mu = (\gamma, \gamma \boldsymbol{v}) \qquad \eta_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

For *massive particles* we may *always* invoke the condition $u \cdot u = -1$.

5.2.1 4-Momenta

We may define the *4-momentum* by

$$p^\mu \equiv (E, \boldsymbol{p}) = mu^\mu,$$

where m is the *rest mass*. Since $u\cdot u = -1$, the normalization of the 4-momentum is

$$p^2 \equiv p\cdot p = p_\mu p^\mu = m^2 u\cdot u = -m^2.$$

Because $u^\mu = (\gamma, \gamma\boldsymbol{v})$, the components of the 4-momentum are

$$p^\mu = (E, \boldsymbol{p}) = (\gamma m, \gamma m\boldsymbol{v}) \quad \longrightarrow \quad p_\mu = \eta_{\mu\nu}p^\nu = (-E, \boldsymbol{p}),$$

with $\gamma = \left(1-\boldsymbol{v}^2\right)^{-1/2}$. Thus, $p^2 = p_\mu p^\mu = -m^2$ implies that

$$p_\mu p^\mu = (-E, \boldsymbol{p}) \begin{pmatrix} E \\ \boldsymbol{p} \end{pmatrix} = -m^2 \quad \longrightarrow \quad E = \sqrt{\boldsymbol{p}^2 + m^2},$$

which is just the most famous equation in physics,

$$E = \sqrt{\boldsymbol{p}^2c^2 + m^2c^4} \quad \longrightarrow \quad E = mc^2 \quad (\boldsymbol{p} \to 0),$$

written in $c = 1$ units.

5.3 Geodesics

A metric allows us to define *geodesics*:

- A geodesic for a manifold is a path that represents the *shortest distance between any two points.*
- A geodesic may also be viewed as the *"straightest possible path"* between two points.
- More technically, a geodesic is a curve that *"parallel-transports its own tangent vector"*.

EUCLIDEAN SPACE:

"The shortest distance between two points is a straight line." Thus, the geodesics in Euclidean space are given by

$$\ddot{\boldsymbol{r}} = 0 \qquad \text{(Newton's 1st law)}$$

MINKOWSKI SPACE:

$$\frac{d^2t}{d\tau^2} = 0 \qquad \frac{d^2\boldsymbol{r}}{d\tau^2} = 0,$$

where τ is the *proper time* (time measured by a clock carried along a worldline).

In both of the above examples, the *geodesics are straight lines* (this generally will not be true in curved spacetime).

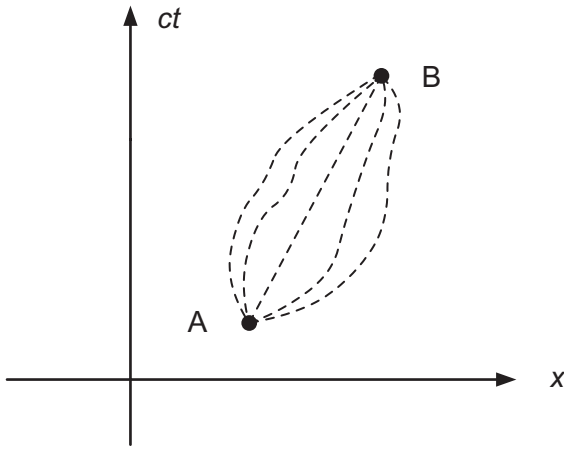


Figure 5.2: Extremizing the proper time to determine the geodesic for a particle.

5.4 Principle of Extremal Proper Time

PRINCIPLE OF EXTREMAL PROPER TIME: the worldline for free particles between timelike separated points *extremizes the proper time* (Fig. 5.2).

From (using $c = 1$ units)

$$d\tau^2 = -ds^2 = -(-dt^2 + dx^2 + dy^2 + dz^2,)$$

the proper time between the points A and B is

$$\tau_{AB} = \int_A^B (dt^2 - dx^2 - dy^2 - dz^2)^{1/2}.$$

We may parameterize the path by a variable σ that varies continuously from 0 to 1 as the particle moves from A to B and

$$\tau_{AB} = \int_0^1 \left[\left(\frac{dt}{d\sigma}\right)^2 - \left(\frac{dx}{d\sigma}\right)^2 - \left(\frac{dy}{d\sigma}\right)^2 - \left(\frac{dz}{d\sigma}\right)^2 \right]^{1/2} d\sigma.$$

The condition for an extremum is that

$$\delta \int d\tau = 0,$$

where the variation is generally of the explicit form

$$\delta f = \frac{\partial f}{\partial x^\mu} \delta x^\mu.$$

Defining a *Lagrangian* for a metric $g_{\mu\nu}$

$$L \equiv \left(-g_{\mu\nu} \frac{dx^\mu}{d\sigma} \frac{dx^\nu}{d\sigma} \right)^{1/2} \qquad \longrightarrow \qquad \tau_{AB} = \int_0^1 L d\sigma,$$

the variation $\delta \int d\tau = 0$ then implies the *Euler–Lagrange equation of motion*

$$-\frac{d}{d\sigma} \left(\frac{\partial L}{\partial (dx^\mu / d\sigma)} \right) + \frac{\partial L}{\partial x^\mu} = 0.$$

(Proved in a Box in the book.)

EXAMPLE: Consider $x^\mu = x^1$. The Euler–Lagrange equation is

$$\underbrace{-\frac{d}{d\sigma}\left(\frac{\partial L}{\partial(dx^\mu/d\sigma)}\right)}_{\text{derivative dependence}} \quad \underbrace{+\frac{\partial L}{\partial x^\mu}}_{\text{coordinate dependence}} = 0$$

For constant $\eta_{\mu\nu}$ the Lagrangian L does not depend on x^1 and the Euler–Lagrange equation reduces to

$$-\frac{d}{d\sigma}\left(\frac{\partial L}{\partial(dx^\mu/d\sigma)}\right) + \underbrace{\frac{\partial L}{\partial x^\mu}}_{=0} = 0 \quad \rightarrow \quad \frac{d}{d\sigma}\left(\frac{1}{L}\frac{dx^1}{d\sigma}\right) = 0.$$

Inserting $1/L = d\sigma/d\tau$ and multiplying by $d\sigma/d\tau$, gives

$$\frac{d^2x^1}{d\tau^2} = 0$$

Applying similar steps to the other terms then gives the general result (Problem)

$$\frac{d^2x^\mu}{d\tau^2} = 0 \qquad \rightarrow \text{ No curvature for geodesic}$$

The principle of extremal proper time implies that *geodesics in Minkowski space are straight lines.*

Principle of Extremal Proper Time (Taylor and Wheeler):

> "Spacetime shouts 'Go straight!' The free stone obeys. ... The stone's wristwatch verifies that its path is straight."

5.5 Light Rays

For particles moving at lightspeed the *rest mass is identically zero*.

- *Light-like particles* such as photons move on the light cone with the proper time between two points given by

$$d\tau^2 = -ds^2 = 0,$$

- Thus photons and other light-like (massless) particles travel any Minkowski distance in *zero proper time.*

> Therefore the *proper time* τ *is not a useful parameterization* for the world line of photons and other massless particles.

However, notice that we may write the curve $x = t$ (corresponding to $v = c$ expressed in $c = 1$ units) parametrically as

$$x^\mu = u^\mu \lambda$$

where $u^\mu = (1,1,0,0)$ is a tangent 4-vector,

$$u^\mu = \frac{dx^\mu}{d\lambda}$$

and λ is a parameter.

With this choice of parameterization the equation of motion for the light ray may be put into the same form as that for a massive particle

$$\frac{du}{d\lambda} = 0$$

which is analogous to Newton's first law.

- Parameters λ for which this is true are termed *affine parameters.*
- Affine parameters generally are not unique.
- For example, if λ is an affine parameter then λ multiplied by any constant is also an affine parameter.

Affine parameters are *convenient for light rays* because they lead to *equations of motion that mimic those for timelike particles*.

For massive particles $u\cdot u=-1$, but since for the photon case $u^\mu=(1,1,0,0)$ for motion on the x-axis, we have

$$u_\mu=\eta_{\mu\nu}u^\nu=(-1,1,0,0).$$

Thus for photons

$$u\cdot u=u_\mu u^\mu=(-1,1,0,0)\times\begin{pmatrix}1\\1\\0\\0\end{pmatrix}=-1+1=0.$$

The primary differences between

- equations governing the motion of massive particles and
- those governing the motion of massless particles (e.g., photons)

in spacetime will be associated with the difference in 4-velocity normalizations

$$u\cdot u=-1 \quad \text{(for massive particles)},$$

$$u\cdot u=0 \quad \text{(for massless particles)}.$$

Otherwise their equations of motion will be similar.

For photons the energy E and momentum $\boldsymbol{p}$ are given by

$$E = \hbar\omega \qquad \boldsymbol{p} = \hbar\boldsymbol{k},$$

where $\hbar$ is Planck's constant, ω is the frequency, and $\boldsymbol{k}$ is the wavevector. Thus,

$$p^\mu = (E, \boldsymbol{p}) = (\hbar\omega, \hbar\boldsymbol{k}) = \hbar k^\mu = \hbar(\omega, \boldsymbol{k}).$$

- Since photons are massless, the 4-momentum and 4-wavevector are normalized such that

  $$p \cdot p = k \cdot k = 0,$$

  which is $E = pc$ in $c = 1$ units.

- The equations of motion for photons may also be expressed in terms of the 4-momentum or 4-wavevector,

  $$\frac{dp}{d\lambda} = 0 \qquad \frac{dk}{d\lambda} = 0,$$

  where λ is an affine parameter.

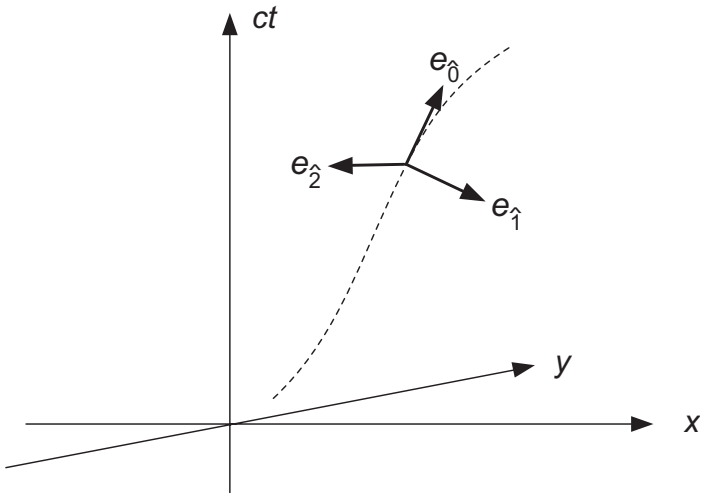


Figure 5.3: Unit vectors of a local coordinate system at a point on an observer's worldline for two space and one time dimension.

5.6 Observers

To test theory against data we must introduce *observers*.

- An observer may be thought of as *occupying a local laboratory* moving on a (timelike) worldline in the spacetime.
- She carries 4 orthogonal unit vectors $e_{\hat{0}}$, $e_{\hat{1}}$, $e_{\hat{2}}$, and $e_{\hat{3}}$ specifying a *local, orthonormal coordinate system* (Fig. 5.3).
- Note: *Hats on indices* indicate explicitly that this is a *local orthonormal coordinate system,* not our usual position-dependent coordinate basis.
- This coordinate system defines (locally) a time direction and three space directions to which the observer will reference all measurements.

- The timelike component $e_{\hat{0}}$ will be *tangent to the observer's worldline* (the observer's clock is moving in that direction if it is at rest in the laboratory).
  - Since the 4-velocity u of the observer is a unit tangent vector ($u \cdot u = -1$),
    $$e_{\hat{0}} = u.$$
  - The observer may choose any mutually orthogonal set of three unit spatial vectors to complete the set, as long as they are orthogonal to $e_{\hat{0}}$.
- Observers refer observations to the axes of their lab and its clocks.
  - Thus, they measure *components of 4-vectors* along their chosen basis vectors.
  - These components may be computed by *taking scalar products* with the orthonormal basis 4-vectors.

Example: For the 4-momentum

$$p = p^{\hat{\mu}} e_{\hat{\mu}},$$

we have in particular that the energy measured by an observer with 4-velocity u_{obs} is given by

$$E = p^0 = -p \cdot e_{\hat{0}}.$$

5.7 Isometries and Killing Vectors

In differential geometry, *Killing vectors* are standard tools for analyzing symmetries such as those that arise as conservation laws in the usual Lagrangian or Hamiltonian formulations of mechanics.

- In all spacetimes, whether flat or not, one constant of motion may be deduced from the normalization of the 4-velocity $u^\mu = dx^\mu/d\tau$

$$g_{\mu\nu}u^\mu u^\nu = -1,$$

  corresponding to the preservation of $u\cdot u$ for a timelike particle.

- If there are additional constants of motion, they must arise from *specific symmetries in the problem.*

- In ordinary mechanics, continuous symmetries imply conservation laws.

  > *Example:* *conservation of angular momentum* follows from a potential that is *spherically symmetric*.

- If a spacetime metric has a symmetry (termed an *isometry)*, that too will generally imply that *some quantity is conserved.*

Suppose the metric is *independent of one of the spacetime coordinates,* say x^0, such that

$$x^0 \to x^0 + \text{constant}$$

leaves the metric unchanged.

- For such an isometry we define a unit vector pointing along the direction in which the metric is constant,

$$K^\mu = (1,0,0,0).$$

- The vector K^μ is termed the *Killing vector* associated with the symmetry.

In flat 3D space

$$ds^2 = dx^2 + dy^2 + dz^2$$

and conservation of the components of linear momentum is associated with three Killing vectors

$$(1,0,0) \quad (0,1,0) \quad (0,0,1)$$

indicating invariance under translations in the x, y, and z directions, respectively.

A symmetry implied by a Killing vector means that

- Some quantity is *conserved along a geodesic.*
- This quantity may be found using the principle of extremal proper time *(Euler–Lagrange equation)*.

Example: The Euler–Lagrange equation is

$$-\frac{d}{d\sigma}\left(\frac{\partial L}{\partial(dx^\mu/d\sigma)}\right)+\frac{\partial L}{\partial x^\mu}=0 \quad L\equiv\left(-g_{\mu\nu}\frac{dx^\mu}{d\sigma}\frac{dx^\nu}{d\sigma}\right)^{1/2}$$

Let $g_{\mu\nu}$ be independent of x^1, corresponding to a Killing vector

$$K^\alpha=(0,1,0,0)$$

Then $\partial L/\partial x^1=0$ and (Problem)

$$\begin{aligned}\frac{\partial L}{\partial(dx^1/d\sigma)}&=-\frac{1}{2L}\left(g_{1\nu}(x)\frac{dx^\nu}{d\sigma}+g_{\mu 1}(x)\frac{dx^\mu}{d\sigma}\right)\\&=-\frac{1}{2L}\left(g_{1\mu}(x)\frac{dx^\mu}{d\sigma}+g_{\mu 1}(x)\frac{dx^\mu}{d\sigma}\right)\\&=-\frac{g_{1\mu}}{L}\frac{dx^\mu}{d\sigma}=-\frac{g_{1\mu}}{L}\frac{dx^\mu}{d\tau}\frac{d\tau}{d\sigma}=-g_{1\mu}\frac{dx^\mu}{d\tau}\\&=-g_{\alpha\mu}K^\alpha u^\mu=-K\cdot u,\end{aligned}$$

where we have used that $g_{\mu\nu}$ is symmetric and

$$L=\left(-g_{\mu\nu}\frac{dx^\mu}{d\sigma}\frac{dx^\nu}{d\sigma}\right)^{1/2}\qquad\frac{d\tau}{d\sigma}=L\qquad g_{1\mu}=g_{\alpha\mu}K^\alpha.$$

Then the Euler–Lagrange equation reduces to

$$\frac{d}{d\sigma}(K\cdot u)=0\quad\longrightarrow\quad K\cdot u\ \text{conserved on geodesic}$$

> The quantity $K\cdot u$ is *conserved along a geodesic* if K is a Killing vector and u is the 4-velocity.

For most of our applications we will be able to guess the Killing vectors.

- However, more formally (proved in the book chapter), Killing vectors satisfy the differential equation

$$\nabla_\nu K_\mu + \nabla_\mu K_\nu = \partial_\nu K_\mu + \partial_\mu K_\nu = 0.$$

- This is known as *Killing's equation.*
- The vector fields that solve it are the *Killing (vector) fields* associated with symmetries of the metric.

Chapter 6

The Principle of Equivalence

The general theory of relativity rests upon two principles that are in fact related:

- The principle of *equivalence*
- The principle of *general covariance*

Let's consider the *equivalence principle.*

6.1 Inertial and Gravitational Mass

1. The *inertial mass* is defined through *Newton's second law*: $m = F/a$.
2. The *gravitational mass* is defined through *Newton's law of gravitation*: $m = r^2F/GM$.
3. The relationship between inertial and gravitational masses was suggested by Galileo: *different objects fall at the same rate in a gravitational field,*
4. which is equivalent to *equality of inertial and gravitational mass*.
5. This was first established to high precision in the *Eötvös experiments* of 1893.

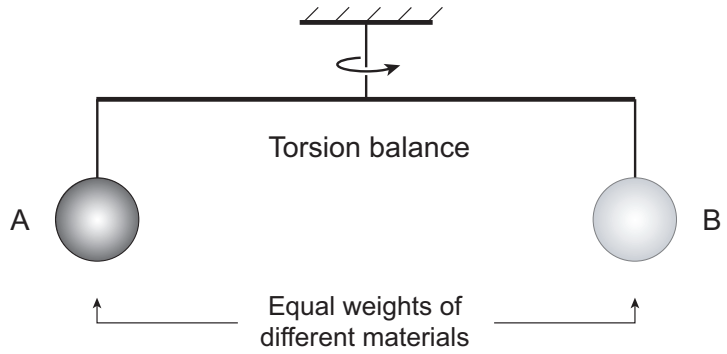


Figure 6.1: Measuring the difference between gravitational and inertial mass.

If the inertial and gravitational masses differ

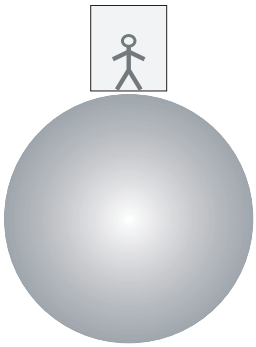
- In Fig. 6.1 a couple will be produced by the action on the inertial mass of the centripetal effects associated with Earth's rotation and
- the *balance will twist* if $m_{\text{inertial}} \neq m_{\text{gravitational}}$.
- Result of the original Eötvös experiment: *inertial and gravitational masses are equivalent,* with a sensitivity of one part in 10^9 (10^{13} in more modern experiments).

Weak Principle of Equivalence:

$$m_{\text{inertial}} = m_{\text{gravitational}}.$$

Equivalent statement: All objects experience the same acceleration in a gravitational field, irrespective of their masses or any other intrinsic property.

(a) Stationary elevator in a gravitational field at the surface of a planet.



(b) Elevator accelerated in interstellar space far from any gravitating masses

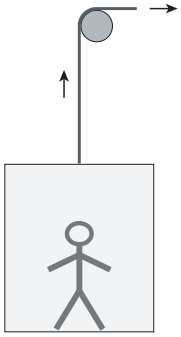


Figure 6.2: The Einstein elevator.

6.2 Strong Equivalence Principle

Einstein extended this idea to the modern equivalence principle (sometimes called the *strong principle of equivalence*), based on a thought experiment. For the elevator illustrated in Fig. 6.2,

the occupant is unable to distinguish

- an acceleration of the elevator at some point in space where no gravitation fields act from
- the effect of a stationary elevator sitting in a planetary gravitational field.

Henceforth, by *equivalence principle* we shall mean the *strong equivalence principle*.

The (strong) *equivalence principle* can be stated in several equally-valid ways:

- For an observer in free fall in a gravitational field, the results of all local experiments are independent of the magnitude of the gravitational field.
- All *local*, freely falling, non-rotating laboratories are fully equivalent for the performance of physical experiments. Such a laboratory is called a *local inertial frame* or *Lorentz frame*.
- In any sufficiently *local* region of spacetime, the effect of gravity can be transformed away.
- In any sufficiently *local* region of spacetime, we may construct a *local inertial system* in which *the special theory of relativity is valid,* even in a very strong gravitational field.
- all forms of mass and energy contribute equivalent quantities of gravitational and inertial mass.

In these statements of equivalence a practical definition of "freely falling" is *weightlessness:*

> Any experiment would reveal any object to be weightless in a freely-falling frame.

We shall give a precise definition of *"local"* shortly.

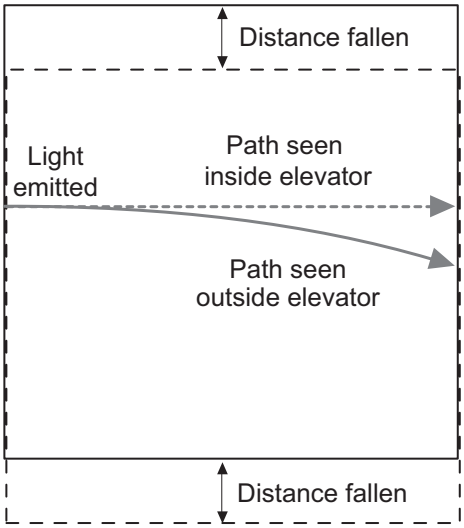


Figure 6.3: Equivalence and deflection of light in a gravitational field.

6.3 Deflection of Light in a Gravitational Field

By applying the principle of equivalence, Einstein obtained important results of the general theory of relativity even before he could solve the corresponding field equations. Consider Fig. 6.3 where *a sealed elevator falls in a gravitational field.*

INTERIOR OBSERVER: Equivalence $\rightarrow$ we may transform away the effect of gravity. Observer in the interior is unaware of any gravitational field and sees light travel in a straight line.

EXTERIOR OBSERVER: Aware of the gravitational field (sees the elevator falling!). The spot at which the light strikes the right wall has fallen by the same amount as the elevator.

RECONCILE (observers must agree on laws of physics): Light follows a *curved path* as it propagates in a gravitational field.

6.3.1 Strength of the Gravitational Field

Bending of the light in the gravitational field may be characterized by a radius of curvature (Problem)

$$r_c = \frac{c^2}{g},$$

Strength of gravitational field at the surface of a gravitating object such as a star quantified through:

$$\frac{R}{r_c} = \frac{GM}{Rc^2} = \frac{\text{Actual radius}}{\text{Light curvature radius}},$$

where $g = GM/R^2$ has been used. If

$$GM/Rc^2 << 1,$$

the field is *weak* (Newtonian gravity). May be expressed as

$$\frac{R}{r_c} = \frac{GM}{Rc^2} \cdot \frac{m}{m} = \frac{GMm/R}{mc^2} = \frac{E_g}{E_0} = \frac{\text{Gravitational energy}}{\text{Rest-mass energy}},$$

- *Weak Field:* gravitational energy of a test particle of mass m is much less than its rest mass energy.
- *EXAMPLE:* White dwarf Sirius B has $\rho \sim 10^6\,\mathrm{g\,cm^{-3}}$, which gives $R/r_c \simeq 10^{-4}$. Even for a white dwarf gravity is *weak on the natural scale set by light curvature*.
- *EXAMPLE:* At the surface of a neutron star or at the event horizon of a black hole, gravitational curvature radius $\sim$ actual radius $\rightarrow$ general relativity.

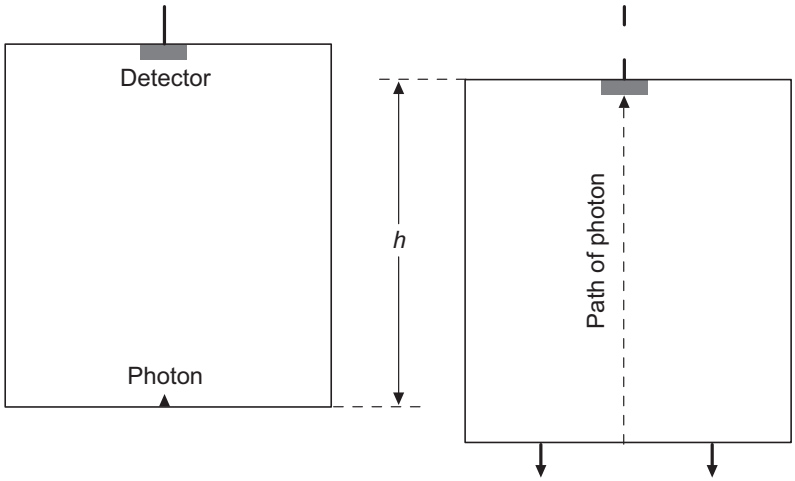


Figure 6.4: Equivalence and the gravitational redshift.

6.4 The Gravitational Redshift

INTERNAL OBSERVER: Free fall, *unaware of gravity*, $\nu = \nu_0$.

EXTERNAL OBSERVER: *Aware of gravity* (sees elevator fall!).

- When light reaches ceiling a time $t = h/c$ has elapsed and
- the elevator has accelerated to a velocity $v = gt = gh/c$.

Doppler shift : $$\frac{\Delta\nu}{\nu} = \frac{v}{c} = \frac{gh/c}{c} = \frac{GMh}{R^2c^2} \qquad \text{since } g = \frac{GM}{R^2}$$

RECONCILE: To avoid contradiction, there must be a *redshift produced by the gravitational field* that exactly cancels the blueshift (produced by velocity, not gravity).

6.4.1 Total Redshift in a Gravitational Field

Integrated redshift assuming *relatively weak gravity*

$$\int_{\nu_0}^{\nu_s} \frac{d\nu}{\nu} = -\int_R^s \frac{GM}{r^2c^2}\,dr,$$

Exercise: Integrating, exponentiating, and expanding the right-side exponential (weak-field assumption) gives at radius s,

$$\frac{\nu_s}{\nu_0} \simeq 1 - \frac{GM}{c^2}\left(\frac{1}{R} - \frac{1}{s}\right).$$

For a distant observer $s \to \infty$ and (field is assumed weak)

$$\frac{\nu_\infty}{\nu_0} \simeq 1 - \frac{GM}{Rc^2} \quad \longrightarrow \quad \frac{\nu_0}{\nu_\infty} = \underbrace{\left(1 - \frac{GM}{Rc^2}\right)^{-1}}_{\text{Expand}} \simeq 1 + \frac{GM}{Rc^2}$$

which is valid if $GM \ll Rc^2$. The corresponding *gravitational redshift* z for the weak field limit is

$$z \equiv \frac{\lambda_\infty - \lambda_0}{\lambda_0} = \frac{\nu_0}{\nu_\infty} - 1 \simeq \frac{GM}{Rc^2} \qquad \text{(weak field limit)}$$

where R is the radius of the star and M is its mass.

The full GR solution for the gravitational redshift in a spherical gravitational field is

$$1 + z = \left(1 - \frac{2GM}{Rc^2}\right)^{-1/2},$$

(derived later). Reduces to above in weak field.

EXAMPLE: for the white dwarf Sirius B

$$R = 5.85 \times 10^8\,\mathrm{cm} \qquad M = 1.95 \times 10^{33}\,\mathrm{g}$$

Inserting these values

$$z \simeq \frac{GM}{Rc^2} \simeq 2.5 \times 10^{-4}$$

The measured redshift is

$$z = 2.7 \pm 0.2 \times 10^{-4},$$

from displacement of spectral lines.

6.4.2 Gravitational Time Dilation

Gravitational redshift also may be viewed as *gravitational time dilation:*

$$\text{time} \propto \frac{1}{\text{frequency}}$$

One period of light wave = One clock tick

$$\frac{\Delta t_0}{\Delta t_\infty} \simeq \frac{\nu_\infty}{\nu_0} \simeq 1 - \frac{GM}{Rc^2} \qquad \text{(weak field limit)},$$

EXAMPLE: For surface of Sirius B

$$\frac{\Delta t_0}{\Delta t_\infty} \simeq 1 - \frac{GM}{Rc^2} \simeq 0.99972 \qquad (\sim \text{One second per hour slow})$$

The full GR solution for a spherical gravitational field gives (derived later)

$$\frac{\Delta t_0}{\Delta t_\infty} = \sqrt{1 - \frac{2GM}{Rc^2}}$$

Reduces to above weak-field result when the second term under the radical is small.

> Purely gravitational effect, independent of any special relativistic time dilation due to relative motion between source and observer *(GPS requires corrections for both)*.

6.5 Equivalence and Riemannian Manifolds

We cannot set up a global cartesian coordinate system on a curved surface. But if the metric takes the quadratic form

$$ds^2 = a(x,y)\,dx^2 + 2b(x,y)\,dxdy + c(x,y)\,dy^2$$

(2D for illustration), the geometry is *locally Euclidean*:

- Near any point *local cartesian coordinates* are valid.
- Circumference of a circle: $C = 2\pi r +$ higher-order terms
- Sum of the angles of a triangle: π + higher-order terms
- with *higher-order terms vanishing smoothly* as circles and triangles are decreased in size.

Such a space is termed a *Riemannian manifold,* with a corresponding *Riemannian metric.*

- The *spacetime metric,*

  $$ds^2 = g_{\mu\nu}dx^\mu dx^\nu,$$

  has such a quadratic form: *it is a Riemannian manifold.*

  - Strictly spacetime is *pseudo-Riemannian* due to indefinite metric (*Lorentzian signature*).
  - But in GR it is common to call it loosely "Riemannian".

- A Riemannian manifold is *locally euclidean* (and a pseudo-Riemannian manifold is *locally Minkowski*).
- Conversely, if the metric is Euclidean (Minkowski) around an arbitrary local point P_0 the space is necessarily Riemannian (pseudo-Riemannian).

The preceding considerations suggest that

$$\text{Gravity} \quad \longleftrightarrow \quad \text{Spacetime curvature}$$

and that

$$\underbrace{\text{Equivalence principle}}_{\text{gravity transformed away locally}} \quad \longleftrightarrow \quad \underbrace{\text{Riemannian geometry}}_{\text{spacetime locally flat}}$$

which in turn implies

$$\text{Gravitation} \quad \longleftrightarrow \quad \text{Riemannian Geometry}$$

This relationship was foreshadowed in the work of *Gauss and Riemann* during the 19th century:

- Gauss: all *inner (intrinsic) properties* of a curved surface are described by the derivatives $\partial\xi^\alpha/\partial x^\mu$ of the functions $\xi^\alpha(x)$ implementing the transformation between
  - a *general coordinate system* x^μ and
  - a *local Cartesian coordinate system* $\xi^\alpha(x)$.
- Because of equivalence, *all effects of a gravitational field* are contained in the *derivatives* $\partial\xi^\alpha/\partial x^\mu$ *of the* $\xi^\alpha(x)$.

Thus, the principle of equivalence finds its natural mathematical expression in Riemannian geometry

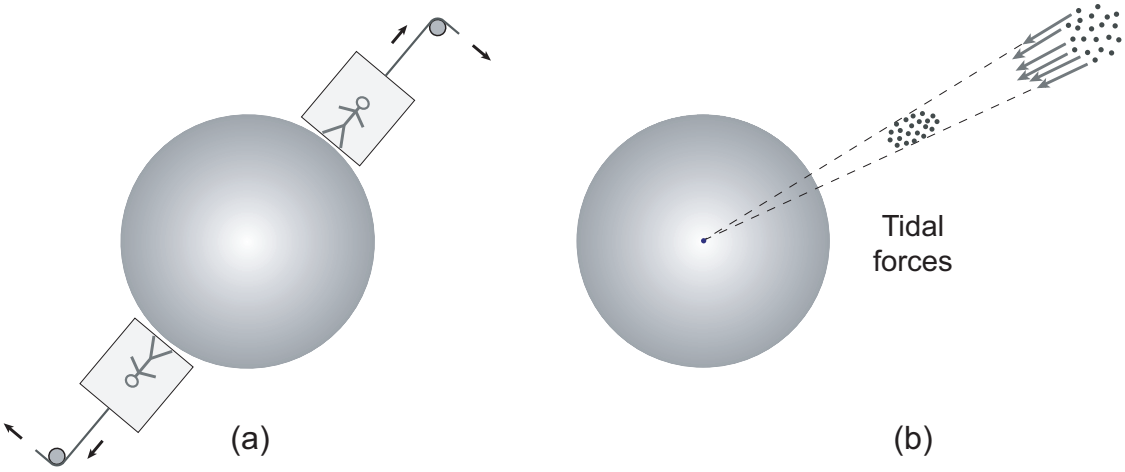


Figure 6.5: (a) The Einstein elevator in two different local inertial frames. (b) Tidal (differential gravitational) forces; an object experiences tidal forces if the gravitational force (magnitude or direction) is not the same for different parts of the object.

6.6 Local Inertial Frames and Inertial Observers

Equivalence: Elevator occupants on opposite sides of Earth may replace gravity by a *local acceleration* (Fig. 6.5a).

No Contradiction: The two elevator occupants in this case *cannot be in the same local inertial frame*.

Operational Definition of Local: Tidal effects (Fig. 6.5b) are negligible. (Later: quantitative definition in terms of spacetime curvature.)

The preceding ideas may be expressed more precisely by specifying exactly what is meant by a *local inertial frame or LIF.*

- By equivalence, at *each point* P of a curved manifold described by a metric $g_{\mu\nu}(x)$ a basis exists where the metric becomes the *constant Minkowski metric:*

$$g'_{\mu\nu}(x'_P) = \eta_{\mu\nu} = \text{diag}\,(-1,1,1,1).$$

- Specifically, because
  - $g_{\mu\nu}(x)$ at some arbitrary point $x = x_P$ is a *real symmetric matrix,*
  - there exists an *orthogonal transformation* that will *diagonalize it*.
- Once diagonalized each coordinate can be rescaled if needed to give the metric tensor $\eta_{\mu\nu} = \text{diag}\,(-1,1,1,1)$.
- Notice two important things, however:
  - Such a transformation *cannot change the signature* of the metric.
  - This transformation is *local to the point* P.

Generally a *different transformation* is required to diagonalize the metric at a different point P' if the space is curved.

Although $g'_{\mu\nu}(x'_P) = \eta_{\mu\nu}$ is valid *only at a point*,

- It is possible to choose the transformation so that the *first derivatives* of the metric evaluated at x_P vanish also:

$$g'_{\mu\nu}(x'_P) = \eta_{\mu\nu} = \text{diag}\,(-1,1,1,1) \qquad \left.\frac{\partial g'_{\mu\nu}}{\partial x'^{\alpha}}\right|_{x=x_P} = 0.$$

- A coordinate system satisfying these conditions is called a *local inertial frame (LIF)*.
- However, no transformation satisfying these conditions makes all *second derivatives* at x_P vanish.
- Thus, *"local"* means that
  - the observer *occupies a sufficiently small laboratory* (in both space and time) that
  - effects depending on *second derivatives of the metric* are negligible.
- Then up to first order *the freely-falling laboratory is flat Minkowski space*.
- Since second derivatives of the metric are associated with tidal forces, an alternative definition is that *in a LIF tidal forces are negligible*.

Thus the gravitational effects of spacetime curvature are expected to appear first in the *second derivatives of the metric*.

The preceding discussion means that *gravity cannot be identified with the Newtonian gravitational force*:

- By the equivalence principle the Newtonian gravitational force can be *transformed to zero* in a suitable reference frame.
- Thus Newtonian gravity is an *inertial force caused by observation in a non-inertial frame*, analogous to fictitious centrifugal or coriolis forces that appear in rotating frames.
- What *cannot be transformed away* are the tidal forces arising from the non-uniformity of the gravitational field, so in GR *tidal forces represent the true effect of gravity*.

In summary, by the equivalence principle there *are* inertial frames in the spacetime of general relativity. However,

- They are *local, freely-falling frames.*
- In the presence of gravity these local inertial frames at different points have *accelerations differing in both magnitude and direction*.
- Thus they *cannot be glued together to form a global inertial frame*.

In a gravitational field there are local inertial frames but *no global inertial frames*.

6.7 Lightcones in Curved Spacetime

Because a *local inertial frame* can be defined at each point of a curved spacetime,

> *General relativity inherits the local lightcone structure of special relativity.*

- This structure is a coordinate-invariant statement that the speed limit is c, so in general relativity
  - velocities (defined locally) satisfy $v \leq c$ and
  - the *local causal classification* of events into timelike, spacelike, and null carries over in curved spacetime.
- However, the *global organization* of lightcones in curved spacetime can lead to causal structure that does not exist for unaccelerated observers in Minkowski space.
- In flat spacetime the lightcone at one point can be obtained from one at another point by translation without change in geometry of the lightcone.
- In curved spacetime this is no longer true since the metric depends on location and the global causal structure can become complicated.

> An extreme consequence of global lightcone organization will be encountered when *event horizons of black holes* are considered in later chapters.

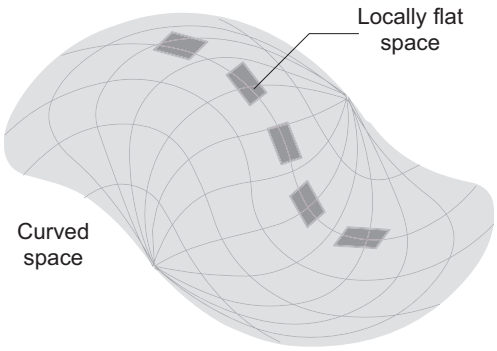


Figure 6.6: Curved spacetime and local flat inertial systems.

6.8 The Road to General Relativity

- *Einstein:* Inhomogeneity of gravitational field due to inhomogeneity of gravitating matter $\rightarrow$ *spacetime is curved*, with *curvature related to the distribution of matter*.
- *Equivalence:* Spacetime is a patchwork of *locally flat frames* meshed smoothly into a curved space (Fig. 6.6).
- *Key to relating curvature and matter distribution:* Connection between *equivalence and Riemannian geometry*.
- *Some Sewing Required:* Local Euclidean patches (where special relativity holds) must be *"stitched together" smoothly* to form a Riemannian manifold (Fig. 6.6).
- *General Relativity:* *A "stitching together"* by finding a (unique) Riemannian metric determined by a nonlinear relationship among mass, curvature, density.

The next two chapters will give the mathematical framework needed to implement this prescription.

Chapter 7

Curved Spacetime and General Covariance

In this chapter we generalize the the preceding discussion to extend covariance to more general curved spacetimes.

7.1 Covariance and Poincaré Transformations

- *Lorentz covariance* makes manifest that the principles of special relativity (invariance under Lorentz transformations) are obeyed by a set of equations.
- *Poincaré transformations:*
  - Six Lorentz transformations, plus
  - four possible uniform translations in space and time.

  Invariance under Poincaré transformations implies that
  - Physics does not depend on choice of coordinate system origin, orientation, ...
  - This implies *conservation laws* (energy, ...).
- Covariance with respect to Poincaré transformations is still insufficient to deal with gravity.
- We seek a more *general covariance,* valid for gravity.

General Covariance: a physical equation holds in a gravitational field provided that

- It holds in the absence of gravity (agrees with special relativity in flat spacetime).
- It maintains its form under the most general coordinate transformation $x \to x'$.

7.2 Curved Spacetime

The deflection of light in a gravitational field suggests that

> *Gravity is associated with the curvature of spacetime.*

Thus, let us consider the more general issue of *covariance in curved spacetime*.

7.3 Curved 2D Spaces and Gaussian Curvature

Gauss demonstrated that for 2-surfaces there is a single invariant (*Gaussian curvature*) characterizing the curvature.

- For a 2-D coordinate system (x^1, x^2) having a diagonal metric with non-zero elements g_{11} and g_{22}, the Gaussian curvature K is

$$K = \frac{1}{2g_{11}g_{22}} \times \left\{ -\frac{\partial^2 g_{22}}{(\partial x^1)^2} - \frac{\partial^2 g_{11}}{(\partial x^2)^2} + \frac{1}{2g_{11}} \left[\frac{\partial g_{11}}{\partial x^1}\frac{\partial g_{22}}{\partial x^1} + \left(\frac{\partial g_{11}}{\partial x^2}\right)^2 \right] + \frac{1}{2g_{22}} \left[\frac{\partial g_{11}}{\partial x^2}\frac{\partial g_{22}}{\partial x^2} + \left(\frac{\partial g_{22}}{\partial x^1}\right)^2 \right] \right\}.$$

The gaussian curvature is generally

- *Position-dependent*.
- An *intrinsic quantity* expressed entirely in terms of the metric for the space and its derivatives.

For the special case of *orthogonal coordinates* (x, y),

$$K(x_0, y_0) = \frac{1}{R_x(x_0) R_y(y_0)},$$

where

- $R_x(x_0)$ is the *radius of curvature* in the x direction and
- $R_y(y_0)$ is the *radius of curvature* in the y direction,

both evaluated at a point (x_0, y_0).

EXAMPLE: For a 2-sphere, $R_x = R_y \equiv R$ and

$$K = \frac{1}{R^2},$$

where R is the constant radius of the sphere.

The highly symmetric sphere has the same curvature at all points, but *in the general case curvature can vary from point to point*.

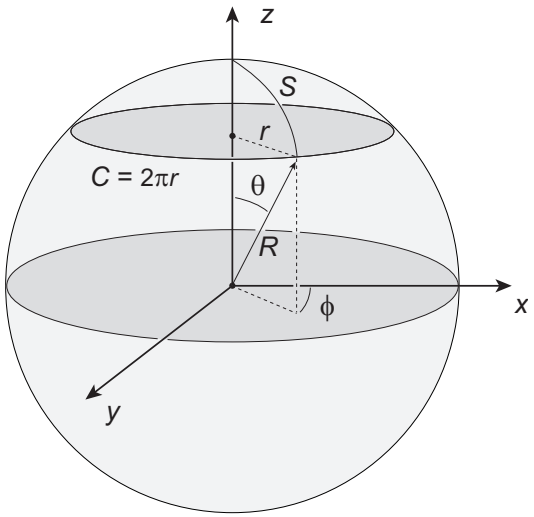


Figure 7.1: Measuring the circumference of a circle in curved space.

Consider the 2-sphere of Fig. 7.1, defined by

$$x^2 + y^2 + z^2 = R^2,$$

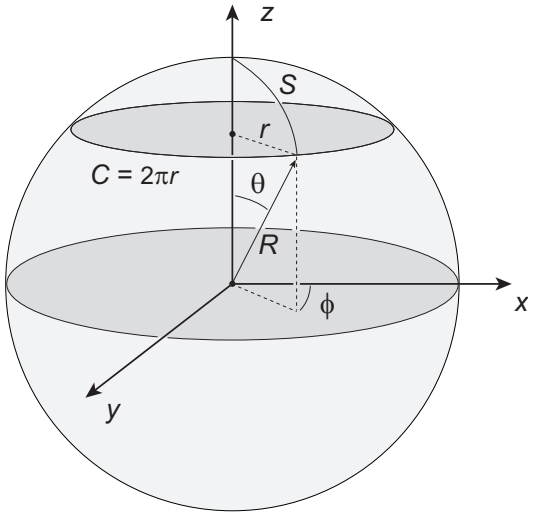
Let us use the circumference of a circle relative to that for flat space to measure deviation from flatness.

- A circle may be drawn in the 2D space by marking a *locus of points lying a constant distance S from a reference point*, chosen as the north pole in Fig. 7.1
- The angle θ subtended by S is S/R and

$$r = R\sin\theta = R\sin\left(\frac{S}{R}\right).$$

- Then the circumference of the circle is

$$C = 2\pi r = 2\pi R\sin\frac{S}{R} \simeq 2\pi S\left(1 - \frac{S^2}{6R^2} + \ldots\right).$$



- For flat space the circumference of the circle would just be $2\pi S$, so *higher-order terms measure the curvature*.
- We have for the gaussian curvature of the sphere $K = 1/R^2$. Substituting $R^2 = 1/K$ in

$$C = 2\pi r = 2\pi R \sin\frac{S}{R} = 2\pi S\left(1 - \frac{S^2}{6R^2} + \ldots\right).$$

  and solving for K in the limit $S \to 0$ gives

$$K = \lim_{S\to 0} \frac{3}{\pi}\left(\frac{2\pi S - C}{S^3}\right) = \lim_{S\to 0} \frac{6}{S^2}\left(1 - \frac{C}{2\pi S}\right).$$

  Thus, we may find the Gaussian curvature for a 2-D surface by *measuring the circumference of small circles*.
- Later we shall generalize the Gaussian curvature parameter for a 2D surface to a set of parameters (elements of the *Riemann curvature tensor*) that describe the curvature of 4D spacetime.

Notice in this and various other discussions that we often use the mental aid of embedding a surface in a higher-dimensional space in order to more easily visualize our arguments. However, it is important to emphasize that

The *intrinsic curvature* of a manifold can be

- determined *entirely by the properties of the manifold* itself,
- without reference to a higher-dimensional embedding space.

We may illustrate determining gaussian curvature intrinsically by using the *geometry of small circles* drawn on the unit 2-sphere (See Box 7.2 in book).

- The line element is

$$ds^2 = d\theta^2 + \sin^2\theta d\varphi^2.$$

in spherical polar coordinates

- In the coordinates (θ,φ) a line segment from $(0,0)$ to $(\lambda,0)$ has a length

$$S = \int \sqrt{ds^2} = \int_0^\lambda d\theta = \lambda,$$

since φ is constant so $d\varphi^2 = 0$.

- The set of points (λ,φ) with φ ranging from 0 to 2π then defines a circle of radius $S = \lambda$, with

$$C = \int \sqrt{ds^2} = \int_0^{2\pi} \sin\theta d\varphi = \sin\lambda \int_0^{2\pi} d\varphi = 2\pi\sin\lambda.$$

for the circumference C.

- Thus, with $S = \lambda$ the Gaussian curvature is

$$K = \operatorname*{Lim}_{S\to 0} \frac{6}{S^2}\left(1 - \frac{C}{2\pi S}\right) = \operatorname*{Lim}_{\lambda\to 0} \frac{6}{\lambda^2}\left(1 - \frac{\sin\lambda}{\lambda}\right).$$

Since our interest is in the limit $\lambda \to 0$, we expand $\sin\lambda$ in a power series,

$$\frac{\sin\lambda}{\lambda} = \frac{1}{\lambda}\left(\lambda - \frac{\lambda^3}{3!} + \ldots\right) \simeq 1 - \frac{\lambda^2}{6}.$$

This gives

$$K = \underset{\lambda\to 0}{\text{Lim}}\, \frac{6}{\lambda^2}\left[1 - \left(1 - \frac{\lambda^2}{6}\right)\right] = 1.$$

This is the expected result, since it was noted earlier that the Gaussian curvature of a 2-sphere having radius R is equal to $1/R^2$ and $R = 1$ (unit sphere) has been assumed.

7.3.1 Distance Intervals in Curved Spacetime

In curved spacetime the interval between two events may be expressed as

$$ds^2 = g_{\alpha\beta}(x)dx^\alpha dx^\beta,$$

where the metric tensor $g_{\alpha\beta}(x)$ in a curved spacetime generally

- has a *more complicated form* than that for Minkowski space, and
- is a *function of the spacetime coordinates*.

7.4 A Covariant Description of Matter

Curved spacetime is responsible for gravity and mass, energy, and pressure are responsible for curving spacetime.

- Therefore, it is critical to describe the distribution of these quantities and their coupling to gravity covariantly.
- To that end, it is convenient to introduce the *stress–energy (*or *energy–momentum) tensor* $T^{\mu\nu}$, with components

  1. $T^{00} = \varepsilon$ (energy density)
  2. $T^{ii} = P^i$ (pressure in i direction; equivalently, momentum components per unit area)
  3. T^{0i} (energy flux in the direction i)
  4. T^{i0} (momentum density in the direction i)
  5. $T^{ij}(i \neq j)$ (shear of the pressure component P^i in the j direction).

- By physical arguments the tensor $T^{\mu\nu}$ is *symmetric, with 10 independent components.*

Physically-meaningful results from general relativity require that the form of $T^{\mu\nu}$ be constrained further by hypotheses that are discussed below.

It is standard to impose a set of *energy conditions* on $T^{\mu\nu}$. basically

- These are *assumptions* about the way that any reasonable form of matter should behave, and that
- are obeyed by all presently-known forms of matter.

Three common energy conditions are

1. ***Weak energy condition:*** $T_{\mu\nu}u^\mu u^\nu \geq 0$ for any unit timelike vector u^μ, which means that *the energy density seen by any observer may not be negative*.
2. ***Strong energy condition:*** $T_{\mu\nu}u^\mu u^\nu + \frac{1}{2}T^\mu{}_\mu \geq 0$ for any unit timelike vector u^μ, implying physically that *the energy density plus the sum of the principle pressures must be non-negative*.
3. ***Null energy condition:*** $T_{\mu\nu}k^\mu k^\nu \geq 0$ for any null vector k^μ, which means physically that *the sum of the energy density plus any of the principle pressures may not be negative*.

These conditions are more in the nature of *recipes based on classical experience* about the way matter should behave rather than physical law.

We simplify by restricting attention to *perfect fluids,* which *neglect energy transport and viscosity effects*.

- For *flat spacetime* the most general perfect-fluid stress–energy tensor consistent with Lorentz invariance is

$$T^{\mu\nu} = (\varepsilon + P)u^\mu u^\nu + P\eta^{\mu\nu} \qquad \text{(flat spacetime)},$$

  where in this equation

  - $\eta^{\mu\nu}$ is the Minkowski metric,
  - P is pressure,
  - $\varepsilon = \rho c^2$ is energy density, and
  - $u^\mu = dx^\mu/d\tau$ is the 4-velocity.

- Conservation of 4-momentum may be expressed by

$$\partial_\mu T^{\mu\nu} = 0 \qquad \text{(flat spacetime)}.$$

- The most general $T^{\mu\nu}$ in *curved spacetime* is

$$T^{\mu\nu} = (\varepsilon + P)u^\mu u^\nu + Pg^{\mu\nu} \qquad \text{(curved spacetime)},$$

  where $g^{\mu\nu}$ is the metric.

- The generalization of $\partial_\mu T^{\mu\nu} = 0$ in curved spacetime is

$$\nabla_\mu T^{\mu\nu} = 0 \qquad \text{(curved spacetime)}.$$

These equations imply a basic difference between sources of gravity in GR and Newtonian theories.

General relativity and Newtonian gravity differ quantitatively in their predictions for physical observables but they also *differ fundamentally in their physical interpretation*.

- The form of the stress–energy tensor

$$T^{\mu\nu} = (\varepsilon + P)u^\mu u^\nu + Pg^{\mu\nu} \quad \text{(curved spacetime)},$$

  points to an essential difference between Einstein gravity and the Newtonian theory.

  - *All components of the stress–energy tensor* contribute to the curvature and thence to gravity.
  - Thus *energy, mass, and pressure* are all sources of the gravitational field.
  - *Only mass* is a source of Newtonian gravity.

  By smuggling in $E = mc^2$ from special relativity we can (by a stretch) view energy as a source for Newton's gravity, but not pressure.

- But what about the role of pressure in stabilizing stars against contraction in Newtonian gravity?

  - In that case forces opposing gravity are not produced by pressure but by a *pressure gradient.*
  - In contrast, *gravity couples directly to the magnitude of the local pressure* in GR.
  - Thus in GR there can be forces associated with pressure even if there is no pressure gradient

- In a universe having a finite but constant pressure the existence of the pressure could still be detected by its (general relativistic) gravitational effect.
- This is precisely the nature of the cosmological *vacuum energy* to be discussed later.
- That increasing the pressure increases the strength of gravity in GR also has *implications for the gravitational stability of stars*.
- Specifically, it suggests a limit beyond which even increasing the pressure by an arbitrary amount cannot stop a massive object from collapsing under the influence of its own gravity.
- This will lead soon to the idea of a *black hole.*

Another important difference implied by Einstein gravity concerns

$$\partial_\mu T^{\mu\nu} = 0 \qquad \text{(flat spacetime)}.$$

$$\nabla_\mu T^{\mu\nu} = 0 \qquad \text{(curved spacetime)},$$

These appear to be *similar formally* but their *physical meanings are different*.

- In Newtonian gravity energy and momentum are conserved and in flat spacetime $\partial_\mu T^{\mu\nu} = 0$ expresses a *conservation law* for 4-momentum.
- In a curved spacetime the constraint $\nabla_\mu T^{\mu\nu} = 0$ does not imply a conservation law because

  > *The gravitational energy is not included in the stress–energy tensor.*

- That is, there is no well-defined concept of local energy and momentum conservation in general relativity, ultimately because it is difficult to construct a sensible local expression for gravitational energy.

  > *Remember:* By equivalence gravity can be *transformed away at a point*.

- The best that one can achieve is *approximate 4-momentum conservation* over a *finite volume*.

7.4.1 Local Energy Conservation

As indicated above, there is no well-defined idea of local energy and momentum conservation in general relativity.

- This is a consequence of the *equivalence principle*,
  - which requires that *all effects of gravity vanish over a small enough region*,
  - but it may also be viewed more fundamentally:
- In non-GR physics conservation laws like for momentum or energy follow from spacetime symmetry (*Noether's theorem*, which is discussed in the book).
  - For example, momentum is conserved locally because of spatial translational invariance.
  - Since all physical systems are expected to be translationally invariant, the law of momentum conservation is *always valid in non-GR physics*.
- But in general relativity *spacetime is the solution,* not a pre-defined stage on which physics plays out.
- Hence *No local symmetries are guaranteed to be common to all GR spacetimes*.
- However, energy is approximately conserved when averaged over a large enough region of spacetime
- Total energy of a spacetime is well-defined if it is asymptotically flat (becomes flat on the boundary).

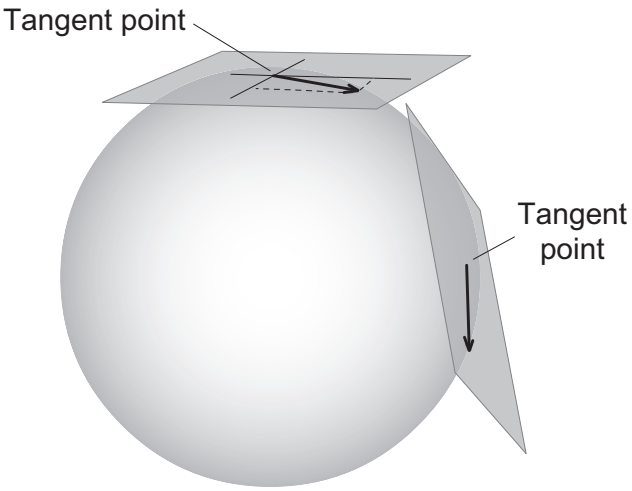


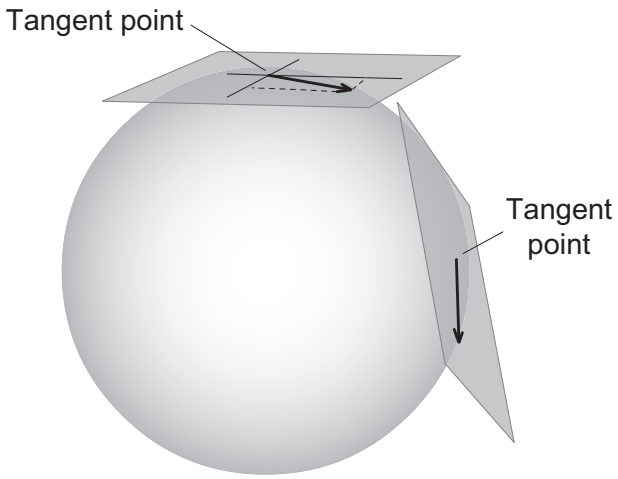
Figure 7.2: Tangent planes and vectors in curved spaces.

7.5 Covariant Derivatives and Parallel Transport

- Covariant derivatives have a geometrical interpretation associated with comparison of vectors located at two different spacetime points.
- Constructing the derivative of a vector requires taking the *difference of vectors at two different points*.

Recall: Vectors are defined

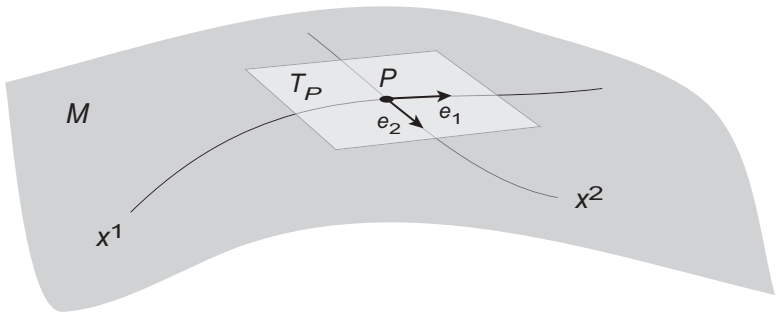
- in the *tangent space*, and each point
- has a *different tangent space* (Fig. 7.2).

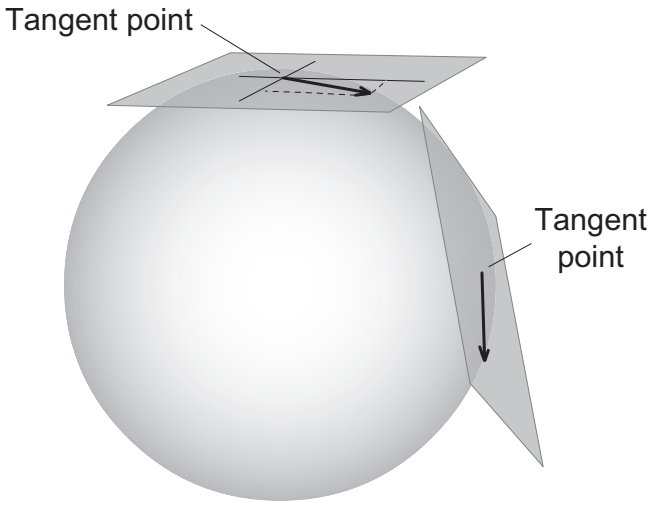


Recall that the figure above is conceptually useful, but

- defining the tangent space by a local flat coordinate system at a point is an *intrinsic process with respect to the original manifold* and
- *does not require embedding* in a higher-dimensional manifold.

Directional derivatives evaluated in the *intrinsic manifold* may be used to define the tangent space.





Parallel transport of vectors is necessary to compare two vectors at different points (e.g., to define derivatives).

- For a flat space the tangent space corresponds with the space itself.
- Thus in a flat space we can just move one vector, keeping its orientation fixed with respect to a global set of coordinate axes, to the position of the other vector and compare.
- On a curved surface this issue is more complicated because the vectors at two points are *defined in different spaces* (see the figure above).

We would like to construct a new derivative operation on tensors that fulfills *three requirements:*

- The operation should exhibit the properties expected of a derivative, such as the *Leibniz rule for the derivative of a product*.
- A derivative of a tensor should *transform as a tensor*.
- The derivative should represent the *change of the whole tensor*, not just its components.

As we will now demonstrate geometrically, a derivative satisfying these requirements corresponds to the *covariant derivative* already introduced.

To be definite, let us illustrate for derivatives of vectors. The formal definition of the derivative that we seek is

$$\nabla_\nu V^\mu u^\nu e_\mu = \operatorname*{Lim}_{\delta\lambda\to 0}\left(\frac{V_\parallel(\lambda+\delta\lambda)-V(\lambda)}{\delta\lambda}\right),$$

- $V_\parallel(\lambda+\delta\lambda)$ represents the vector $V(\lambda)$ parallel transported along the curve parameterized by λ from λ to $\lambda+\delta\lambda$.
- A basis vector field $e_\mu(\lambda)$ is assumed defined in the vicinity of the curve.
- $u^\nu \equiv dx^\nu/d\lambda$ (chain rule conversion)
- We use the symbol ∇_ν, anticipating equivalence with the covariant derivative already defined.

We must now understand how to transport a vector from λ to $\lambda+\delta\lambda$ while preserving its intrinsic properties.

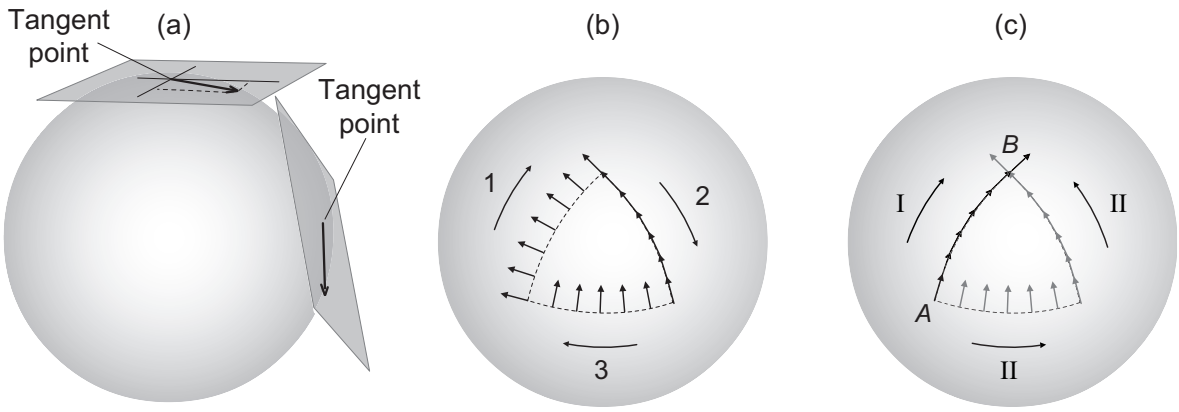


Figure 7.3: (a) Tangent spaces and vectors in curved spaces (see Fig. 3.1). (b) Parallel transport of a vector in a closed path on a curved surface. The vector rotates by 90° for parallel transport on the closed path $1 \to 2 \to 3$. (c) Dependence of parallel transport on the path. Parallel transport from A to B on the direct path labeled I rotates the vector by a different amount than for parallel transport from A to B on the two-segment path labeled II.

- Natural notion of parallel transport: *keep the vector parallel to itself in infinitessimal steps; see Fig. 7.3 (space locally euclidean)*.
- As Fig. 7.3 illustrates, parallel transport of vectors on a curved surface is generally *path-dependent*.
- Hence parallel transport in curved spaces is *not unique* and requires a *prescription*.
- The apparent rotation of a vector when parallel transported around a closed path *measures curvature of the manifold.*
- For 2D, *rotation is proportional to Gaussian curvature*.
- In 4D spacetime we will find that the rotation depends on the *Riemann curvature tensor*.

7.5.1 The Affine Connection and Covariant Derivatives

Generalizing the discussion in Chapter 2 to curved spacetime,

- Differentiation of a 4-vector $V = V^\mu e_\mu$ with respect to a parameter λ gives two contributions,

$$\frac{dV}{d\lambda} = \frac{d}{d\lambda}\left(V^\mu e_\mu\right) = \frac{dV^\mu}{d\lambda} e_\mu + \frac{de_\mu}{d\lambda} V^\mu,$$

- where the first term represents the *change in the vector components* in a fixed basis e_μ and
- the second term represents the *change in the basis* in moving from one point to another.

Introducing $u^\mu \equiv dx^\mu/d\lambda$ and using $\frac{dV}{d\lambda} = \frac{\partial}{\partial x^\nu}\frac{dx^\nu}{d\lambda}$,

$$\begin{aligned}\frac{dV}{d\lambda} &= \frac{\partial}{\partial x^\nu}\frac{dx^\nu}{d\lambda} V^\mu e_\mu + V^\mu \frac{dx^\nu}{d\lambda}\frac{\partial}{\partial x^\nu} e_\mu \\ &= \partial_\nu V^\mu u^\nu e_\mu + V^\mu u^\nu \partial_\nu e_\mu.\end{aligned}$$

For infinitesimal separation between x and x' the second term will be linear in V^μ, so expand in the vector basis e_μ,

$$\frac{dV}{d\lambda} = \left(\partial_\nu V^\mu + \Gamma^\mu_{\alpha\nu} V^\alpha\right) u^\nu e_\mu,$$

- $\Gamma^\mu_{\alpha\nu}$ is called the *affine connection coefficient* or often just *the connection.*
- The $\Gamma^\mu_{\alpha\nu}$ define a *connection between tangent spaces at two different points* of the manifold.

- This permits a vector in the tangent space at one point to be parallel transported and compared with a vector defined in the (different) tangent space at another point.
- Use of the same notation for the connection coefficient as for the Christoffel symbol is deliberate because *the connection coefficient and the Christoffel symbol may be viewed as equivalent*.

  Specifically, the Christofel symbols are *connection coefficients expressed in a coordinate basis*.

- Despite its indices, the Christoffel symbol *does not transform as a tensor*.

  (That's the point! If it *did* transform as a tensor it would be of no use to us.)

It will be shown later that

- The affine connection $\Gamma^{\mu}_{\alpha\nu}$ can be *constructed from the metric tensor and its derivatives*,
- $\Gamma^{\mu}_{\alpha\nu}$ *vanishes if the metric is constant*, and
- The *Riemann tensor* describing the local intrinsic curvature of the spacetime may be constructed from $\Gamma^{\mu}_{\alpha\nu}$.

Thus the affine connection is central to developing the theory of general relativity.

Assuming equivalence of the Christoffel coefficients and connection coefficients,

- comparison of with the expression from Chapter 3 for the covariant derivative,

$$\nabla_\mu A^\lambda = \partial_\mu A^\lambda + \Gamma^\lambda_{\alpha\mu} A^\alpha,$$

with the above definition

$$\frac{dV}{d\lambda} = \underbrace{\left(\partial_\nu V^\mu + \Gamma^\mu_{\alpha\nu} V^\alpha\right)}_{\text{Covariant derivative}} u^\nu e_\mu,$$

- indicates that the quantity in parentheses in the above equation corresponds to the *covariant derivative of the vector* V^μ,

$$\nabla_\nu V^\mu = \partial_\nu V^\mu + \Gamma^\mu_{\alpha\nu} V^\alpha.$$

- The covariant derivative represents the change of the *whole vector* V, not only its components V^μ.

Thus the derivative representing the change of a vector under parallel transport along a path is the covariant derivative evaluated along that path.

7.5.2 Absolute Derivatives and Parallel Transport

Consider in more detail parallel transport of a vector.

- For *euclidean space*, parallel transport along a path parameterized by λ means that
  - The length and direction of the vector (referenced to a universal cartesian coordinate system) don't change.
  - Thus *in flat space* the components of the vector satisfy $dV^\mu/d\lambda = 0$ under parallel transport
- For a curved manifold *both the components and the basis vectors change* and this condition generalizes to

$$\nabla_\nu V^\mu u^\nu = \partial_\nu V^\mu u^\nu + \Gamma^\mu_{\alpha\nu} V^\alpha u^\nu = 0,$$

  where $V = V^\mu e_\mu$ and $u^\nu \equiv dx^\nu/d\lambda$.
- But $\partial_\nu V^\mu u^\nu = dV^\mu/d\lambda$ by the chain rule and comparison with the definition of the *absolute derivative* (Ch. 3),

$$\frac{DV^\mu}{D\lambda} \equiv \frac{dV^\mu}{d\lambda} + \Gamma^\mu_{\alpha\nu} V^\alpha \frac{dx^\nu}{d\lambda},$$

  indicates that along a path parameterized by λ,

$$\frac{DV^\mu}{D\lambda} = 0$$

  is the condition for parallel transport of a vector.
- By this prescription, at each infinitesimal step the vector is parallel transported.

Summary of properties of the affine connection $\Gamma^{\nu}_{\mu\alpha}$.

- $\Gamma^{\nu}_{\mu\alpha}$ is called the
  - *affine connection*, or
  - the *connection coefficient*, or
  - the *metric connection,* or just
  - *the connection*.
- The equivalence of notation for the affine connection and the Christoffel symbol introduced earlier is deliberate because they *may be viewed as equivalent*.
- $\Gamma^{\nu}_{\mu\alpha}$ can be constructed from the metric and its derivatives.
- $\Gamma^{\nu}_{\mu\alpha}$ vanishes in a space with constant metric.
- $\Gamma^{\nu}_{\mu\alpha}$ does not follow from differential geometry of the manifold but is *additional imposed structure* that specifies how tangent spaces at different points are related.
- The *Riemann curvature tensor* describing the local intrinsic curvature of the spacetime may be constructed from the affine connection.

Thus the affine connection is central to

- defining the covariant derivative,
- implementing parallel transport of tensors, and
- measuring quantitatively the curvature of a manifold.

It will play an important role in general relativity.

7.6 Gravity and Curved Spacetime

Free Particle: A particle moving solely under the influence of gravity is termed a *free particle* in general relativity, because

- The classical gravitational force will be replicated by particles propagating with *no forces acting on them*, but
- the propagation is in a *curved spacetime*.

Equivalence Principle: in a freely-falling coordinate system labeled by coordinates ξ^μ, the special theory of relativity is valid and the equation of motion is given by

$$\frac{d^2\xi^\mu}{d\tau^2} = 0,$$

(the special relativistic generalization of Newton's second law). The proper time interval $d\tau$ is

$$d\tau^2 = \eta_{\mu\nu} d\xi^\mu d\xi^\nu$$

and the Minkowski metric is defined by

$$\eta_{\mu\nu} = \text{diag}\,(-1,1,1,1),$$

which *by equivalence* is valid in the coordinates ξ^μ.

Introduce another *arbitrary* coordinate system x^μ (*not necessarily inertial*).

- The freely-falling coordinates ξ^μ are functions of the new coordinates, $\xi^\mu = \xi^\mu(x)$, and by the chain rule

$$\frac{d^2\xi^\alpha}{d\tau^2} = \frac{d}{d\tau}\left(\frac{d\xi^\alpha}{d\tau}\right) = \underbrace{\frac{d}{d\tau}\left(\overbrace{\frac{\partial\xi^\alpha}{\partial x^\mu}}^{\text{Chain rule}}\frac{dx^\mu}{d\tau}\right)}_{\text{Derivative of product}} = 0$$

$$\frac{\partial\xi^\alpha}{\partial x^\mu}\frac{d^2x^\mu}{d\tau^2} + \frac{d}{d\tau}\left(\frac{\partial\xi^\alpha}{\partial x^\mu}\right)\frac{dx^\mu}{d\tau} = 0$$

$$\frac{\partial\xi^\alpha}{\partial x^\mu}\frac{d^2x^\mu}{d\tau^2} + \underbrace{\frac{\partial}{\partial x^\nu}\left(\frac{\partial\xi^\alpha}{\partial x^\mu}\right)\frac{dx^\nu}{d\tau}}_{\text{Chain rule}}\frac{dx^\mu}{d\tau} = 0$$

$$\frac{\partial\xi^\alpha}{\partial x^\mu}\frac{d^2x^\mu}{d\tau^2} + \frac{\partial^2\xi^\alpha}{\partial x^\mu\partial x^\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau} = 0.$$

- Multiply by $\partial x^\lambda/\partial\xi^\alpha$ (note the implied sum on α)

$$\underbrace{\frac{\partial x^\lambda}{\partial\xi^\alpha}\frac{\partial\xi^\alpha}{\partial x^\mu}}_{\delta^\lambda_\mu}\frac{d^2x^\mu}{d\tau^2} + \underbrace{\frac{\partial x^\lambda}{\partial\xi^\alpha}\frac{\partial^2\xi^\alpha}{\partial x^\mu\partial x^\nu}}_{\equiv\Gamma^\lambda_{\mu\nu}}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau} = 0$$

to obtain the *geodesic equation*,

$$\frac{d^2x^\lambda}{d\tau^2} + \Gamma^\lambda_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau} = 0 \qquad \Gamma^\lambda_{\mu\nu} \equiv \frac{\partial x^\lambda}{\partial\xi^\alpha}\frac{\partial^2\xi^\alpha}{\partial x^\mu\partial x^\nu} \quad \text{(connection)}$$

The *proper time interval* in this coordinate system is

$$\begin{aligned} d\tau^2 &= \eta_{\alpha\beta} d\xi^\alpha d\xi^\beta \\ &= \eta_{\alpha\beta} \underbrace{\frac{\partial \xi^\alpha}{\partial x^\mu} dx^\mu \frac{\partial \xi^\beta}{\partial x^\nu}}_{\text{chain rule}} dx^\nu \\ &= \underbrace{\eta_{\alpha\beta} \frac{\partial \xi^\alpha}{\partial x^\mu} \frac{\partial \xi^\beta}{\partial x^\nu}}_{\equiv g_{\mu\nu}} dx^\mu dx^\nu . \end{aligned}$$

Thus, the proper time interval may be written as

$$d\tau^2 = g_{\mu\nu} dx^\mu dx^\nu ,$$

where the metric tensor is defined by

$$g_{\mu\nu} = \eta_{\alpha\beta} \frac{\partial \xi^\alpha}{\partial x^\mu} \frac{\partial \xi^\beta}{\partial x^\nu} .$$

It is clear that

- $g_{\mu\nu} \neq \eta_{\mu\nu}$, and generally
- $g_{\mu\nu} = g_{\mu\nu}(x)$ is a *function of the spacetime coordinates* because
- $\partial \xi^\alpha / \partial x^\mu$ generally *depend on the spacetime coordinates*.

Much of the mathematics of general relativity lies in the discipline of *differential geometry.* However,

- The affine connection is *not part of differential geometry* since
  - it is *not a natural consequence of differential structure on the manifold,* and
  - it is in fact *not even a tensor*.
- The affine connection is an *augmentation of differential geometry* that
  - gives "shape and curvature" to a manifold;
  - it is a *defined rule* for parallel transport on curved surfaces.
- The affine connection generally is *not unique* because we can define many notions of parallel transport for a curved surface.
- Nevertheless, we shall see that

> Under conditions that are *assumed* to be satisfied in general relativity, the affine connection *is uniquely determined by the metric tensor.*

7.7 The Local Inertial Coordinate System

We now demonstrate explicitly that *the affine connection* $\Gamma^{\lambda}_{\mu\nu}$ *in an arbitrary coordinate system* x^{μ} *defines the local inertial coordinates* ξ^{α} at any point X. Multiply

$$\Gamma^{\lambda}_{\mu\nu} = \frac{\partial x^{\lambda}}{\partial \xi^{\alpha}} \frac{\partial^2 \xi^{\alpha}}{\partial x^{\mu} \partial x^{\nu}}.$$

through by $\partial \xi^{\beta} / \partial x^{\lambda}$ and utilize

$$\frac{\partial \xi^{\beta}}{\partial x^{\lambda}} \frac{\partial x^{\lambda}}{\partial \xi^{\alpha}} = \delta^{\beta}_{\alpha}$$

to give the differential equation for the inertial coordinates

$$\frac{\partial^2 \xi^{\beta}}{\partial x^{\mu} \partial x^{\nu}} = \frac{\partial \xi^{\beta}}{\partial x^{\lambda}} \Gamma^{\lambda}_{\mu\nu}.$$

This has a *power series solution* near the point X

$$\xi^{\alpha}(x) = \xi^{\alpha}(X) + \frac{\partial \xi^{\alpha}(X)}{\partial x^{\mu}} (x^{\mu} - X^{\mu}) + \frac{1}{2} \frac{\partial \xi^{\alpha}(X)}{\partial x^{\lambda}} \Gamma^{\lambda}_{\mu\nu} (x^{\mu} - X^{\mu})(x^{\nu} - X^{\nu}) + \ldots$$

Thus, $\Gamma^{\lambda}_{\mu\nu}$ and the partial derivatives $\partial \xi / \partial x$ at the point X *determine the local inertial coordinates* ξ^{α} up to order $(x - X)^2$.

> *This is sufficient*, since the inertial coordinates are valid only in the vicinity of the point X.

7.8 The Affine Connection and the Metric Tensor

We have seen from the preceding derivation that

- The affine connection $\Gamma^{\lambda}_{\mu\nu}$ determines the gravitational force through the *geodesic equation*.
- Thus, *the affine connection is the gravitational field.*
- The *metric tensor* determines the properties of the interval $d\tau$ through
$$d\tau^2 = g_{\mu\nu} dx^{\mu} dx^{\nu}.$$
- This suggests that the effect of gravity is determined by
  - the *affine connection* $\Gamma^{\lambda}_{\mu\nu}$ and
  - the *metric tensor* $g_{\mu\nu}$.
- Now we show that, in fact,
  - the metric tensor $g_{\mu\nu}$ alone *determines the full effect of gravity* because
  - the connection $\Gamma^{\lambda}_{\mu\nu}$ can be expressed in terms of *the metric tensor and its derivatives*.
- Thus, we shall now show that

> *The metric tensor may be viewed as the gravitational potential.*

Let's first differentiate

$$g_{\mu\nu} = \eta_{\alpha\beta}\frac{\partial\xi^\alpha}{\partial x^\mu}\frac{\partial\xi^\beta}{\partial x^\nu}$$

with respect to x^λ,

$$\frac{\partial g_{\mu\nu}}{\partial x^\lambda} = \frac{\partial^2\xi^\alpha}{\partial x^\lambda \partial x^\mu}\frac{\partial\xi^\beta}{\partial x^\nu}\eta_{\alpha\beta} + \frac{\partial\xi^\alpha}{\partial x^\mu}\frac{\partial^2\xi^\beta}{\partial x^\lambda \partial x^\nu}\eta_{\alpha\beta}.$$

But we have shown that the *inertial coordinates* ξ^α obey the differential equation

$$\frac{\partial^2\xi^\beta}{\partial x^\mu \partial x^\nu} = \frac{\partial\xi^\beta}{\partial x^\lambda}\Gamma^\lambda_{\mu\nu}.$$

Inserting this in the preceding equation gives

$$\begin{aligned}\frac{\partial g_{\mu\nu}}{\partial x^\lambda} &= \Gamma^\rho_{\lambda\mu}\underbrace{\frac{\partial\xi^\alpha}{\partial x^\rho}\frac{\partial\xi^\beta}{\partial x^\nu}\eta_{\alpha\beta}}_{g_{\rho\nu}} + \Gamma^\rho_{\lambda\nu}\underbrace{\frac{\partial\xi^\beta}{\partial x^\rho}\frac{\partial\xi^\alpha}{\partial x^\mu}\eta_{\alpha\beta}}_{g_{\rho\mu}} \\ &= \Gamma^\rho_{\lambda\mu}g_{\rho\nu} + \Gamma^\rho_{\lambda\nu}g_{\rho\mu},\end{aligned}$$

where

$$g_{\mu\nu} = \eta_{\alpha\beta}\frac{\partial\xi^\alpha}{\partial x^\mu}\frac{\partial\xi^\beta}{\partial x^\nu}$$

has been used.

As described in more detail in the book chapter, this equation may be solved by *switching indices and exploiting that* $\Gamma^{\sigma}_{\lambda\mu}$ *and* $g_{\mu\nu}$ *are symmetric under exchange of lower indices* to give

$$\begin{aligned}\Gamma^{\sigma}_{\lambda\mu} &= \frac{1}{2}g^{\nu\sigma}\left(\frac{\partial g_{\mu\nu}}{\partial x^{\lambda}} + \frac{\partial g_{\lambda\nu}}{\partial x^{\mu}} - \frac{\partial g_{\mu\lambda}}{\partial x^{\nu}}\right) \\ &= \frac{1}{2}g^{\nu\sigma}\left(g_{\mu\nu,\lambda} + g_{\lambda\nu,\mu} - g_{\mu\lambda,\nu}\right).\end{aligned}$$

Therefore the connection—and hence the gravitational field—are *determined entirely by the metric tensor and its derivatives*:

> *In general relativity, the metric tensor is the source of the gravitational field.*

7.9 Uniqueness of the Affine Connection

The affine connection is

- an additional feature *imposed on the differential structure of a manifold*
- through a definition that generally *is not unique.*
- However, if a manifold has both
  - a *metric* and
  - a *connection*

  defined for it,
- one usually makes certain *compatibility demands* that constrain the connection.
- If the manifold has a metric tensor, the divergence of a vector field and the metric are said to be *compatible* if the *inner (scalar) product of two arbitrary vectors is preserved under parallel transport.*
- In proving preceding results we have assumed *symmetry of* $\Gamma^{\lambda}_{\mu\nu}$ *in its lower indices*.
- The *torsion tensor* is defined by

$$T^{\lambda}_{\mu\nu} = \overbrace{\underbrace{\Gamma^{\lambda}_{\mu\nu}}_{\text{not tensor}} - \underbrace{\Gamma^{\lambda}_{\nu\mu}}_{\text{not tensor}}}^{\text{tensor}} .$$

  It measures *deviation from symmetry in the lower indices*.

The connection defined on a manifold with metric $g_{\mu\nu}$ is *unique* and *determined completely by the metric* if

1. The manifold is *torsion-free:* $T^{\lambda}_{\mu\nu} = 0$.
2. The *covariant derivative of the metric tensor vanishes on the entire manifold*,

$$\nabla_\alpha g_{\mu\nu} = 0,$$

   which is sufficient to ensure that
3. *the scalar product of vectors is preserved under parallel transport.*

In this case the connection is termed a *metric connection*.

The previous result that

$$\Gamma^{\sigma}_{\lambda\mu} = \frac{1}{2} g^{\nu\sigma} \left(\frac{\partial g_{\mu\nu}}{\partial x^\lambda} + \frac{\partial g_{\lambda\nu}}{\partial x^\mu} - \frac{\partial g_{\mu\lambda}}{\partial x^\nu} \right)$$

determines the affine connection uniquely in terms of the metric tensor

- is a consequence of the *assumptions* (1) and (2) above.
- These assumptions are then *justified after the fact* by concordance of the resulting theory and observations.

Chapter 8

The General Theory of Relativity

We shall now employ the central ideas introduced in the previous two chapters:

- The metric and curvature of spacetime
- The principle of equivalence
- The principle of general covariance

to construct the general theory of relativity and the corresponding theory of gravitation.

> We know that the weak-field, low-velocity limit of this theory must be Newtonian gravitation, so we begin by asking what the weak-field limit of a covariant theory of gravity would look like.

8.1 Weak Field Limit

We begin by considering the weak field limit of Einstein's theory as a guide to what the full theory should look like.

In the *weak field limit* we should recover *Newton's gravitational theory*.

The Newtonian gravitational field may be derived from a scalar potential φ that obeys the Poisson equation,

$$\nabla^2\varphi = 4\pi G\rho \qquad \nabla \equiv \hat{\imath}\frac{\partial}{\partial x} + \hat{\jmath}\frac{\partial}{\partial y} + \hat{k}\frac{\partial}{\partial z}$$

- The Newtonian equation of motion is then

$$\frac{d^2x^i}{dt^2} = F^i = -\frac{\partial\varphi}{\partial x^i},$$

  where $\boldsymbol{F}$ is the gravitational force.

- For a point-like mass M the potential is

$$\varphi = -\frac{GM}{r}$$

Earlier we showed that

$$\frac{d^2x^\lambda}{d\tau^2} + \Gamma^\lambda_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau} = 0 \qquad \text{(geodesic equation)}$$

For the special case of *vanishing gravity:*

1. The metric is flat

$$g_{\mu\nu}(x) = \eta_{\mu\nu} = \text{constant},$$

   in which case $\partial g_{\mu\nu}(x)/\partial x^\alpha = 0$.

2. The affine connection vanishes

$$\Gamma^\sigma_{\lambda\mu} = \tfrac{1}{2}g^{\nu\sigma}\left(\frac{\partial g_{\mu\nu}}{\partial x^\lambda} + \frac{\partial g_{\lambda\nu}}{\partial x^\mu} - \frac{\partial g_{\mu\lambda}}{\partial x^\nu}\right) = 0.$$

3. Covariant derivatives equal partial derivatives.

4. The equation of motion becomes that of a free particle in Minkowski space:

$$\frac{d^2x^\lambda}{d\tau^2} = 0.$$

Generally though, *space is curved by mass*,

- which produces gravity, and
- the second term in the geodesic equation does not vanish.

Assume for the moment gravitational fields that

- are *varying slowly* in time
- are *weak*, and
- produce *low velocities*, $v << c$.

This implies the conditions

$$\underbrace{\frac{\partial g_{\mu\nu}}{\partial x^0} = 0}_{\text{Slowly varying}} \qquad \underbrace{\frac{dx^i}{d\tau} << 1}_{\text{Weak}} \qquad \underbrace{\frac{dx^0}{d\tau} \simeq 1 \quad (\rightarrow dt \simeq d\tau).}_{v<<c}$$

The *geodesic equation of motion* in this limit reduces to

$$\frac{d^2x^\mu}{d\tau^2} + \Gamma^\mu_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau} = 0 \quad \longrightarrow \quad \frac{d^2x^\mu}{d\tau^2} + \Gamma^\mu_{00}\left(\frac{dx^0}{d\tau}\right)^2 = 0,$$

and the *connection* reduces to

$$\Gamma^\mu_{00} = \tfrac{1}{2}g^{\mu\rho}\left(\underbrace{\frac{\partial g_{0\rho}}{\partial x^0} + \frac{\partial g_{0\rho}}{\partial x^0}}_{\text{Neglect}} - \frac{\partial g_{00}}{\partial x^\rho}\right) = -\tfrac{1}{2}g^{\mu\rho}\frac{\partial g_{00}}{\partial x^\rho}.$$

Since the field is weak, *expand the metric around the flat one*,

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$$

where $h_{\mu\nu}$ is a small correction. Then, $\partial g_{00}/\partial x^\rho = \partial h_{00}/\partial x^\rho$ and to lowest order in $h_{\mu\nu}$

$$\Gamma^\mu_{00} = -\frac{1}{2} g^{\mu\rho} \frac{\partial g_{00}}{\partial x^\rho} \quad \longrightarrow \quad \Gamma^\mu_{00} = -\frac{1}{2} \eta^{\mu\rho} \frac{\partial h_{00}}{\partial x^\rho}.$$

From the metric $\eta_{\mu\nu} = \mathrm{diag}\,(-1,1,1,1)$ the connection components are explicitly

$$\Gamma^0_{00} = \frac{1}{2}\frac{\partial h_{00}}{\partial x^0} = 0 \qquad \Gamma^i_{00} = -\frac{1}{2}\frac{\partial h_{00}}{\partial x^i}.$$

and we thus obtain (restoring c)

$$\frac{d^2x^0}{d\tau^2} = 0 \qquad \frac{d^2x^i}{dt^2} = \frac{1}{2}c^2\frac{\partial h_{00}}{\partial x^i}.$$

Comparing with the Newtonian equation

$$\frac{d^2x^i}{dt^2} = F^i = -\frac{\partial \varphi}{\partial x^i},$$

we conclude that $h_{00} = -2\varphi/c^2$ and thus that

$$g_{00} = \eta_{00} + h_{00} = -\left(1 + \frac{2\varphi}{c^2}\right).$$

This implies a *scalar-field source for weak gravity*

$$\varphi = -\tfrac{1}{2}c^2(g_{00}+1).$$

Thus we obtain in the weak-gravity limit a clear manifestation of the Einstein conjecture that

- gravity derives from the *geometry of space-time*, with
- the *metric tensor* $g_{\mu\nu}$ *as its source*.

At the surface of spherical gravitating object of mass M and radius R, the potential is $\varphi = -GM/R$ and

$$g_{00} \simeq -\left(1+\frac{2\varphi}{c^2}\right) = -\left(1-\frac{2GM}{Rc^2}\right).$$

The second term $2GM/Rc^2$ measures the strength of the gravitational field at the surface of the object.

- It is about 10^{-6} for the Sun.
- It is only of order 10^{-4} even for a white dwarf.
- It is about 0.3 for the surface of a neutron star, which invalidates the assumptions of the weak-gravity derivation.

Neutron star densities imply significant curvature of spacetime and non-negligible general relativistic corrections to Newtonian gravity.

8.2 Recipe for Motion in a Gravitational Field

The preceding discussion suggests a general recipe for writing the equations of motion in a gravitational field.

- Invoke the equivalence principle to justify a local Minkowski coordinate system ξ^μ and formulate the appropriate equations of motion for flat Minkowski spacetime in tensor form.
- Replace the Minkowski coordinates ξ^μ by general curvilinear coordinates x^μ in all equations.
- Replace all derivatives by the corresponding covariant derivatives.
- Replace all integral volume elements by invariant volume elements.

The resulting equations describe physics in a gravitational field.

- Because of the structure of the covariant derivatives, this procedure clearly implies a relationship

$$\text{gravity} \leftrightarrow \text{spacetime curvature} \leftrightarrow \text{mass/energy}.$$

that we will now exploit.

8.3 Towards a Covariant Theory of Gravity

Combining the Poisson equation

$$\nabla^2\varphi = 4\pi G\rho$$

with the density expressed in terms of the time–time component of the stress–energy tensor,

$$T_{00} = \rho c^2 \quad \longrightarrow \quad \rho = \frac{1}{c^2}T_{00},$$

and the weak-gravity scalar field,

$$\varphi = -\tfrac{1}{2}c^2(g_{00}+1),$$

gives

$$\nabla^2 g_{00} = \frac{8\pi G}{c^4}T_{00} \qquad \text{(First stab at covariant gravity)}$$

This expression is clearly not yet satisfactory:

- *Not covariant:* it is expressed in tensor components, not tensors.
- *Not generally valid:* It was derived assuming weak, slowly-varying fields.

But the *limit is correct*, so let's use it as a guide to guessing the form of a fully covariant gravitational theory.

1. The right side of
$$\nabla^2 g_{00} = \frac{8\pi G}{c^4} T_{00}$$
is *not a tensor*, but
   - since the Newtonian limit is *proportional to one component of the stress–energy tensor*,
   - we guess that the right side should be *modified by the replacement* $T_{00} \to T_{\mu\nu}$.
2. The right side now transforms as a *rank-2 tensor*, so covariance requires that the left side be replaced by something having the same transformation properties.
3. We assume for requirements on the new left side:
   - The weak-field limit is proportional to a curvature $\nabla^2 g_{00}$, so the left side should be a *covariant measure of spacetime curvature.*
   - It must be a *rank-2 covariant tensor* to match the right side.
   - It must be *symmetric in its lower indices* to match the corresponding property of $T_{\mu\nu}$ on the right side.
   - It must be *divergenceless with respect to covariant differentiation* since $T_{\mu\nu;\nu} = 0$.
4. The candidate field equations must *reduce to the Poisson equation* describing Newtonian gravitation for weak, slowly-varying fields and non-relativistic velocities.

8.4 The Riemann Curvature Tensor

We must first *generalize Gaussian curvature to 4D spacetime*.

- Let's introduce the rank-4 *Riemann curvature tensor*,

$$R^{\sigma}_{\mu\nu\lambda} = \Gamma^{\sigma}_{\mu\lambda,\nu} - \Gamma^{\sigma}_{\mu\nu,\lambda} + \Gamma^{\sigma}_{\alpha\nu}\Gamma^{\alpha}_{\mu\lambda} - \Gamma^{\sigma}_{\alpha\lambda}\Gamma^{\alpha}_{\mu\nu}.$$

  Lowering an index by contraction with $g_{\mu\nu}$, it has the symmetries

$$R_{\sigma\mu\nu\lambda} = -R_{\mu\sigma\nu\lambda} = -R_{\sigma\mu\lambda\nu}$$

$$R_{\sigma\mu\nu\lambda} = R_{\nu\lambda\sigma\mu} \qquad R_{\sigma\mu\nu\lambda} + R_{\sigma\lambda\mu\nu} + R_{\sigma\nu\lambda\mu} = 0$$

  and also satisfies the *Bianchi Identity*:

$$R^{\sigma}_{\mu\nu\lambda;\rho} + R^{\sigma}_{\mu\rho\nu;\lambda} + R^{\sigma}_{\mu\lambda\rho;\nu} = 0.$$

Because of the symmetries, only 20 of the $4^4 = 256$ components of the Riemann tensor are independent in 4-D spacetime.

> 2-D: 15 symmetry relations on 2^4 components $\rightarrow$ 1 independent component (*Gaussian curvature*).

- All components of $R^{\sigma}_{\mu\nu\lambda}$ vanish in flat spacetime.
- Conversely, if $R^{\sigma}_{\mu\nu\lambda}$ vanishes, spacetime is flat.

> 20 independent components of the Riemann tensor *generalize Gaussian curvature to 4-D spacetime*.

8.5 Instrinsic and Extrinsic Curvature

The curvature tensor has n^4 *components in* n *dimensions*.

- However, symmetries (discussed below) reduce that to $n^2(n^2-1)/12$ *independent components*.
- Thus in 4D the curvature tensor has *20 independent components* and
- in 3D it has *6 independent components*.
- In 2D there are 15 symmetry relations on the $2^4 = 16$ components of the Riemann curvature tensor.
- This leaves only one independent component of curvature—the *Gaussian curvature* already introduced.
- In 1D we find that *curvature cannot be defined*.

The assertion that curvature cannot be defined in 1D may not seem right—*what about a curved line?*

- But curvature is meant here as an *intrinsic property of a manifold* and in 1D there can be no intrinsic curvature.
- What is really meant by a "curved line" is the *embedding of a 1D surface in a higher-dimensional manifold*.
- The perceived curvature of the line then is a *property of the embedding*, not an intrinsic property of the 1D space.

A manifold embedded in a higher-dimensional manifold

- inherits an *induced metric* from the embedding.
- The *apparent curvature* of embedded spaces having no intrinsic curvature *results from this induced metric*.
- This is termed *extrinsic curvature.*

In a similar vein, a cylinder is a *flat 2D surface:*

- Its *Gaussian curvature vanishes* and it is constructed by
- identifying two *opposite edges of a flat surface*.
- This also may be verified by *parallel transporting a vector* in a closed rectangular path on the cylinder.
- Unlike for a sphere, the *vector remains unchanged* under parallel transport on the cylinder.
- The curvature perceived for the cylinder is an *artifact of embedding* the 2D cylinder in a 3D space.

GR usually deals with *intrinsic curvature*.

- Intrinsic curvature is specified by the *Riemann curvature tensor*.
- Intrinsic curvature is *independent of any embedding* in higher dimensions.

Our sole concern here will be *intrinsic curvature*.

8.6 The Einstein Equations

Having now a *covariant description* of

- *matter*,
- *energy*,
- *pressure*, and
- *spacetime curvature*,

we possess the tools to implement a covariant theory of gravity.

- The Riemann curvature tensor is *rank-4*, but
- our previous reasoning indicates that *we need rank-2 tensors* to describe gravity.
- This suggests that we need *contractions of the Riemann tensor with the metric tensor* to form a covariant theory of gravity.

First form the symmetric *Ricci tensor* $R_{\mu\nu}$ by *contracting the Riemann tensor*,

$$\begin{aligned} R_{\mu\nu} &= R_{\nu\mu} = g^{\lambda\sigma} R_{\lambda\mu\sigma\nu} = R^{\sigma}{}_{\mu\sigma\nu}, \\ &= \Gamma^{\lambda}_{\mu\nu,\lambda} - \Gamma^{\lambda}_{\mu\lambda,\nu} + \Gamma^{\lambda}_{\mu\nu}\Gamma^{\sigma}_{\lambda\sigma} - \Gamma^{\sigma}_{\mu\lambda}\Gamma^{\lambda}_{\nu\sigma} \qquad \text{(Ricci tensor)}, \end{aligned}$$

and the *Ricci scalar* R by a further contraction,

$$R = g^{\mu\nu} R_{\mu\nu} \qquad \text{(Ricci scalar)}.$$

Finally multiply the Bianchi identity

$$R^{\sigma}_{\mu\nu\lambda;\rho} + R^{\sigma}_{\mu\rho\nu;\lambda} + R^{\sigma}_{\mu\lambda\rho;\nu} = 0.$$

by $g^{\mu\nu}$ and $g^{\sigma\rho}$ and do the implied sums to give (Problem)

$$\underbrace{\nabla_\lambda R_{\mu\nu\alpha\beta} + \nabla_\beta R_{\mu\nu\lambda\alpha} + \nabla_\alpha R_{\mu\nu\beta\lambda} = 0}_{\text{Bianchi identity}}$$

$$\longrightarrow \left(\underbrace{R^{\mu\nu} - \tfrac{1}{2} g^{\mu\nu} R}_{\text{Einstein tensor}} \right)_{;\nu} = 0$$

where the symmetric *Einstein tensor* is defined by the quantity in parentheses,

$$G^{\mu\nu} \equiv R^{\mu\nu} - \tfrac{1}{2} g^{\mu\nu} R \qquad \text{(Einstein tensor)}.$$

and has *vanishing convariant 4-divergence:*

$$G^{\mu\nu}_{;\nu} = \nabla_\nu G^{\mu\nu} = 0.$$

The *Einstein tensor*

$$G^{\mu\nu} \equiv R^{\mu\nu} - \tfrac{1}{2} g^{\mu\nu} R \qquad \text{(Einstein tensor).}$$

is in fact the *tensor that we seek* to replace the left side of the weak-field equation:

- It is a *rank-2 tensor*.
- It is *symmetric* in its indices.
- It is a *covariant measure of spacetime curvature* because
- it is formed by *contractions of the Riemann tensor*.
- It has *vanishing covariant 4-divergence*.

Thus, we may express the covariant theory of gravitation in terms of the *Einstein equation*,

$$G_{\mu\nu} \equiv \left(R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R\right) = \frac{8\pi G}{c^4} T_{\mu\nu}.$$

or even more compactly in $c = 1$ or $c = G = 1$ units,

$$G_{\mu\nu} = 8\pi G T_{\mu\nu} = 8\pi T_{\mu\nu}.$$

The *tensors are symmetric* so this expression represents *10 coupled, partial, non-linear differential equations* that must be solved to determine the effect of gravitation.

> We shall see later that additional symmetries can reduce this to fewer equations to be solved in some favorable cases.

By contraction with the metric tensor the Einstein equation can also be written in the *alternative form* (Problem)

$$R_{\mu\nu} = \frac{8\pi G}{c^4}(T_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}T^{\alpha}_{\alpha}),$$

where the full contraction T^{α}_{α} represents the *trace of the stress–energy tensor* expressed as a mixed rank-2 tensor.

Vacuum Solutions: If the region where the solution is valid is a *vacuum*,

$$T_{\mu\nu} = T^{\alpha}_{\alpha} = 0.$$

- Then the Einstein equations reduce to the *vacuum Einstein equation*

$$R_{\mu\nu} = 0,$$

- which involves *only the Riemann curvature tensor*, not the full Einstein tensor.

Vacuum solutions of the Einstein equation satisfy the differential equation

$$R_{\mu\nu} = 0.$$

Even though the Ricci tensor $R_{\mu\nu}$ vanishes for this solution, that *does not necessarily mean that spacetime curvature is zero*, as we now discuss.

Only $R_{\mu\nu}$ is needed to construct the vacuum Einstein equation,.

- However, only the full Riemann curvature tensor $R_{\sigma\mu\nu\lambda}$ with its 20 independent components contains *complete information about the spacetime curvature*.
- Because they are *contracted quantities*
  - the Ricci tensor $R_{\mu\nu}$ has only *10 independent components* and
  - the Ricci scalar R only *one independent component*.

When $R_{\sigma\mu\nu\lambda}$ vanishes for the entire space then so do $R_{\mu\nu}$ and R, but the *converse need not hold*. For example,

- A manifold for which $R_{\mu\nu} = 0$ is termed *Ricci flat*.
- But such a manifold *need not be geometrically flat*.
- Only the *vanishing of the full curvature tensor* $R_{\sigma\mu\nu\lambda}$ ensures geometrical flatness.

Example: the *Schwarzschild metric* (see Ch. 9):

- It satisfies $R_{\mu\nu} = 0$ (*Ricci flat*), but corresponds to a *curved spacetime manifold*.
- This is because $R_{\sigma\mu\nu\lambda}$ *has non-vanishing components*, even though $R_{\mu\nu} = 0$.

Indeed, the curvature is so strong that it can lead to a *black hole with an event horizon*.

To fix ideas, lets illustrate calculation of the *quantities that enter the Einstein tensor* in a simple *2D case* with *uniform curvature*.

Example: Consider a *2D spherical surface* in *3D Euclidean space*. Let's find the components of

- the *metric tensor*,
- the non-zero *connection coefficients*,
- the *Riemann curvature tensor*,
- the *Ricci tensor*, and
- the *Ricci scalar curvature*.

In standard polar coordinates the *line element* is

$$ds^2 = a^2(d\theta^2 + \sin^2\theta d\varphi^2),$$

where a is the radius of the sphere. This corresponds to a *diagonal metric* with

$$g_{\theta\theta} = a^2 \qquad g_{\varphi\varphi} = a^2\sin^2\theta \qquad g_{\theta\varphi} = g_{\varphi\theta} = 0,$$

and since $g^{\mu\nu}$ is the matrix inverse of $g_{\mu\nu}$,

$$g^{\theta\theta} = \frac{1}{a^2} \qquad g^{\varphi\varphi} = \frac{1}{a^2\sin^2\theta}.$$

The *connection coefficients* are given by

$$\Gamma^{\sigma}_{\lambda\mu} = \tfrac{1}{2} g^{\nu\sigma}\left(\frac{\partial g_{\mu\nu}}{\partial x^{\lambda}} + \frac{\partial g_{\lambda\nu}}{\partial x^{\mu}} - \frac{\partial g_{\mu\lambda}}{\partial x^{\nu}}\right).$$

Thus from the metric

$$g_{\theta\theta} = a^2 \qquad g_{\varphi\varphi} = a^2 \sin^2\theta \qquad g_{\theta\varphi} = g_{\varphi\theta} = 0,$$

the connection coefficients with θ as an upper index are

$$\Gamma^{\theta}_{\varphi\varphi} = -\sin\theta\cos\theta \qquad \Gamma^{\theta}_{\theta\theta} = \Gamma^{\theta}_{\theta\varphi} = \Gamma^{\theta}_{\varphi\theta} = 0$$

and the connection coefficients with φ as an upper index are

$$\Gamma^{\varphi}_{\theta\varphi} = \Gamma^{\varphi}_{\varphi\theta} = \cot\theta \qquad \Gamma^{\varphi}_{\theta\theta} = \Gamma^{\varphi}_{\varphi\varphi} = 0.$$

This then permits us to calculate the *Riemann tensor and its contractions*.

The *Riemann curvature tensor* is given by

$$R^{\sigma}{}_{\mu\nu\lambda} = \Gamma^{\sigma}_{\mu\lambda,\nu} - \Gamma^{\sigma}_{\mu\nu,\lambda} + \Gamma^{\sigma}_{\alpha\nu}\Gamma^{\alpha}_{\mu\lambda} - \Gamma^{\sigma}_{\alpha\lambda}\Gamma^{\alpha}_{\mu\nu}.$$

This is a 2D space so there will be only *one independent component*, which (up to symmetries) can be taken as

$$\begin{aligned} R^{\theta}{}_{\varphi\theta\varphi} &= \frac{\partial\Gamma^{\theta}_{\varphi\varphi}}{\partial\theta} - \frac{\partial\Gamma^{\theta}_{\varphi\theta}}{\partial\varphi} + \Gamma^{\theta}_{\alpha\theta}\Gamma^{\alpha}_{\varphi\varphi} - \Gamma^{\theta}_{\alpha\varphi}\Gamma^{\alpha}_{\varphi\theta} \\ &= \frac{\partial\Gamma^{\theta}_{\varphi\varphi}}{\partial\theta} - \Gamma^{\theta}_{\varphi\varphi}\Gamma^{\varphi}_{\varphi\theta} = \sin^2\theta. \end{aligned}$$

The metric tensor may be used to *lower the upper index*, $R_{\mu\nu\alpha\beta} = g_{\mu\lambda}R^{\lambda}{}_{\nu\alpha\beta}$, which leads to

$$R_{\theta\varphi\theta\varphi} = R_{\varphi\theta\varphi\theta} = a^2\sin^2\theta,$$

where the symmetry $R_{\sigma\mu\nu\lambda} = R_{\nu\lambda\sigma\mu}$ has been used. The *Ricci tensor* is then given by

$$R_{\mu\nu} = \Gamma^{\lambda}_{\mu\nu,\lambda} - \Gamma^{\lambda}_{\mu\lambda,\nu} + \Gamma^{\lambda}_{\mu\nu}\Gamma^{\sigma}_{\lambda\sigma} - \Gamma^{\sigma}_{\mu\lambda}\Gamma^{\lambda}_{\nu\sigma},$$

and its non-vanishing components are

$$R_{\varphi\varphi} = g^{\theta\theta}R_{\theta\varphi\theta\varphi} = \sin^2\theta \qquad R_{\theta\theta} = g^{\varphi\varphi}R_{\varphi\theta\varphi\theta} = 1.$$

Finally, the *Ricci scalar curvature* is the full contraction of the Ricci tensor with the metric tensor,

$$R = g^{\mu\nu}R_{\mu\nu} = g^{\varphi\varphi}R_{\varphi\varphi} + g^{\theta\theta}R_{\theta\theta} = \frac{2}{a^2},$$

which is, up to a multiplicative factor, the *Gaussian curvature*.

8.6.1 Limiting Cases of the Einstein Tensor

It is not hard to show that the Einstein tensor

$$G_{\mu\nu} \equiv R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R$$

has the following limiting behavior:

- For *weak, non-relativistic fields*,

$$G_{00} \to \nabla^2 g_{00} \qquad \boldsymbol{\nabla} \equiv \hat{i}\frac{\partial}{\partial x} + \hat{j}\frac{\partial}{\partial y} + \hat{k}\frac{\partial}{\partial z},$$

  as required in our derivation of the weak-field limit.

- If *spacetime is flat* (no curvature), $G_{\mu\nu} \to 0$.

- If there were no
  - *matter*
  - *energy*
  - *pressure*

  in the universe, then $G_{\mu\nu} \to 0$.

These are exactly the properties expected from a theory of gravity

- in which *gravity is curved spacetime* and
- *mass–energy–pressure curves spacetime*

that *reduces to Newtonian gravity* for weak fields.

8.7 Solving the Einstein equations

The foregoing development has produced a set of *field equations describing gravity covariantly*.

- However, to apply this formalism systematically it is necessary to find *solutions* for the field equations.
- This is no easy task, given that
  - the Einstein equations represent a set of *coupled, nonlinear, partial differential equations*, and that
  - the appropriate *boundary conditions* may involve tricky issues, particularly in the limit of strong gravity.

In subsequent chapters it will be found that two assumptions can be used to find important solutions analytically that appear to represent physically-observable objects:

- *Assume the field to be weak*, or
- Assume the metric describing the field to have a *high degree of symmetry*.

8.7.1 Solutions in the Limit of Weak Fields

Analytical solutions may be obtained by *positing a relatively weak gravitational field*.

- Then it is justified to *expand the metric* about the Minkowski-space metric.
- Since *most* gravitational fields are weak, this can be very useful.
- Earlier this approach was used to show that Newtonian gravity is the weak-field limit of general relativity.
- Later, it will be used to predict the existence of gravitational waves.

8.7.2 Solutions with a High Degree of Symmetry

A second assumption that can lead to meaningful analytical solutions of the field equations is to *idealize the problem by assuming a high degree of symmetry*.

- Then it is possible to decouple the Einstein equations and reduce the problem to solving a subset of the original equations.
- In this case it may be possible to obtain solutions without making a weak-field assumption.
- This will be the approach that we shall take to finding analytical solutions relevant for black holes and for cosmology in later chapters.

8.7.3 Solutions by Numerical Relativity

If gravity is strong without a high degree of symmetry for the source it is necessary to resort to *numerical relativity,* where solutions are obtained from *large-scale computer simulations*.

- Standard approaches to solution of partial differential equations on a computer often require that the usual textbook representation of general relativity be reformulated.
- Ideally one would like to use a model of an object described by general relativity to supply some "initial data".
- Then these initial data are evolved numerically according to the Einstein field equations to some final equilibrium situation.
- The problem is that the coordinate independence of the usual formulation of GR means that
  - there is no natural notion of "time" and
  - "initial data" has no clearly-defined meaning.
- A typical approach is to reformulate general relativity by splitting spacetime into 3+1 dimensions of space and time.
- In this formulation Einstein's equations take a form better adapted to solving the initial value problem numerically on a computer.

- Full-blown numerical relativity is *beyond the scope of our presentation* because it involves advanced issues in
  - *general relativity*,
  - *numerical analysis*, and
  - *computational science*.
- However, it is *assuming increasing importance* with the emphasis on
  - *black holes* and
  - *gravitational waves*.

  in modern astrophysics.

Although we will not cover numerical relativity in any depth, we will

- give some allusions to it,
- use some of its results, and
- cite some literature references

in the course of our discussion.

Chapter 9

The Schwarzschild Spacetime

One of the *simplest solutions* of the Einstein equations corresponds to

- a metric that describes the gravitational field *exterior to a mass* that is
  - *static*,
  - *spherical*,
  - *uncharged*,
  - *without angular momentum*, and
  - *isolated* from all other mass.
- It was obtained by Schwarzschild in 1916.

> Schwarzschild found the solution while serving in World War I, before Einstein could find any exact solutions for his field equations.

The Schwarzschild solution is

- A solution of the vacuum Einstein equations

$$R_{\mu\nu} = 0.$$

- Only valid in the absence of matter and non-gravitational fields ($T_{\mu\nu} = 0$).
- Spherically symmetric and
- time independent.
- (Later we will see that time independence and spherical symmetry are related to each other for the Schwarzschild solution.)

Thus, the Schwarzschild solution

- is valid *outside spherical mass distributions*, but
- the interior of a star will be described by a different metric.

since it is a *vacuum solution* valid only in the *absence of matter or energy*.

9.1 The Form of the Metric

Let's work in spherical coordinates (r, θ, φ) and seek a time-independent solution assuming that

- The angular part of the metric will be unchanged from its form in flat space because of the spherical symmetry.
- The parts of the metric describing dt and dr will be modified by functions that depend *only on the radial coordinate* r.

Therefore, let us write the 4-D line element as

$$ds^2 = \underbrace{-B(r)\,dt^2 + A(r)\,dr^2}_{\text{Modified from flat space}} \quad \underbrace{+r^2 d\theta^2 + r^2 \sin^2\theta\, d\varphi^2}_{\text{Same as flat space}},$$

where $A(r)$ and $B(r)$ are unknown functions that may depend on r but not time. They may be determined by

1. Inserting this metric in the Einstein field equations for $T_{\mu\nu} = 0$ (*vacuum Einstein equations*).
2. Solving the resulting equations to determine the unknown functions $A(r)$ and $B(r)$.

Substitute the metric form in the vacuum Einstein equation, $R_{\mu\nu} = 0$, and carry out the following steps (Problem):

1. With the assumed form of the metric,

$$g_{\mu\nu} = \begin{pmatrix} -B(r) & 0 & 0 & 0 \\ 0 & A(r) & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & r^2 \sin^2\theta \end{pmatrix}.$$

compute the non-vanishing connection coefficients $\Gamma^{\lambda}_{\mu\nu}$.

$$\Gamma^{\sigma}_{\lambda\mu} = \tfrac{1}{2} g^{\nu\sigma} \left(\frac{\partial g_{\mu\nu}}{\partial x^{\lambda}} + \frac{\partial g_{\lambda\nu}}{\partial x^{\mu}} - \frac{\partial g_{\mu\lambda}}{\partial x^{\nu}} \right)$$

2. Use the $\Gamma^{\lambda}_{\mu\nu}$ to construct the Ricci tensor $R_{\mu\nu}$.

$$R_{\mu\nu} = \Gamma^{\lambda}_{\mu\nu,\lambda} - \Gamma^{\lambda}_{\mu\lambda,\nu} + \Gamma^{\lambda}_{\mu\nu}\Gamma^{\sigma}_{\lambda\sigma} - \Gamma^{\sigma}_{\mu\lambda}\Gamma^{\lambda}_{\nu\sigma},$$

(Only need $R_{\mu\nu}$, not full $G_{\mu\nu}$.)

3. Solve the resulting set of equations for $A(r)$ and $B(r)$.

The final solution is remarkably simple:

$$B(r) = 1 - \frac{2M}{r} \qquad A(r) = B(r)^{-1},$$

($G = c = 1$ units), where M is the single parameter.

The *Schwarzschild line element* is then

$$ds^2 = -\left(1-\frac{2M}{r}\right)dt^2 + \left(1-\frac{2M}{r}\right)^{-1}dr^2 + r^2 d\theta^2 + r^2\sin^2\theta\, d\varphi^2,$$

where $d\tau^2 = -ds^2$. The corresponding *Schwarzschild metric tensor* is

$$g_{\mu\nu} = \begin{pmatrix} -\left(1-\frac{2M}{r}\right) & 0 & 0 & 0 \\ 0 & \left(1-\frac{2M}{r}\right)^{-1} & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & r^2\sin^2\theta \end{pmatrix}.$$

which is *diagonal* but obviously *not constant*.

By comparing

$$\underbrace{g_{00} = -\left(1 - \frac{2GM}{rc^2}\right)}_{\text{Weak gravity (earlier)}} \quad \longleftrightarrow \quad \underbrace{g_{00} = -\left(1 - \frac{2GM}{rc^2}\right)}_{\text{Schwarzschild } (G \text{ \& } c \text{ restored})}$$

- we see that M (the single free parameter of the solution, which arises mathematically as an integration constant)
- may be identified with the *total mass,* consisting of
  - *rest mass,*
  - contributions from *energy densities and pressure,*
  - energy from *spacetime curvature,*

  that is the *source of the gravitational curvature*.

From the structure of the metric

$$ds^2 = -B(r)dt^2 + A(r)dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\varphi^2,$$

- θ and φ have similar interpretations as for flat space.
- The *coordinate radius* r generally cannot be interpreted as a physical radius because $A(r) \neq 1$.
- The *coordinate time* t generally cannot be interpreted as a physical clock time because $B(r) \neq 1$.

The important quantity defined by

$$r_S \equiv 2M = 2GM/c^2$$

is called the *Schwarzschild radius.*

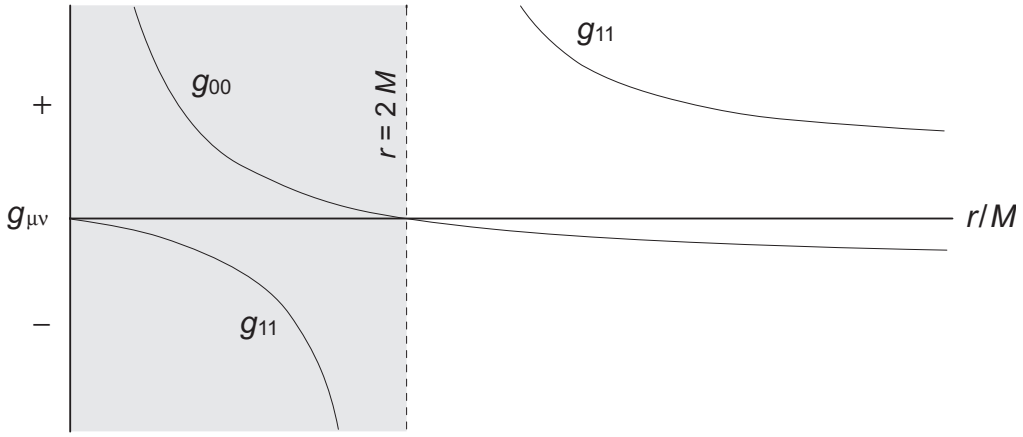


Figure 9.1: The components g_{00} and g_{11} in the Schwarzschild metric.

The line element (metric)

$$ds^2 = \underbrace{-\left(1-\frac{2M}{r}\right)}_{g_{00}} dt^2 + \underbrace{\left(1-\frac{2M}{r}\right)^{-1}}_{g_{11}} dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\varphi^2$$

appears to contain *two singularities*

1. A singularity at $r = 0$ arising from the g_{00} term (an *essential singularity*).
2. A singularity at
$$r = r_S = 2M$$
arising from the g_{11} term (a *coordinate singularity*).

Coordinate Singularity: A place where a chosen set of coordinates *does not describe the geometry properly*.

Example: At the *North Pole*

- the azimuthal angle φ takes a continuum of values 0–2π, so
- all those values correspond to a single point.

But this has *no physical significance.*

Example: Consider the *2D line element*

$$ds^2 = dr^2 + r^2 d\varphi^2$$

and introduce the transformation $r = a^2/\rho$, giving

$$ds^2 = \frac{a^4}{\rho^4}\left(d\rho^2 + \rho^2 d\varphi^2\right).$$

This is *singular at* $\rho = 0$, but

- Nothing is actually pathological at that point,
- so this is a *coordinate problem:*
  - The transformation has *mapped all points at infinity into* $\rho = 0$, and
  - the *geometry is not uniquely represented* by the (ρ, φ) coordinates at $\rho = 0$.

Coordinate singularities are

- *not essential* and
- can be removed by a *different choice of coordinate system.*

Conversely, *physical singularities* cannot be removed by a coordinate transformation.

9.2 Measuring Distance and Time

What is the physical meaning of the coordinates (t, r, θ, φ)?

- We may assign a *practical definition to* r by
  1. Enclosing the origin of our Schwarzschild spacetime in a *series of concentric spheres*,
  2. Measuring for each sphere a *surface area* (conceptually by laying measuring rods end to end),
  3. Assigning a *radial coordinate* r *to that sphere* using
  $$\text{Area} = 4\pi r^2.$$
- Then we can use *distances and trigonometry* to define the *angular coordinate variables* θ and φ.
- Finally we can define *coordinate time* t in terms of *clocks attached to the concentric spheres*.

For Newtonian theory with its implicit assumption that events occur on a passive background of

- euclidean space and
- constantly flowing time,

that's the whole story.

But in curved Schwarzschild spacetime

- The coordinates (t, r, θ, φ) provide a global reference frame for an observer making measurements *at an infinite distance* from the gravitational source.
- However, physical quantities measured by *arbitary observers* are not specified directly by these coordinates but rather

> *Physical quantitites must be computed from the spacetime metric.*

Proper and Coordinate Distances

Consider *distance measured in the radial direction*. Set

$$dt = d\theta = d\varphi = 0$$

in the line element to obtain an *interval of radial distance:*

$$ds^2 = \underbrace{-\left(1-\frac{2M}{r}\right)dt^2 + \left(1-\frac{2M}{r}\right)^{-1}dr^2 + r^2 d\theta^2 + r^2\sin^2\theta d\varphi^2}_{\text{set } t,\theta,\varphi \text{ to constants } \rightarrow dt=d\theta=d\varphi=0}$$

$$\longrightarrow \quad ds = \frac{dr}{\sqrt{1-\dfrac{2GM}{rc^2}}},$$

- In this expression we term
  1. ds the *proper distance* and
  2. dr the *coordinate distance.*

  The *physical radial interval* measured by a local observer is the *proper distance* ds, *not* dr.

- GM/rc^2 measures the strength of gravity, so the *proper distance and coordinate distance are equivalent* only if *gravity is weak*, either because
  1. The source M is small, or
  2. We are large coordinate distance r from the source.

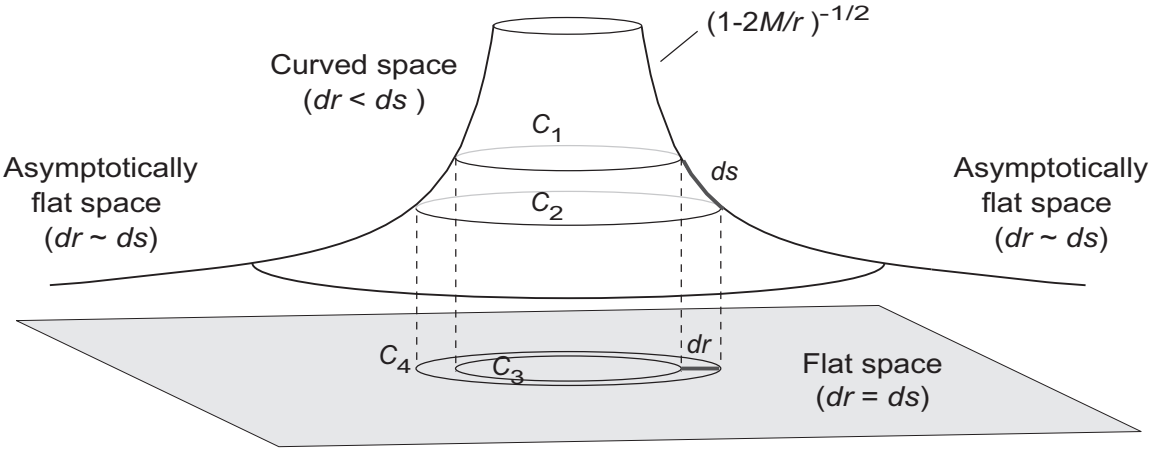


Figure 9.2: Relationship between radial coordinate distance dr and proper distance ds in Schwarzschild spacetime.

The relationship between coordinate distance interval dr and proper distance interval ds is illustrated further in Fig. 9.2.

- Circles C_1 and C_3 represent euclidean spheres of radius r.
- The circles C_2 and C_4 represent spheres having an infinitesimally larger radius $r+dr$ in euclidean space.
- In *euclidean space* the distance between the spheres is dr.
- But in *the curved space* the measured distance between the spheres is ds, which is larger than dr, by virtue of

$$ds = \frac{dr}{\sqrt{1-\dfrac{2GM}{rc^2}}},$$

- Notice however that at large distances from the source of the gravitational field the Schwarzschild spacetime becomes flat and then $dr \sim ds$.

Proper and Coordinate Times

Likewise, to measure a time interval for a stationary clock at r

- set $dr = d\theta = d\varphi = 0$ in the line element and
- use $ds^2 = -d\tau^2 c^2$

to obtain

$$ds^2 = \underbrace{-\left(1-\frac{2M}{r}\right)dt^2 + \left(1-\frac{2M}{r}\right)^{-1}dr^2 + r^2 d\theta^2 + r^2\sin^2\theta\, d\varphi^2}_{\text{set } r,\theta,\varphi \text{ to constants } \rightarrow\, dr=d\theta=d\varphi=0}$$

$$\longrightarrow \quad d\tau = \sqrt{1-\frac{2GM}{rc^2}}\,dt.$$

- In this expression
  - $d\tau$ is termed the *proper time* and
  - dt is termed the *coordinate time*.

> The *physical time interval* measured by a local observer is given by the proper time $d\tau$, *not* by the coordinate time dt.

- dt and $d\tau$ coincide *only if the gravitational field is weak*.

Thus we see that for the gravitational field outside a spherical mass distribution

- The coordinates r and t correspond directly to physical distance and time in Newtonian gravity.
- In general relativity
  - The physical (proper) distances and times *must be computed from the metric.*
  - They generaly are *not given directly by the coordinates*.
- Only in regions of spacetime where gravity is very weak do we recover the Newtonian interpretation.

This is as it should be: *The goal of relativity is to make the laws of physics independent of the coordinate system in which they are formulated.*

The coordinates in a physical theory are *like street numbers*.

- They provide a *labeling* that locates points in a space, but knowing the street numbers is not sufficient to determine distances.
- We can't answer the question of whether the distance between 36th Street and 37th street is the same as the distance between 40th Street and 41st Street until we know how the streets are spaced.
- We must compute distances from a *metric* that gives a *distance-measuring prescription*.
  - Streets that are always equally spaced correspond to a "flat" space.
  - Streets with irregular spacing correspond to a position-dependent metric and thus to a "curved" space.

For the flat space the difference in street number corresponds directly (up to a scale) to a physical distance, but in the more general (curved) case it does not.

9.2.1 Embedding Diagrams

It is sometimes useful to form a mental image of the structure for a curved space by embedding the space or a subset of its dimensions in 3-D euclidean space.

Such mental images are called *embedding diagrams*.

We can embed only 2 dimensions of Schwarzschild spacetime in 3D euclidean space.

- Choose $\theta = \frac{\pi}{2}$ and $t = 0$, to give a 2-D metric

$$d\ell^2 = \left(1 - \frac{2M}{r}\right)^{-1} dr^2 + r^2 d\varphi^2.$$

- The metric of the 3-D embedding space is conveniently represented in cylindrical coordinates as

$$d\ell^2 = dz^2 + dr^2 + r^2 d\varphi^2$$

- This can be written on $z = z(r)$ as

$$d\ell^2 = \left(\frac{dz}{dr}\right)^2 dr^2 + dr^2 + r^2 d\varphi^2 = \left[1 + \left(\frac{dz}{dr}\right)^2\right] dr^2 + r^2 d\varphi^2$$

- Comparing

$$d\ell^2 = \left[1 + \left(\frac{dz}{dr}\right)^2\right] dr^2 + r^2 d\varphi^2$$

with

$$d\ell^2 = \left(1 - \frac{2M}{r}\right)^{-1} dr^2 + r^2 d\varphi^2$$

and solving for $z(r)$ implies that

$$z(r) = 2\sqrt{2M(r - 2M)},$$

which is an *embedding surface* $z(r)$ with the same geometry as the Schwarzschild metric in the $(r - \varphi)$ plane.

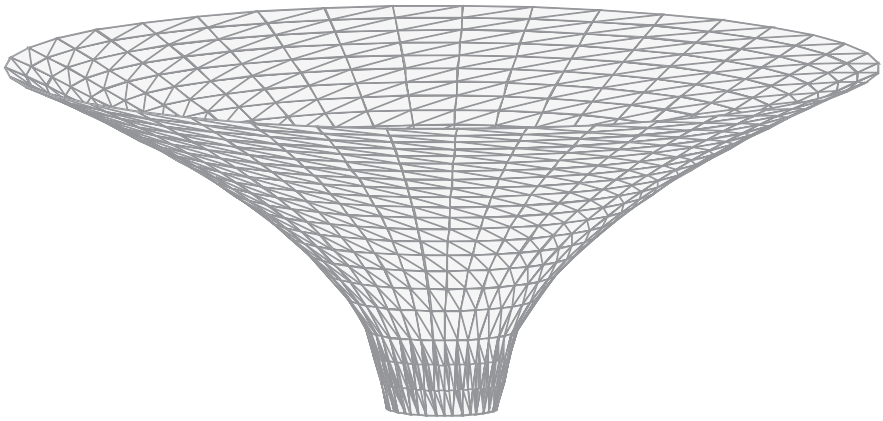


Figure 9.3: An embedding diagram for the Schwarzschild $(r-\varphi)$ plane.

Figure 9.3 shows a plot of the *embedding function* for the Schwarzschild metric

$$z(r) = 2\sqrt{2M(r-2M)}.$$

- Figure 9.3 is not what a Schwarzschild spacetime "looks like" physically, but
- it is a useful visualization of the Schwarzschild geometry.

Thus such embedding diagrams are a standard representation of the Schwarzschild metric in popular-level discussion.

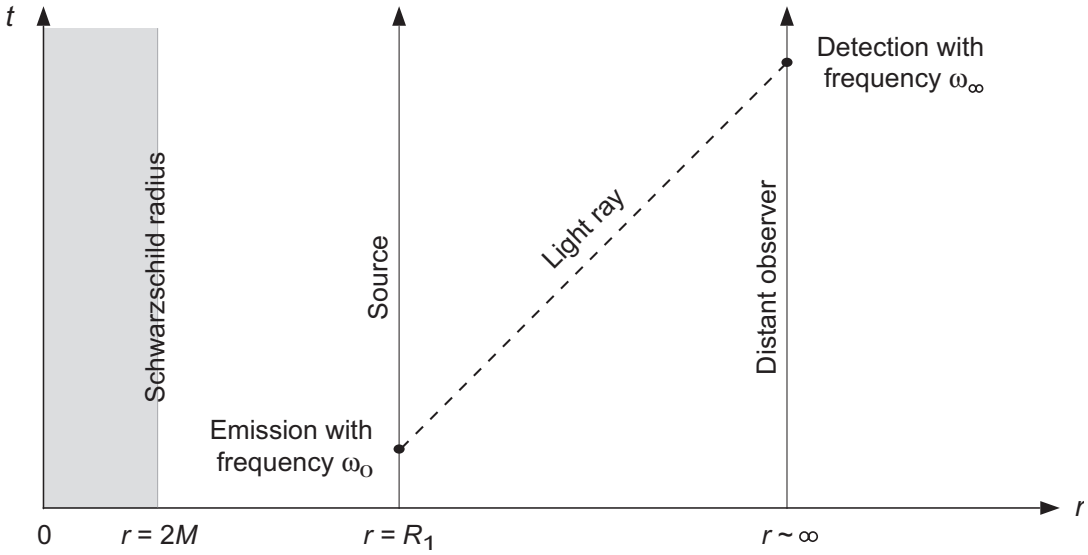


Figure 9.4: Gravitational redshift in the Schwarzschild metric.

9.2.2 The Gravitational Redshift

Let's now return to the *gravitational redshift*, which we treated earlier with a weak-field approximation.

Consider emission of light from R_1 that is detected by a stationary observer at $r >> R_1$ (Fig. 9.4).

For an observer with 4-velocity u, the energy measured for a photon with 4-momentum p is

$$E = \hbar\omega = -p\cdot u,$$

Observers are stationary in space but not time so

$$u^i(r) = 0 \qquad u^0 \neq 0$$

Thus the 4-velocity normalization gives

$$u\cdot u = g_{\mu\nu}(x)\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau} = \underbrace{g_{00}(x)u^0(r)u^0(r) = -1}_{\text{Solve for } u^0(r)}$$

and we obtain

$$u^0(r) = \sqrt{\frac{-1}{g_{00}}} = \left(1-\frac{2M}{r}\right)^{-1/2}.$$

Symmetry: The Schwarzschild metric is *independent of time,* implying a *Killing vector*

$$K^\mu = (t,r,\theta,\varphi) = (1,0,0,0)$$

associated with *time-displacement isometry*.

Thus, for a stationary observer at a distance r,

$$u^\mu(r) = \left(\left(1-\frac{2M}{r}\right)^{-1/2}, 0, 0, 0\right) = \left(1-\frac{2M}{r}\right)^{-1/2} K^\mu,$$

and the energy of the photon measured at r by the observer is

$$\hbar\omega(r) = -p\cdot u = -\left(1-\frac{2M}{r}\right)^{-1/2}(K\cdot p)_r.$$

But K is a *Killing vector*, so

- $K \cdot p$ is *conserved along the geodesic* and thus
- $K \cdot p$ is *independent of* r.

Therefore, it follows that

$$\hbar\omega_0 \equiv \hbar\omega(r = R_1) = -\left(1 - \frac{2M}{R_1}\right)^{-1/2} (K \cdot p)$$

$$\hbar\omega_\infty \equiv \hbar\omega(r \to \infty) = -(K \cdot p).$$

Taking the ratio $\hbar\omega_\infty / \hbar\omega_0$ gives a *gravitational redshift*

$$\omega_\infty = \omega_0 \left(1 - \frac{2M}{R_1}\right)^{1/2}.$$

> *No weak-field assumption was made* so this result should be *generally valid* for situations corresponding to the Schwarzschild metric.

For the special case of *weak fields,*

- $2M/R_1$ is small,
- the square root can be expanded, and
- the G and c factors restored to give

$$\omega_\infty \simeq \omega_0 \left(1 - \frac{GM}{R_1 c^2}\right) \qquad \text{(valid for weak fields).}$$

This is the result found earlier *from the equivalence principle*.

By viewing ω as defining clock ticks, the redshift may also be interpreted as a *gravitational time dilation:* clocks run slower in stronger gravitational fields.

9.2.3 Particle Orbits in the Schwarzschild Metric

Symmetries of the Schwarzschild metric:

1. *Time independence* $\rightarrow$ Killing vector $K_t = (1,0,0,0)$
2. *No dependence on* $\varphi \rightarrow$ Killing vector $K_\varphi = (0,0,0,1)$
3. *Additional Killing vectors* associated with full rotational symmetry (won't need in following).

Conserved quantities associated with these Killing vectors:

$$\varepsilon \equiv -K_t \cdot u = -g_{\mu\nu} K_t^\mu u^\nu = g_{00} u^0 = \left(1 - \frac{2M}{r}\right) \frac{dt}{d\tau}$$

$$\ell \equiv K_\varphi \cdot u = g_{\mu\nu} K_\varphi^\mu u^\nu = g_{33} u^3 = r^2 \sin^2\theta \frac{d\varphi}{d\tau}.$$

Physical interpretation:

- At *low velocities* $\ell \sim$ (orbital angular momentum / unit rest mass)
- Since $E = p^0 = mu^0 = mdt/d\tau$,

$$\lim_{r\to\infty} \varepsilon = \lim_{r\to\infty} \left(1 - \frac{2M}{r}\right) \frac{dt}{d\tau} = \frac{dt}{d\tau} = u^0 = \frac{E}{m}$$

  and $\varepsilon \sim$ energy / (unit rest mass) *at large distance*.

Also we have the usual *velocity normalization constraint* for timelike particles,

$$u \cdot u = g_{\mu\nu} u^\mu u^\nu = -1.$$

Conservation of angular momentum confines the particle motion to a plane, which we conveniently take to be

- the equatorial plane with $\theta = \frac{\pi}{2}$,
- implying that $u^2 \equiv u^\theta = 0$ and $d\theta = 0$.

Then writing the velocity constraint for timelike particles,

$$g_{\mu\nu} u^\mu u^\nu = -1.$$

out explicitly using the metric implied by the Schwarzschild line element,

$$ds^2 = -\left(1 - \frac{2M}{r}\right) dt^2 + \left(1 - \frac{2M}{r}\right)^{-1} dr^2 + \underbrace{r^2 d\theta^2}_{=0} + r^2 \underbrace{\sin^2\theta}_{=1} d\varphi^2,$$

gives that

$$-\left(1 - \frac{2M}{r}\right)(u^0)^2 + \left(1 - \frac{2M}{r}\right)^{-1} (u^1)^2 + r^2 (u^3)^2 = -1,$$

which we may rewrite using

$$u^\mu = \left(\frac{dx^0}{d\tau}, \frac{dx^1}{d\tau}, \frac{dx^2}{d\tau}, \frac{dx^3}{d\tau}\right),$$

$$\varepsilon = \left(1 - \frac{2M}{r}\right)\frac{dt}{d\tau} \qquad \ell = r^2 \sin^2\theta \frac{d\varphi}{d\tau},$$

in the form

$$\frac{\varepsilon^2 - 1}{2} = \frac{1}{2}\left(\frac{dr}{d\tau}\right)^2 + \frac{1}{2}\left[\left(1 - \frac{2M}{r}\right)\left(\frac{\ell^2}{r^2} + 1\right) - 1\right].$$

We can put this in the form

$$E = \frac{1}{2}\left(\frac{dr}{d\tau}\right)^2 + V_{\text{eff}}(r),$$

where we define a fictitious "energy"

$$E \equiv \frac{\varepsilon^2 - 1}{2}$$

and an effective potential

$$\begin{aligned} V_{\text{eff}}(r) &= \frac{1}{2}\left[\left(1 - \frac{2M}{r}\right)\left(\frac{\ell^2}{r^2} + 1\right) - 1\right] \\ &= \underbrace{-\frac{M}{r} + \frac{\ell^2}{2r^2}}_{\text{Newtonian}} - \underbrace{\frac{M\ell^2}{r^3}}_{\text{correction}} . \end{aligned}$$

This is analogous to the energy integral of Newtonian mechanics with

- an effective potential V_{eff} and
- a proper time interval $d\tau$.

The effective potential $V_{\text{eff}}(r)$ is of Newtonian form except that

- The GR potential has an *additional term proportional to* r^{-3} that is not present in the Newtonian potential.
- The Schwarzschild coordinate r and the Newtonian coordinate r *don't have the same physical interpretations*.

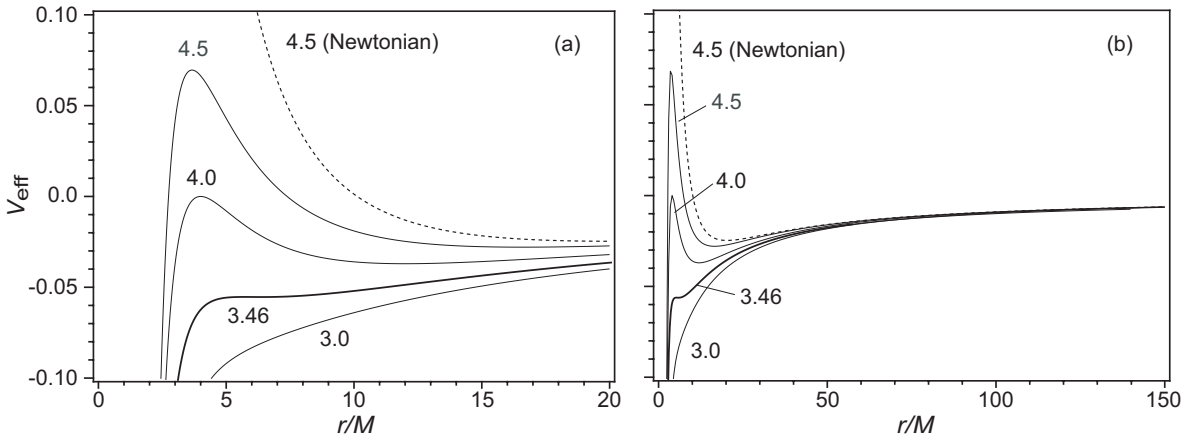


Figure 9.5: (a) Effective potentials for timelike particles in the Schwarzschild metric, with curves labeled by values of ℓ/M. The Schwarzschild curves with $\ell/M > \sqrt{12} \simeq 3.46$ have one maximum and one minimum. The innermost stable circular orbit corresponding to $\ell/M = \sqrt{12} \simeq 3.46$ is indicated by a heavier curve. For $\ell/M = 4.5$ both the Schwarzschild potential (solid) and the corresponding Newtonian potential (dotted) are displayed. (b) Behavior of the solutions in (a) at large distances.

Figure 9.5 compares the *Schwarzschild effective potential* with an *effective Newtonian potential*.

- The Schwarzschild potential generally has *one maximum and one minimum* if $\ell/M > \sqrt{12}$.
- Note the very different behavior of Schwarzschild and Newtonian mechanics at the origin because of the *correction term* in

$$V_{\text{eff}}(r) = \underbrace{-\frac{M}{r} + \frac{\ell^2}{2r^2}}_{\text{Newtonian}} - \underbrace{\frac{M\ell^2}{r^3}}_{\text{correction}}$$

Figure 9.6 on the following page summarizes *possible orbits in the Schwarzschild spacetime*.

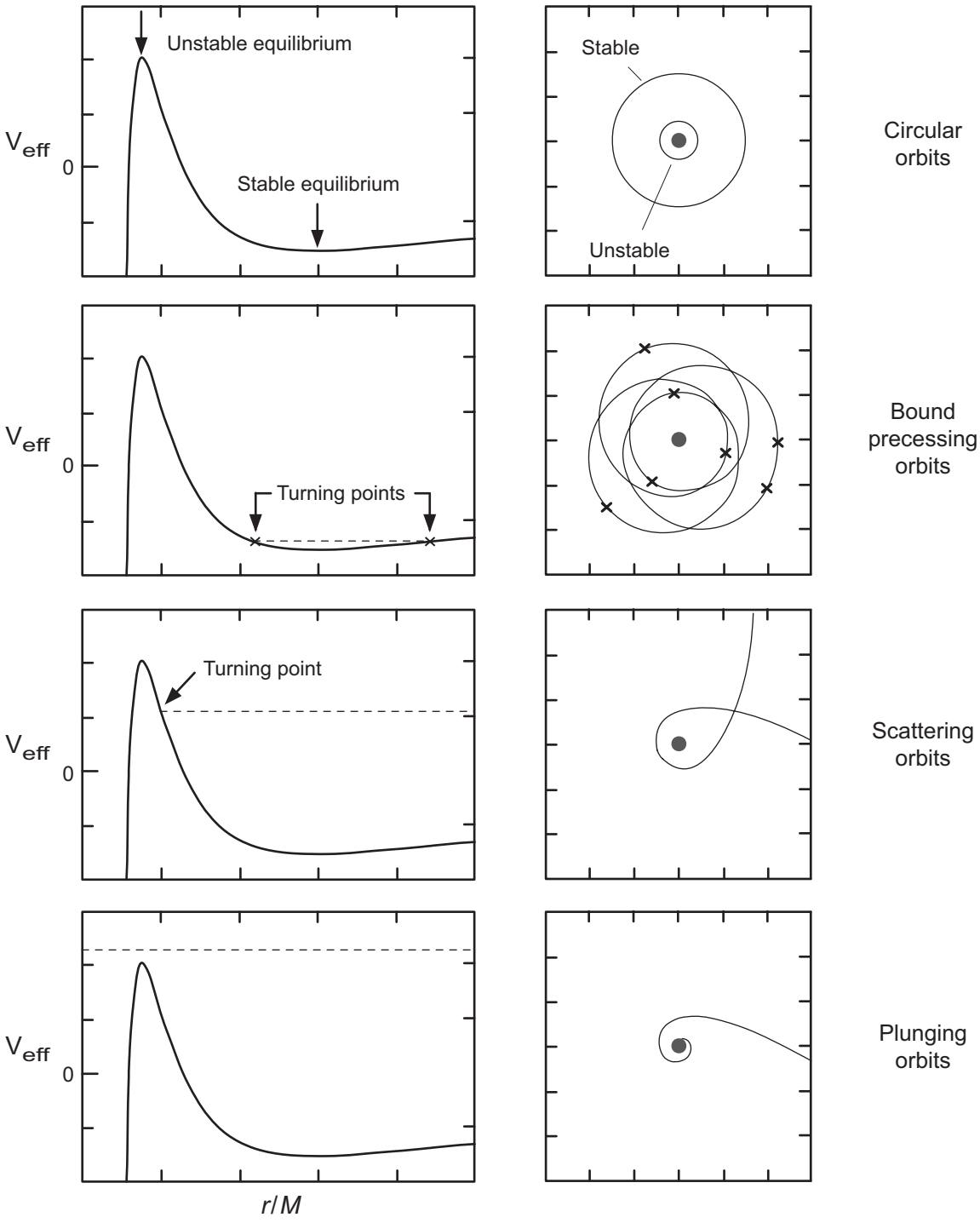


Figure 9.6: Orbits in a Schwarzschild spacetime. Effective potential on left and corresponding classes of orbits on right.

9.2.4 Stable Circular Orbits

Accretion onto compact objects is a major energy source for various astrophysical phenomena.

- Accretion typically occurs through an *accretion disk*, and
- *tidal forces* on the particles in an accretion disk tend to *circularize orbits*.

Therefore, the *stable and marginally-stable circular orbits* for a spacetime are of particular interest in astrophysics.

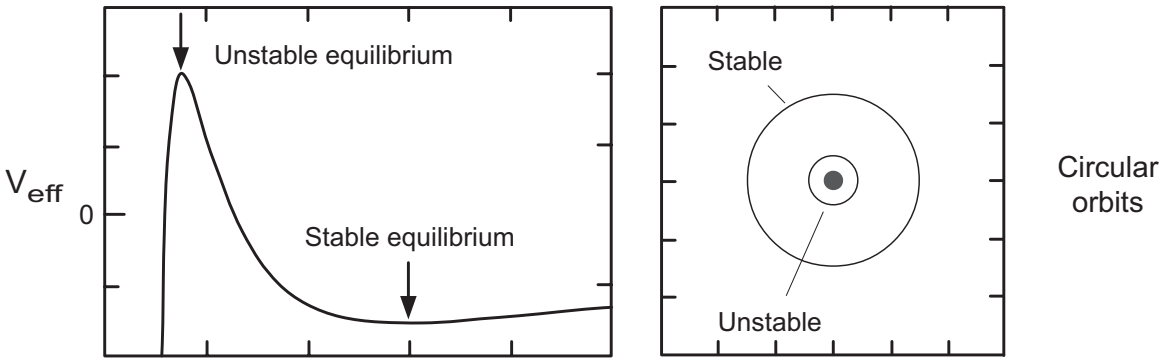


Figure 9.7: Circular orbits in a Schwarzschild spacetime. Effective potential on left and corresponding classes of orbits on right.

A *circular orbit* occurs in the Schwarzschild metric when E is equal to a minimum or maximum of the effective potential, as illustrated in arrows in Fig. 9.7.

- The radial coordinates of these orbits satisfy

$$r = \frac{\ell^2}{2M} \pm \frac{1}{2}\sqrt{\frac{\ell^4}{M^2} - 12\,\ell^2}.$$

- This has two solutions if if $\ell^2/M^2 > 12$, with
  - the plus sign corresponding to a minimum and
  - the minus sign to a maximum of the potential.

- The requirement that $E = V_{\rm eff}$ at a minimum implies that

$$\varepsilon^2 = \left(1 - \frac{2M}{r}\right)\left(1 + \frac{\ell^2}{r^2}\right).$$

The angular velocity Ω of a particle as seen by a distant observer in a $\theta = \frac{\pi}{2}$ Schwarzschild orbit is given by

$$\Omega = \frac{d\varphi}{dt} = \frac{1}{r^2}\left(1 - \frac{2M}{r}\right)\left(\frac{\ell}{\varepsilon}\right),$$

which implies that

$$\frac{\ell}{\varepsilon} = \sqrt{Mr}\left(1 - \frac{2M}{r}\right)^{-1},$$

which further implies that

$$\Omega = \sqrt{\frac{M}{r^3}}.$$

Since the period is given by $P = 2\pi/\Omega$, this is

- *equivalent formally to Kepler's 3rd law*, but
- it is *expressed in terms of Schwarzschild coordinates* r and t rather than in terms of proper distances and times.

It will be useful for later applications to write out explicit *components of the velocity 4-vector for a circular Schwarzschild orbit* in the equatorial plane.

- The components are

$$u = (u^t, 0, 0, \Omega u^t),$$

- where the relation

$$\Omega = \frac{d\varphi}{d\tau}\frac{d\tau}{dt} = \frac{1}{u^t}\frac{d\varphi}{d\tau}$$

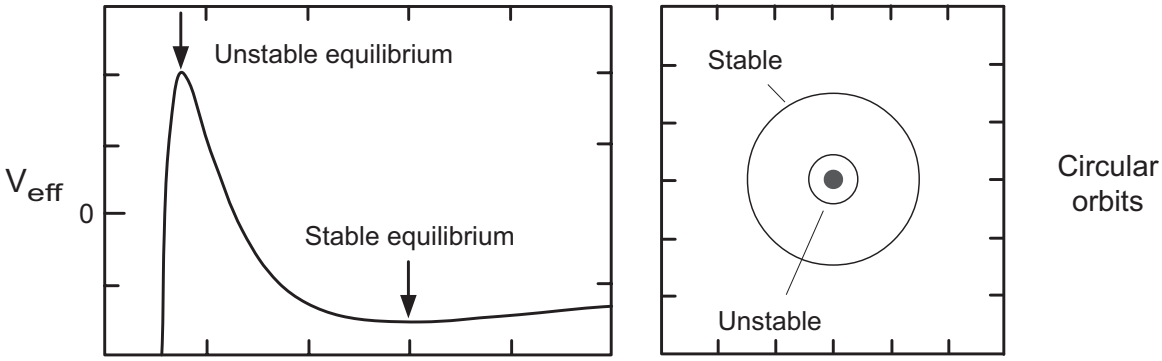
was used.

The timelike component $u^t = dt/d\tau$ may be evaluated explicitly using the 4-velocity normalization

$$g_{\mu\nu}u^\mu u^\nu = -1,$$

which gives

$$u^t = \left(1 - \frac{3M}{r}\right)^{-1/2}.$$



Stable circular orbits

- *do not exist at arbitrarily small radial coordinates* in the Schwarzschild spacetime, as illustrated in the figure above.
- The minimum radius for a stable orbit occurs for $\ell^2/M^2 = 12$.
- The corresponding radius for the *innermost stable circular orbit (ISCO)* in the Schwarzschild spacetime is then

$$R_{\rm ISCO} = 6M.$$

The *innermost stable circular orbit* is important in determining how much gravitational energy can be extracted from *accretion onto compact objects*.

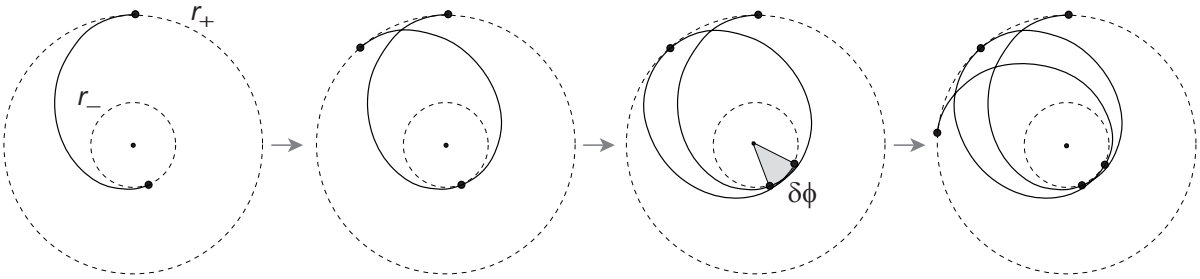


Figure 9.8: Precession of orbits in a Schwarzschild metric (highly-exaggerated). The radial coordinate of the inner turning point r_- and of the outer turning point r_+ are represented by dashed circles. Dots on the inner circle indicate perihelion for each orbit and dots on the outer circle indicate the corresponding aphelion. The quantity $\delta\varphi$ indicates the shift in angle of the perihelion for one orbital period.

9.3 Precession of Orbits

An orbit *closes* (on itself) if the angle φ

- *sweeps out exactly* 2π in the passage between two successive inner or two successive outer radial turning points.
- In Newtonian gravity the central potential is $1/r$, implying *closed elliptical orbits* (leading to *Kepler's laws*).
- In the Schwarzschild metric the effective potential deviates from $1/r$ and *orbits precess:*

  The angle φ changes by *more than* 2π between successive radial turning points.

Precession is illustrated in Fig. 9.8.

To investigate this precession quantitatively we require an *expression for* $d\varphi/dr$. From the *energy equation*

$$E = \frac{1}{2}\left(\frac{dr}{d\tau}\right)^2 + V_{\text{eff}}(r) \quad \longrightarrow \quad \frac{dr}{d\tau} = \pm\sqrt{2(E - V_{\text{eff}}(r))},$$

and from the *conservation equation for* ℓ,

$$\ell = r^2 \sin^2\theta \frac{d\varphi}{d\tau} \quad \longrightarrow \quad \frac{d\varphi}{d\tau} = \frac{\ell}{r^2 \sin^2\theta}.$$

Combining, recalling that we chose an orbital plane $\theta = \frac{\pi}{2}$,

$$\begin{aligned}\frac{d\varphi}{dr} = \frac{d\varphi/d\tau}{dr/d\tau} &= \pm\frac{\ell}{r^2\sqrt{2(E - V_{\text{eff}}(r))}} \\ &= \pm\frac{\ell}{r^2}\left[2E - \left(1 - \frac{2M}{r}\right)\left(1 + \frac{\ell^2}{r^2}\right) + 1\right]^{-1/2} \\ &= \pm\frac{\ell}{r^2}\left[\varepsilon^2 - \left(1 - \frac{2M}{r}\right)\left(1 + \frac{\ell^2}{r^2}\right)\right]^{-1/2},\end{aligned}$$

where we have used $E = \frac{1}{2}(\varepsilon^2 - 1)$.

The change in φ per orbit, $\Delta\varphi$, can be obtained by integrating over one orbit,

$$\begin{aligned}\Delta\varphi &= \int_{r_-}^{r_+}\frac{d\varphi}{dr}dr + \int_{r_+}^{r_-}\frac{d\varphi}{dr}dr = 2\int_{r_-}^{r_+}\frac{d\varphi}{dr}dr \\ &= 2\ell\int_{r_-}^{r_+}\frac{dr}{r^2}\left[\varepsilon^2 - \left(1-\frac{2M}{r}\right)\left(1+\frac{\ell^2}{r^2}\right)\right]^{-1/2} \\ &= 2\ell\int_{r_-}^{r_+}\frac{dr}{r^2}\left(\underbrace{c^2(\varepsilon^2-1)+\frac{2GM}{r}-\frac{\ell^2}{r^2}}_{\text{Newtonian}}+\underbrace{\frac{2GM\ell^2}{c^2r^3}}_{\text{correction}}\right)^{-1/2}\end{aligned}$$

where in the last step G and c have been reinserted through

$$M \to \frac{GM}{c^2} \qquad \ell \to \frac{\ell}{c},$$

Evaluation of the integral requires some care because *the integrand tends to* ∞ *at the integration limits:* From an earlier expressions

$$\frac{dr}{d\tau} = \pm\left[\varepsilon^2 - \left(1-\frac{2M}{r}\right)\left(1+\frac{l^2}{r^2}\right)\right]^{1/2},$$

which is the denominator of our integrand. But the limits are turning points of the radial motion and $dr/d\tau = 0$ at r_+ or r_-.

In the Solar System and most other applications the values of $\Delta\varphi$ are very small.

- Thus it is sufficient to keep only terms of order $1/c^2$ beyond Newtonian approximation.
- Expanding the integrand and evaluating the integral with due care yields

$$\begin{aligned}\text{Precession angle} &= \delta\varphi \equiv \Delta\varphi - 2\pi \\ &\simeq 6\pi \left(\frac{GM}{c\ell}\right)^2 \text{ rad/orbit.}\end{aligned}$$

This may be expressed in more familiar orbital parameters:

- In Newtonian mechanics $L = mr^2\omega$, where L is the angular momentum and ω the angular frequency.
- For Kepler orbits

$$\ell^2 = \left(\frac{L}{m}\right)^2 = \left(r^2\frac{d\varphi}{d\tau}\right)^2 \simeq \left(r^2\frac{d\varphi}{dt}\right)^2 = GMa(1-e^2),$$

where e is the eccentricity and a is the semimajor axis.

This permits us to write

$$\begin{aligned}\delta\varphi &= \frac{6\pi GM}{ac^2(1-e^2)} \\ &= 1.861\times 10^{-7}\left(\frac{M}{M_\odot}\right)\left(\frac{\text{AU}}{a}\right)\frac{1}{1-e^2} \text{ rad/orbit,}\end{aligned}$$

The form of

$$\delta\varphi = \frac{6\pi GM}{ac^2(1-e^2)}$$

shows explicitly that *relativistic precession is enhanced by*

- large M for the central mass,
- tight orbits (small values of a),
- large eccentricities e.

The precession observed for most objects is small.

- *Precession of Mercury's orbit* in the Sun's gravitational field because of general relativistic effects is *43 arcseconds per century.*
- The *orbit of the Binary Pulsar* precesses by about *4.2 degrees per year.*

The *precise agreement of both of these observations* with the predictions of general relativity is a *strong test of the theory*.

9.3.1 Escape Velocity in the Schwarzschild Metric

Consider a stationary observer at a Schwarzschild radial coordinate R who launches a projectile radially with a velocity v such that the projectile reaches infinity with zero velocity.

- This defines the *escape velocity* in the Schwarzschild metric.
- The projectile follows a *radial geodesic* since there are no forces acting on it
- The energy per unit rest mass is ε and it is conserved (time invariance of metric).
- At infinity $\varepsilon = 1$, since then the particle is at rest and the entire energy is rest mass energy.
- Thus $\varepsilon = 1$ at all times since it is conserved.

If u_{obs} is the 4-velocity of the stationary observer, the energy measured by the observer is

$$\begin{aligned} E &= -p \cdot u_{\text{obs}} = -mu \cdot u_{\text{obs}} \\ &= -m g_{\mu\nu} u^{\mu} u^{\nu}_{\text{obs}} \\ &= -m g_{00} u^{0} u^{0}_{\text{obs}}, \end{aligned}$$

where $p = mu$, with p the 4-momentum and m the rest mass, and the last step follows because the observer is stationary (in space but not in time).

But we have that

$$\underbrace{g_{00} = -\left(1 - \tfrac{2M}{r}\right)}_{\text{From metric}} \quad \underbrace{u^0_{\text{obs}} = \left(1 - \tfrac{2M}{R}\right)^{-1/2}}_{\text{Stationary observer}} \quad \underbrace{u^0 = \left(1 - \tfrac{2M}{r}\right)^{-1}}_{\varepsilon = (1 - \frac{2M}{r})u^0 = 1}$$

Therefore,

$$\begin{aligned} E &= -m g_{00} u^0 u^0_{\text{obs}} \\ &= m\left(1 - \tfrac{2M}{r}\right)\left(1 - \tfrac{2M}{r}\right)^{-1}\left(1 - \tfrac{2M}{r}\right)^{-1/2} \\ &= m\left(1 - \tfrac{2M}{R}\right)^{-1/2}. \end{aligned}$$

But in the observer's rest frame

$$E = m\gamma = m(1 - v^2)^{-1/2},$$

so comparison yields $2M/R = v^2$ and thus

$$v_{\text{esc}} = \sqrt{\frac{2M}{R}} = \sqrt{\frac{2GM}{R}}.$$

Notice that

- This, coincidentally, is the *same result as for Newtonian theory*.
- At the Schwarzschild radius $R = r_{\text{S}} = 2M$, the *escape velocity is equal to* c.

This is the first hint of an *event horizon* in the Schwarzschild spacetime.

9.4 Radial Fall of a Test Particle

It will be instructive for later discussion to consider a *radial plunge orbit* that *starts from infinity* with

- *zero kinetic energy* ($\varepsilon = 1$) and
- *zero angular momentum* ($\ell = 0$).

First, let's find an expression for the *proper time* as a function of the coordinate r.

- From earlier expressions

$$E = \frac{\varepsilon^2 - 1}{2} = \frac{1}{2}\left(\frac{dr}{d\tau}\right)^2 - \frac{M}{r} + \frac{\ell^2}{2r^2} - \frac{M\ell^2}{r^3},$$

- which implies for $\ell = 0$ and $\varepsilon = 1$,

$$\frac{dr}{d\tau} = \pm\left(\frac{2M}{r}\right)^{1/2}.$$

- Choosing the negative sign (infalling orbit) and integrating with initial condition $\tau(r = 0) = 0$ gives (Problem)

$$\frac{\tau}{2M} = -\frac{2}{3}\frac{r^{3/2}}{(2M)^{3/2}}$$

for the *proper time* τ *to reach the origin* as a function of the initial Schwarzschild coordinate r.

That was the *proper time* τ. To find an expression for the *coordinate time* t as a function of r,

- we note that $\varepsilon = 1$ and is conserved. Then from

$$\varepsilon = 1 = \left(1 - \frac{2M}{r}\right)\frac{dt}{d\tau} \qquad \frac{dr}{d\tau} = \pm\left(\frac{2M}{r}\right)^{1/2}$$

  we have that

$$\frac{dt}{dr} = \frac{dt/d\tau}{dr/d\tau} = -\left(1 - \frac{2M}{r}\right)^{-1}\left(\frac{2M}{r}\right)^{-1/2}.$$

- This may be integrated to give (Problem)

$$t = t_0 - 2M\left(-\frac{2}{3}\left(\frac{r}{2M}\right)^{3/2} + 2\left(\frac{r}{2M}\right)^{1/2} + \ln\left|\frac{(r/2M)^{1/2} - 1}{(r/2M)^{1/2} + 1}\right|\right).$$

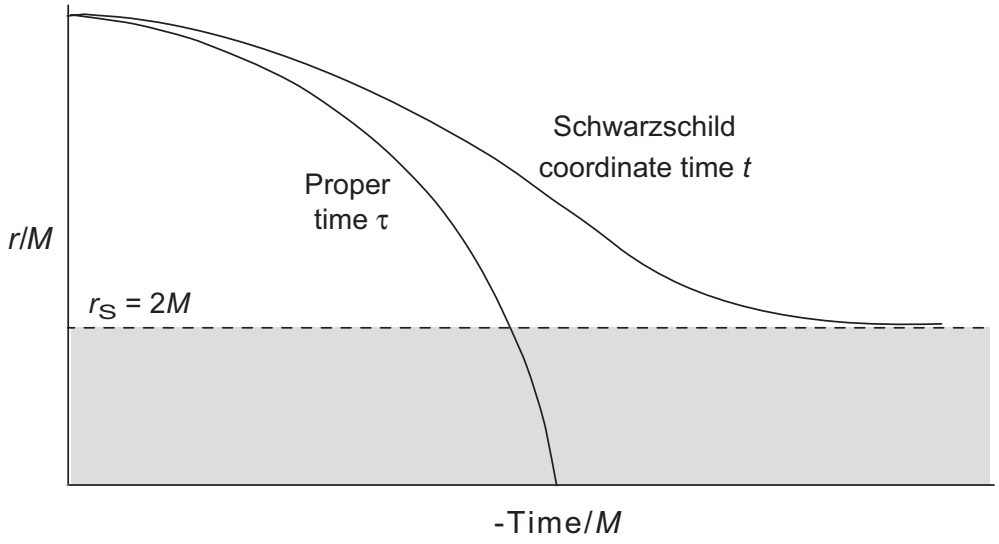


Figure 9.9: Comparison of proper time and Schwarzschild coordinate time for a particle falling radially in the Schwarzschild geometry.

- The proper time τ to fall to the origin is finite.
- For the same trajectory an infinite amount of coordinate time t elapses to reach the Schwarzschild radius.
- The smooth trajectory of the proper time through r_S suggests that the apparent singularity of the metric there is not real.

Later we shall introduce alternative coordinates that explicitly remove the singularity at $r = 2M$ (but not at $r = 0$).

9.5 Orbits for Light Rays

Calculation of light ray orbits in the Schwarzschild metric largely parallels that of particle orbits,

- except that

$$u\cdot u = g_{\mu\nu}\frac{dx^\mu}{d\lambda}\frac{dx^\nu}{d\lambda} = 0,$$

  where λ is an *affine parameter*.

- For motion in the equatorial plane ($\theta = \frac{\pi}{2}$), this becomes

$$-\left(1-\frac{2M}{r}\right)\left(\frac{dt}{d\lambda}\right)^2+\left(1-\frac{2M}{r}\right)^{-1}\left(\frac{dr}{d\lambda}\right)^2+r^2\left(\frac{d\varphi}{d\lambda}\right)^2=0.$$

- By analogy with the arguments for particle motion

$$\varepsilon \equiv -K_t\cdot u = \left(1-\frac{2M}{r}\right)\frac{dt}{d\lambda},$$
$$\ell = K_\varphi\cdot u = r^2\sin^2\theta\frac{d\varphi}{d\lambda},$$

  are *conserved along the orbits of light rays*.

With a proper choice of normalization for λ,

- ε may be interpreted as the *photon energy*
- ℓ is the *photon angular momentum* at infinity.

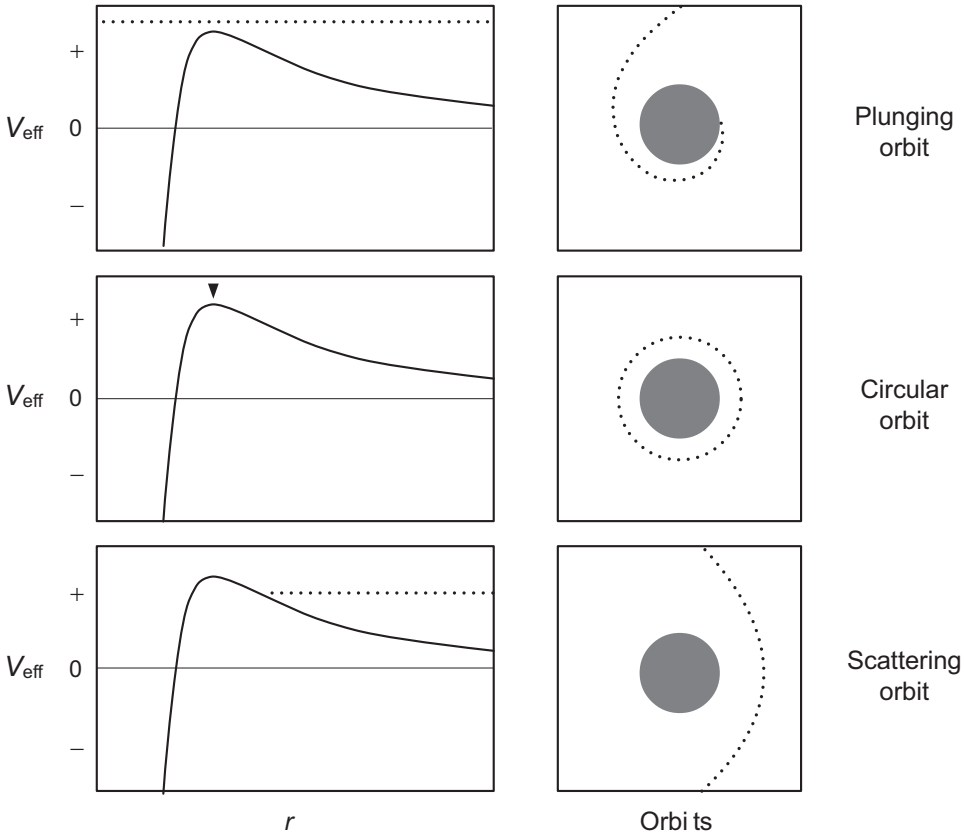


Figure 9.10: Effective potential for photons and light ray orbits in a Schwarzschild metric. Dotted lines on the left side give $1/b^2$ for each orbit.

By following steps analogous to the derivation for particle orbits the equation of motion is

$$\frac{1}{b^2} = \frac{1}{\ell^2}\left(\frac{dr}{d\lambda}\right)^2 + V_{\text{eff}}(r)$$

$$V_{\text{eff}}(r) \equiv \frac{1}{r^2}\left(1 - \frac{2M}{r}\right) \qquad b^2 \equiv \frac{\ell^2}{\varepsilon^2}.$$

The effective potential for photons and some classes of orbits in the Schwarzschild geometry are illustrated in Fig. 9.10.

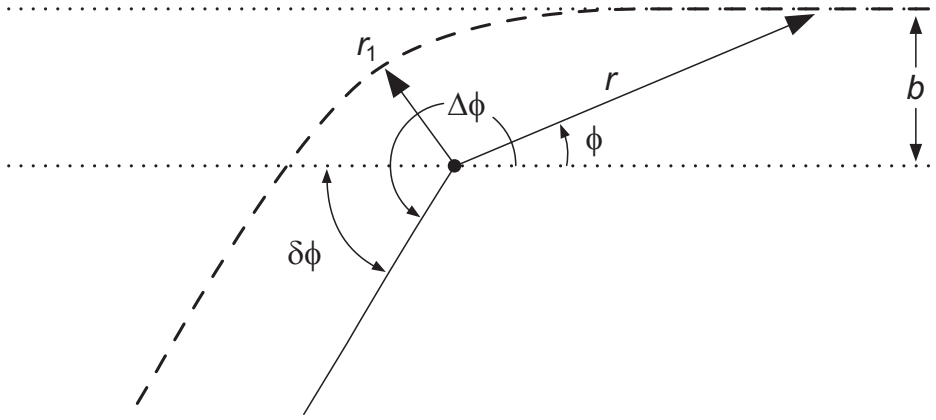


Figure 9.11: Deflection of light by an angle $\delta\varphi$ in a Schwarzschild metric.

9.6 Deflection of Light in a Gravitational Field

Proceeding in a manner similar to that for the calculation of the precession angle for orbits of massive objects, we may calculate the *deflection* $d\varphi/dr$ *for a light ray* in the Schwarzschild metric.

$$\delta\varphi = \frac{4GM}{c^2 b} = 8.477 \times 10^{-6} \left(\frac{M}{M_\odot}\right)\left(\frac{R_\odot}{b}\right) \text{ radians.}$$

For a *photon grazing the Sun's surface*,

- $M = 1\,M_\odot$ and
- $b = 1\,R_\odot$,

which gives $\delta\varphi \simeq 1.75$ arcseconds.

> *Observation of this deflection* during a total solar eclipse catapulted Einstein to worldwide fame almost overnight in the early 1920s.

9.7 Shapiro Time Delay of Light

Light passing near a gravitating body follows a curved path and the time for light to travel between two points depends on this curvature.

- The deviation in travel time between that in the curved spacetime and the travel time if there were no curvature is termed the *Shapiro time delay.*
- This does not mean that the speed of light varies.
- The local speed of light is always c,
- but the observed elapsed time for light to go between two points in spacetime *depends on the metric*.

Thus, measurement of this time delay is a test of general relativity.

To determine the *time delay of light* over a given path it is necessary to evaluate the integral of dt/dr.

- Proceeding in a similar manner as the earlier discussion of light deflection, we may use

$$\varepsilon = \left(1 - \frac{2M}{r}\right)\frac{dt}{d\lambda} \qquad \frac{1}{b^2} = \frac{1}{\ell^2}\left(\frac{dr}{d\lambda}\right)^2 + V_{\text{eff}}(r)$$

  to write

$$\frac{dt}{dr} = \frac{dt/d\lambda}{dr/d\lambda} = \pm\varepsilon\left(1 - \frac{2M}{r}\right)^{-1}\left[\ell^2\left(\frac{1}{b^2} - V_{\text{eff}}\right)\right]^{-1/2}$$
$$= \pm\frac{1}{b}\left(1 - \frac{2M}{r}\right)^{-1}\left(\frac{1}{b^2} - V_{\text{eff}}(r)\right)^{-1/2}.$$

- This may be integrated to give the light travel time.

In a typical *Shapiro-delay experiment*,

- radar waves are *bounced off a planet* and
- the time to go and return is measured for paths that pass *very close to the surface of the Sun*, or
- the *delay in transmitting signals from space probes* to Earth is measured as the signals *pass near the Sun*.

> The results of such experiments are consistent with the equation above, thereby providing further confidence in the validity of general relativity.

9.8 Gyroscopes in Curved Spacetime

Consider the behavior of gyroscopes in free fall.

- These will follow a *timelike geodesic*, with a 4-velocity $u(\tau)$ governed by the geodesic equation

$$\frac{du^\lambda}{d\tau} + \Gamma^\lambda_{\mu\nu} u^\mu u^\nu = 0.$$

- In addition the gyroscope will have a spacelike spin 4-vector $s^\mu = (0, \boldsymbol{s})$ in this frame.
- In the freely-falling local inertial frame the 4-velocity components of the gyroscope are $u = (1,0,0,0)$, so

$$s \cdot u = 0,$$

  which is *a tensor equation and thus true in all frames*.
- In flat spacetime or a local inertial frame $ds^\mu/d\tau = 0$.
- In curved spacetime the appropriate covariant generalization gives an equation *analogous to the geodesic equation*,

$$\frac{ds^\mu}{d\tau} + \Gamma^\mu_{\alpha\beta} s^\alpha u^\beta = 0.$$

  - This equation describes how the components of the gyroscopic spin s^μ change along a geodesic and
  - *preserves the scalar product* $s \cdot u$ on the geodesic.

As in classical mechanics the magnitude of the spin is constant but the direction can *precess*.

Let us now use the equation

$$\frac{ds^\mu}{d\tau} + \Gamma^\mu_{\alpha\beta} s^\alpha u^\beta = 0.$$

to investigate two predictions of general relativity that lead to precession of the spin vector for gyroscopes in gravitational fields,

- *geodetic precession* and
- *dragging of inertial frames.*

9.9 Geodetic Precession

Consider a gyroscope in a circular orbit around a *non-rotating* gravitating sphere of mass M.

- A comoving observer in orbital free fall will see the *gyroscope precess*, even if the source of the field is not rotating.
- This is called *geodetic precession*.

Assume that

- Spacetime is described by the Schwarzschild metric,
- The radius for the orbit is R, and
- The spin points initially in the direction of a distant star.

For an observer at rest in the gyroscope's frame

- The spin has only spatial components.
- By symmetry it must remain in the same plane.
- Choosing $\theta = \frac{\pi}{2}$ for this plane in Schwarzschild coordinates (t, r, θ, φ), any precession occurs in the φ direction.

For the 4-velocity $(u^0, u^1, u^2, u^3) \equiv (u^t, u^r, u^\theta, u^\varphi)$ the component in the φ direction is

$$u^\varphi = \frac{d\varphi}{d\tau} = \frac{d\varphi}{dt}\frac{dt}{d\tau} = \Omega u^t \qquad \Omega \equiv \frac{d\varphi}{dt} = \sqrt{\frac{M}{R^3}} \qquad u^t \equiv \frac{dt}{d\tau},$$

where Ω is the classical orbital angular velocity.

Time evolution for spin components $(s^t, s^r, s^\theta, s^\varphi)$ is given by

$$\frac{ds^\mu}{d\tau} + \Gamma^\mu_{\alpha\beta} s^\alpha u^\beta = 0.$$

evaluated in the Schwarzschild basis.

- Choose $s^\theta = 0$. It will remain zero because of symmetry.
- Because $s \cdot u = 0$ along the geodesic, (Problem)

$$s^t = \frac{R^2\Omega}{1 - 2M/R} s^\varphi.$$

- The remaining spin components s^r and s^φ require solving

$$\frac{ds^r}{d\tau} + \Gamma^r_{\alpha\beta} s^\alpha u^\beta = 0,$$
$$\frac{ds^\varphi}{d\tau} + \Gamma^\varphi_{\alpha\beta} s^\alpha u^\beta = 0.$$

- The solutions are (Problem)

$$s^\varphi(t) = -s_0 \sqrt{1 - \frac{2M}{R}} \left(\frac{\Omega}{\omega R} \right) \sin(\omega t)$$

$$s^r(t) = s_0 \sqrt{1 - \frac{2M}{R}} \cos(\omega t),$$

where $s_0 = (s \cdot s)^{1/2}$ is the invariant spin magnitude and

$$\omega \equiv \left(1 - \frac{3M}{R} \right)^{1/2} \Omega.$$

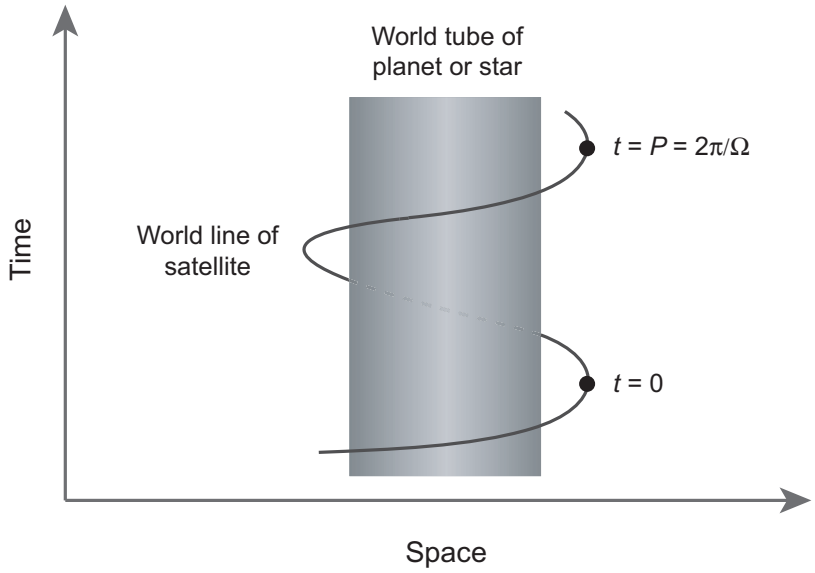
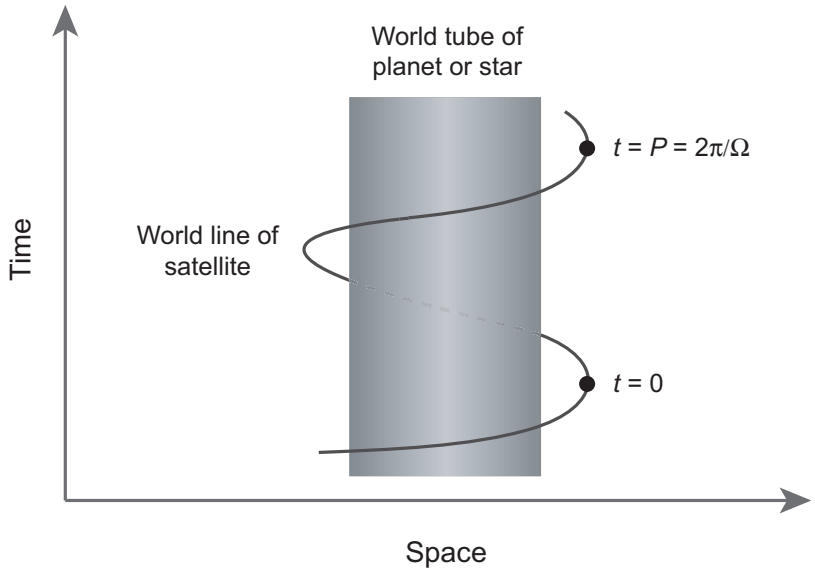


Figure 9.12: Geodetic world line for gyroscope in orbit around a spherical mass.

Imagine a gyroscope on a satellite in a circular Earth orbit.

- Assume that the spin of the gyroscope starts off at $t = 0$ pointing in the radial direction.
- The world lines for the gyroscope and Earth are illustrated in Fig. 9.12.



- After one complete orbit with a period $t = P = 2\pi/\Omega$,

$$\begin{aligned} s^r(t=P) &= s_0(1-2M/R)^{1/2}\cos(\omega P) \\ &= s_0(1-2M/R)^{1/2}\cos\left(2\pi\frac{\omega}{\Omega}\right). \end{aligned}$$

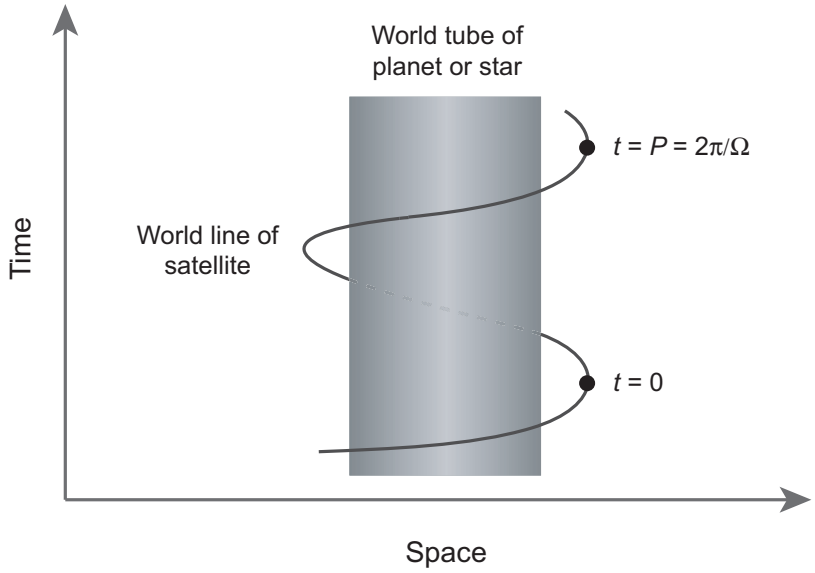
- In the absence of geodetic spin precession (so that $\omega = \Omega$), the angle φ would change by 2π for one orbit.
- Thus the additional spin precession angle for each orbit is

$$\Delta\varphi = 2\pi - 2\pi\frac{\omega}{\Omega} = 2\pi\left(1-\frac{\omega}{\Omega}\right) = 2\pi\left[1-\left(1-\frac{3M}{R}\right)^{1/2}\right],$$

in the direction of the orbital motion, where

$$\omega \equiv \left(1-\frac{3M}{R}\right)^{1/2}\Omega.$$

was used in the last step.



For objects in the Solar System $M/R = GM/Rc^2$ is small, so

- The square root can be expanded to give

$$\Delta\varphi = 2\pi\left[1 - \underbrace{\left(1 - \frac{3M}{R}\right)^{1/2}}_{\text{expand}}\right] \simeq \frac{3\pi M}{R} = \frac{3\pi GM}{c^2 R}\,\text{rad orbit}^{-1},$$

  for the geodetic precession per orbit for gyroscopes on a satellite in Earth orbit.

- The radial direction is perpendicular to the direction of orbital motion.

- Hence, this expression also gives the precession measured by an observer comoving with the gyroscope in orbit.

The precession is small, but *cumulative for successive orbits*.

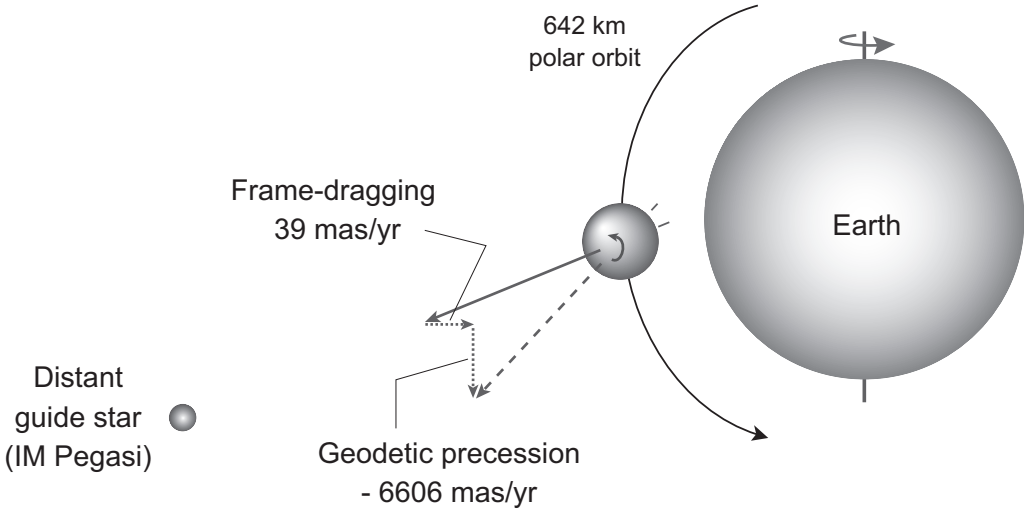


Figure 9.13: Geodetic precession and frame dragging for a gyroscope on Gravity Probe B. Precession angles are exaggerated. IM Pegasi was chosen as the directional reference because it was approximately in the desired direction for the gyroscopic spin axis and its proper motion on the celestial sphere was known precisely.

Gravity Probe B (GP-B) tested geodetic precession for

- gyroscopes aboard a satellite in an almost circular orbit
- that averaged 642 km above the surface of the Earth.

Figure 9.13 illustrates measurement of geodetic precession.

- The expected geodetic precession per orbit is 1.22×10^{-3} arcsec orbit^{-1}.
- This corresponds to a predicted geodetic precession rate of $\Delta\varphi/\Delta t = 6.6$ arcsec yr^{-1}.
- The geodetic precession rate measured by GP-B was within 0.07% of this value.

9.10 Gyroscopes in Rotating Spacetimes

The Schwarzschild solution gives a spacetime valid outside any spherical, static, non-rotating body.

- But *astronomical bodies are typically spinning*.
- Thus it is important physically to ask about *solutions of the Einstein equations for rotating spacetimes*.
- Some objects are spinning slowly, which suggests that their exterior metric might be approximated by an expansion about the Schwarzschild spherical spacetime.
- On the other hand, in later chapters black holes will be encountered that are spinning with an angular momentum comparable to the maximum allowed by the laws of physics.
- These cannot in any sense be understood in terms of perturbations of the Schwarzschild metric.
- In the remainder of this chapter the simpler topic of very *slowly rotating spherical spacetimes* will be taken up;
- in later chapters the more complex issue of strongly-deformed metrics implying potentially large angular momentum will be addressed.

9.10.1 Slow Rotation in the Schwarzschild Metric

As an example of a slowly-rotating astronomical body, consider the Sun.

- It rotates differentially with a period of about a month, somewhat faster at the equator than at the poles.
- The spacetime outside the Sun is described by the Schwarzschild solution provided that
  1. it is a vacuum,
  2. gravity from all other bodies can be neglected,
  3. the Sun is static with no spin, and
  4. the Sun is spherical.
- To a very good approximation these conditions are satisfied, so the Schwarzschild metric is *almost (but not quite)* valid outside the Sun.

Let us now parse the *"not quite"* part of the preceding statement.

Take the exterior of the Sun to be a vacuum and ignore all other masses.

- Thus deviation from the Schwarzschild metric outside the Sun is caused only by its
  1. *angular momentum* and by its
  2. *deviation from spherical symmetry*.
- Deviations from spherical symmetry are very small and caused classically by *centripetal effects* of its rotation.
- If the Schwarzschild metric is expanded in powers of the angular momentum J,
  - Centripetal forces vary as ω^2 and so come in *only at the level of the* J^2 *term* in the expansion.
  - Thus *to 1st order in* J the rotating Sun *is spherical*.
- However, recall that general relativity has many formal similarities with electromagnetism and that electromagnetic forces can arise from *motion of charge.*
- These are *magnetic effects.*
- In general relativity *mass acts as "gravitational charge"*.
- This suggests that gravitational "forces" (curvature of spacetime) may arise from *motion of mass* ("mass currents"), in addition to arising from the mass itself.

This is the case, and forces arising from mass currents are called *gravitomagnetic effects* in general relativity.

Assuming a slowly-rotating spherical field, expansion about the Schwarzschild metric to first order in J gives a metric

$$ds^2 = -\left[1 - \frac{2M}{r} + \mathscr{O}\left(\frac{1}{r^2}\right)\right] dt^2 - \left[\frac{4J\sin^2\theta}{r} + \mathscr{O}\left(\frac{1}{r^2}\right)\right] d\varphi\, dt + \left[1 + \frac{2M}{r} + \mathscr{O}\left(\frac{1}{r^2}\right)\right]\left(dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2)\right).$$

Upon restoring factors of G and c, this may be written as

$$ds^2 = ds_0^2 - \frac{4GJ}{c^3 r^2}\sin^2\theta (r d\varphi)(c dt) + \mathscr{O}\left(J^2\right),$$

where

- ds_0^2 is the contribution from the unperturbed Schwarzschild metric and
- $\mathscr{O}\left(J^2\right)$ indicates that terms of order J^2 and higher have been discarded.

For Newtonian theory $J \sim Mrv$, where v is the linear velocity. Therefore, for the coefficient of the term proportional to J,

$$\frac{GJ}{c^3 r^2} \sim \frac{GMrv}{c^3 r^2} = \frac{v}{c} \times \frac{GM}{c^2 r},$$

which shows that

> The effect of mass motion on curvature is of order v/c relative to the primary effect caused by the mass itself, which is proportional to $GM/c^2 r$.

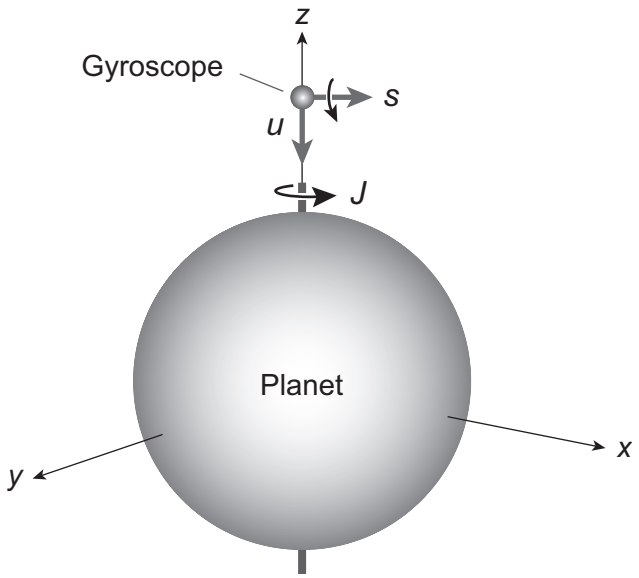


Figure 9.14: A gyroscope in free fall on the rotation axis of a spherical planet in slow rotation. The initial 4-spin s of the gyroscope is perpendicular to the 4-velocity u, and the angular momentum of the planet is J.

9.10.2 Dragging of Inertial Frames

Consider the following thought experiment:

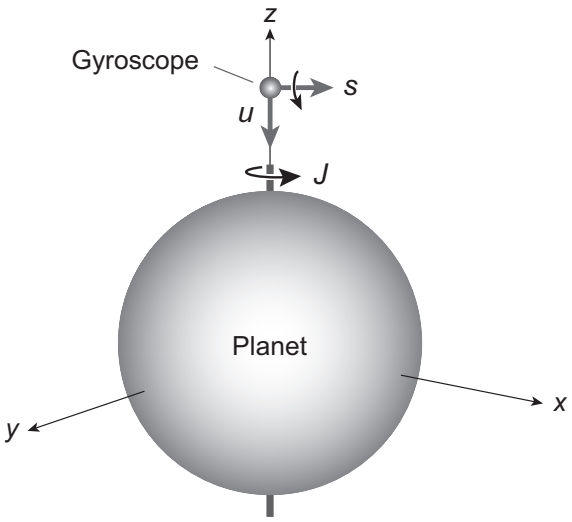
- Imagine a spherical body in slow rotation, with a metric

$$ds^2 = ds_0^2 - \frac{4GJ}{c^3 r^2}\sin^2\theta(rd\varphi)(cdt) + \mathscr{O}\left(J^2\right),$$

  valid approximately in the spacetime surrounding it.

- Now imagine dropping a gyroscope from above the North Pole with an initial spin perpendicular to the rotation axis.

Figure 9.14 illustrates.



Spherical coordinates are singular on the z axis, so it is convenient to work in cartesian coordinates (see figure above).

- In cartesian coordinates the Schwarzschild metric is

$$ds^2 = ds_0^2(\text{cartesian}) - \frac{4GJ}{c^3 r^2}(cdt)\left(\frac{xdy - ydx}{r}\right),$$

  where $ds_0^2(\text{cartesian})$ is the unperturbed metric.

- Only terms up to order $1/c^3$ need be retained and terms involving the mass M will contribute at order $1/c^5$.

- Thus for small J the unperturbed Schwarzschild metric can be replaced by its $M \to 0$ limit, which is just the *flat Minkowski metric*.

- Thus, it is sufficient to work with the approximate metric

$$ds^2 = \underbrace{-(cdt)^2 + dx^2 + dy^2 + dz^2}_{\text{Minkowski metric}} - \underbrace{\frac{4GJ}{c^3 r^2}(cdt)\left(\frac{xdy - ydx}{r}\right)}_{\text{1st-order gravitomagnetic}}.$$

Taking the spin to lie in the x–y plane, initially

$$u^\mu = (u^t, 0, 0, u^z) \qquad s^\mu = (0, s^x, s^y, 0),$$

so clearly

- $s \cdot u = 0$, and *by symmetry arguments*
- the *spin will remain in the x–y plane*.

The *spin of the freely-falling gyroscope* will be governed by

$$\frac{ds^\mu}{d\tau} + \Gamma^\mu_{\alpha\beta} s^\alpha u^\beta = 0.$$

To leading order in $1/c$ the only contributions to the second term will involve

$$\Gamma^1_{02} \equiv \Gamma^x_{ty} = \frac{2GJ}{c^2 z^3} \qquad \Gamma^2_{01} \equiv \Gamma^y_{tx} = -\frac{2GJ}{c^2 z^3}$$

where $r = z$, since the gyroscope lies on the z axis. Then

$$\frac{ds^\mu}{d\tau} + \Gamma^\mu_{\alpha\beta} s^\alpha u^\beta = 0$$

yields the equations

$$\frac{ds^x}{d\tau} = -\Gamma^x_{yt} s^y u^t = -\frac{2GJ}{c^2 z^3} s^y u^t \qquad \frac{ds^y}{d\tau} = -\Gamma^y_{xt} s^x u^t = \frac{2GJ}{c^2 z^3} s^x u^t.$$

Utilizing $u^t = dt/d\tau$, the equations

$$\frac{ds^x}{d\tau} = -\Gamma^x_{yt} s^y u^t = -\frac{2GJ}{c^2 z^3} s^y u^t \qquad \frac{ds^y}{d\tau} = -\Gamma^y_{xt} s^x u^t = \frac{2GJ}{c^2 z^3} s^x u^t .$$

may be written

$$\frac{ds^x}{dt} = -\frac{2GJ}{c^2 z^3} s^y \qquad \frac{ds^y}{dt} = \frac{2GJ}{c^2 z^3} s^x ,$$

and the *angular rate of precession* in the x–y plane is

$$\Omega_{\mathrm{LT}} = \frac{2GJ}{c^2 z^3} .$$

The result

$$\Omega_{\mathrm{LT}} = \frac{2GJ}{c^2 z^3}.$$

was obtained in a Lorentz frame for which

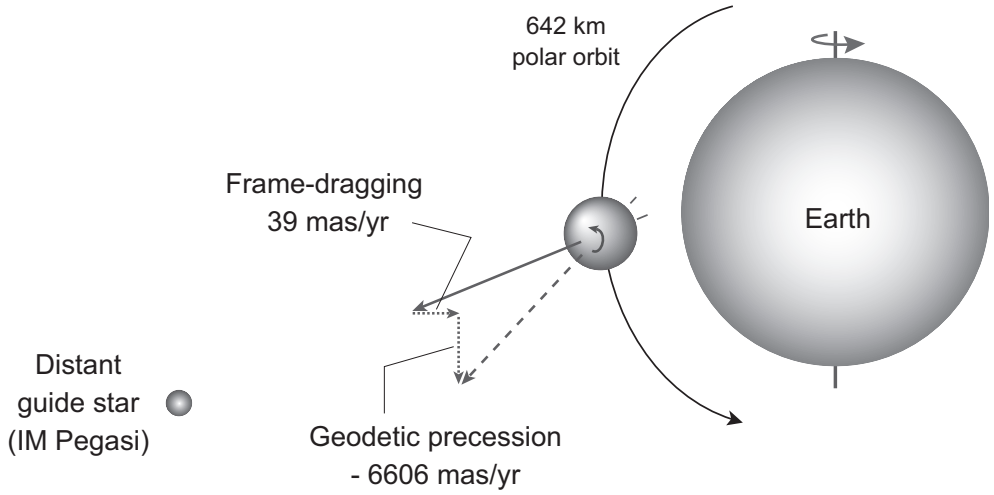
- the source of the spherical field is at rest and the gyroscope is falling.
- However, it is *valid also in the frame of the gyroscope* because
  - Lorentz boosts along the z axis do not affect the transverse spin components s^x and s^y, and
  - there is no time dilation to leading order in c.

This gyroscopic precession effect is called the *Lense–Thirring effect* or *frame dragging*.

Frame dragging should be distinguished from *geodetic precession*, which is

- much larger and
- occurs even if the gravitational field source is not rotating.

Gravity Probe B measured frame dragging for Earth's slowly-rotating field and found a value *within 5% of the GR prediction*.



Effect	GP-B measurement	General relativity
Geodetic precession	$-6601.8 \pm 18.3\,\text{mas yr}^{-1}$	$6606.1\,\text{mas yr}^{-1}$
Frame dragging	$-37.2 \pm 7.2\,\text{mas yr}^{-1}$	$-39.2\,\text{mas yr}^{-1}$

The table above compares the predictions of general relativity with measurements from Gravity Probe B for geodetic precession and frame dragging in a polar Earth orbit.

Thus GP-B found

- *geodetic precession* within 0.07% and
- *frame dragging* within 5%

of GR predictions.

Frame dragging has been illustrated here specifically for a *slowly-rotating Schwarzschild metric*.

- However, dragging of inertial frames is expected for any metric having *a field source that depends on angular momentum*.
- It is a very small effect for Earth's gravitational field.
- However, in later chapters extreme frame-dragging effects will be discussed that can occur in
  - much stronger,
  - much more rapidly-rotating

  gravitational fields.

These may be of *large astrophysical importance* because frame dragging of spacetime around rotating black holes may help power some of the most energetic events observed in the Universe.

Chapter 10

Neutron Stars and Pulsars

Neutron stars are relevant to our discussion of general relativity on two levels.

- They are of *considerable intrinsic interest* because their quantitative description requires solution of the Einstein equations in the presence of matter.
- In addition, they also explain the *existence of pulsars* and these in turn provide the most stringent observational tests of general relativity.

10.1 A Qualitative Picture of Neutron Stars

Objects of white dwarf density are described reasonably well by Newtonian gravity.

- For *white dwarfs*, general relativity corrections appear typically at the $\sim 10^{-4}$ level.
- For *neutron stars* the densities are much higher;
- Newtonian gravity works *qualitatively* for neutron stars, but a correct *quantitative description* requires GR.
- However, many of their *basic properties* can be *estimated using Newtonian concepts* and simple reasoning.
- For example, the assumption that in a neutron star gravity packs the neutrons down to their hard-core radius of $\sim 10^{-13}$ cm yields immediately that
  - the most massive neutron stars contain about 3×10^{57} baryons (mostly neutrons)
  - within a radius of about 7 km,
  - with a corresponding mass of about $2.3\,M_\odot$.
- This implies an average density greater than $10^{15}\,\mathrm{g\,cm^{-3}}$, which is several times nuclear matter density, and
- a (gravitational) binding energy of order 100 MeV per nucleon, which is an order of magnitude larger than the binding energy of nucleons in nuclear matter.

- The *total gravitational binding energy* of a neutron star is *within an order of magnitude of the rest mass energy*, and
- the *escape velocity* is about 50% *of the speed of light*.
- The binding energy and the escape velocity both signal that *GR effects are likely to be significant*.

General relativity is important for the overall properties of neutron stars.

- However, over a microscopic scale characteristic of nuclear interactions the *metric is essentially constant*.
- This implies that the *microphysics* (nuclear and elementary particle interactions) of the neutron star can be described by *special-relativistic quantum field theory*.

Thus it is possible to decouple

- gravity, which governs the overall structure, from
- quantum mechanics, which governs the microscopic properties.

in the study of neutron stars.

10.2 The Oppenheimer–Volkov Equations

Neutron stars have $\rho \sim 10^{14} - 10^{15}\,\mathrm{g\,cm^{-3}}$.

- This produces gravitational fields that are of *moderate strength by GR standards* (enormous by Earth standards).
- *Escape velocity* at the surface is around $\frac{1}{3}c - \frac{1}{2}c$.
- Thus *GR is necessary* for a correct description.
- Unlike vacuum solutions, we must now deal with *mass distributions* and a *finite stress–energy tensor*.
- We shall, however, simplify by assuming a *static, spherically symmetric configuration* for the matter.
- With these assumptions we may assume that
  - the solution *outside* the neutron star corresponds to the Schwarzschild solution, so the interior solution
  - must match Schwarzschild at the surface.

We consider the general solution of the Einstein equations for the gravitational field produced by

- a static, spherical mass distribution that
- matches the exterior (vacuum) solution at the surface of the spherical mass distribution.

- Assume the matter inside the star is a *perfect fluid*, with

$$T^{\mu}{}_{\nu} = (\varepsilon + P)u^{\mu}u_{\nu} + P\delta^{\mu}_{\nu}.$$

- We assume *spherical symmetry*, with a line element

$$ds^2 = -e^{\sigma(r)}dt^2 + e^{\lambda(r)}dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\varphi^2,$$

  implying non-vanishing metric components

$$g_{00}(r) = -e^{\sigma(r)} \quad g_{11}(r) = e^{\lambda(r)}$$
$$g_{22}(r) = r^2 \quad g_{33}(r,\theta) = r^2 \sin^2\theta.$$

- *Assume equilibrium*, so
  - $\sigma(r)$ and $\lambda(r)$ depend on r but not t, and
  - the 4-velocity *has no space components*:

$$u^{\mu} = (e^{-\sigma/2}, 0, 0, 0) = (g_{00}^{-1/2}, 0, 0, 0).$$

- Inserting these 4-velocity components, the *stress–energy tensor* takes the diagonal form

$$T^{\mu}{}_{\nu} = (\varepsilon + P)u^{\mu}u_{\nu} + P\delta^{\mu}_{\nu} = \begin{bmatrix} -\varepsilon & 0 & 0 & 0 \\ 0 & P & 0 & 0 \\ 0 & 0 & P & 0 \\ 0 & 0 & 0 & P \end{bmatrix}$$

  where $\varepsilon = \rho$ in $c = 1$ units.

- For the vacuum Einstein equation we need only the Ricci tensor to construct the Einstein tensor.
- However, in the general non-vacuum case we need both the *Ricci tensor* $R_{\mu\nu}$ and the *Ricci scalar* R.
- It is convenient to express the *Einstein equation* as

$$G^{\nu}{}_{\mu} \equiv R^{\nu}{}_{\mu} - \tfrac{1}{2}\delta^{\nu}_{\mu} R = 8\pi T^{\nu}{}_{\mu}.$$

  $T^{\nu}{}_{\mu}$ is diagonal, so only diagonal components of $G^{\nu}{}_{\mu}$.
- Because of the Bianchi identity and the Einstein equations

$$G^{\mu}{}_{\nu} = 8\pi T^{\mu}{}_{\nu}$$

  the stress–energy tensor obeys

$$T^{\mu}{}_{\nu;\mu} = 0.$$

- Thus we can choose to solve the equation $T^{\mu}{}_{\nu;\mu} = 0$ in place of solving one of the Einstein equations.

We shall employ that strategy here, using *two Einstein equations* and the *constraint equation* $T^{\mu}{}_{\nu;\mu} = 0$ to obtain a solution.

The constraint equation has been solved in a Problem, where you were asked to show that

$$T^{\mu}{}_{\nu;\mu} = 0 \quad \longrightarrow \quad P' + \tfrac{1}{2}(P+\rho)\sigma' = 0.$$

(primes denoting partial derivatives with respect to r) for a line element and stress–energy tensor

$$ds^2 = -e^{\sigma(r)}dt^2 + e^{\lambda(r)}dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\varphi^2,$$

$$T^{\mu}{}_{\nu} = \begin{bmatrix} -\varepsilon & 0 & 0 & 0 \\ 0 & -P & 0 & 0 \\ 0 & 0 & -P & 0 \\ 0 & 0 & 0 & -P \end{bmatrix}$$

We require *two additional equations;* the simplest choices are the Einstein equations

$$G^0{}_0 = 8\pi T^0{}_0 \qquad G^1{}_1 = 8\pi T^1{}_1$$

- The Einstein tensors G_{00} and G_{11} were derived for this metric in a Problem.
- Using contraction with the metric tensor to raise an index we obtain from those results

$$G^0{}_0 = g^{00}G_{00} = -e^{-\sigma}G_{00} = e^{-\lambda}\left(\frac{1}{r^2} - \frac{\lambda'}{r}\right) - \frac{1}{r^2}$$

$$G^1{}_1 = g^{11}G_{11} = e^{-\lambda}G_{11} = e^{-\lambda}\left(\frac{1}{r^2} + \frac{\sigma'}{r}\right) - \frac{1}{r^2}.$$

Collecting the preceding equations, we find that we must solve the set of equations

$$-e^{-\lambda}\left(\frac{1}{r^2}-\frac{\lambda'}{r}\right)+\frac{1}{r^2}=8\pi\varepsilon(r)$$
$$e^{-\lambda}\left(\frac{1}{r^2}-\frac{\sigma'}{r}\right)-\frac{1}{r^2}=8\pi P(r)$$
$$P'+\tfrac{1}{2}(P+\rho)\sigma'=0.$$

To proceed we note that the first equation above may be rewritten as

$$G^0{}_0=\frac{1}{r^2}\frac{d}{dr}\left[r\left(1-e^{-\lambda}\right)\right]=\frac{2}{r^2}\frac{dm}{dr}=8\pi\varepsilon,$$

where we have defined a new parameter

$$2m(r)\equiv r(1-e^{-\lambda}).$$

At this point $m(r)$ is only a reparameterization of the metric coefficient e^{λ} since, upon multiplying by e^{λ},

$$e^{\lambda}=\frac{r}{r-2m(r)}=\left(1-\frac{2m(r)}{r}\right)^{-1},$$

but $m(r)$ will be interpreted below as *the total mass–energy enclosed within the radius* r.

From the first Einstein equation

$$G^0{}_0 = \frac{2}{r^2}\frac{dm}{dr} = 8\pi\varepsilon \quad \rightarrow \quad dm = 4\pi r^2 \varepsilon dr,$$

and thus

$$m(r) = 4\pi \int_0^r \varepsilon(r) r^2 \, dr,$$

with an integration constant $m(0) = 0$ chosen on physical grounds. Now consider the second Einstein equation

$$e^{-\lambda}\left(\frac{1}{r^2} - \frac{\sigma'}{r}\right) - \frac{1}{r^2} = 8\pi P(r)$$

Solving it for $\sigma' = d\sigma/dr$ gives

$$\frac{d\sigma}{dr} = e^{\lambda}\left(8\pi r P(r) + \frac{1}{r}\right) - \frac{1}{r},$$

and substitution of

$$e^{\lambda} = \frac{r}{r - 2m(r)}$$

leads to

$$\frac{d\sigma}{dr} = \frac{8\pi r^3 P(r) + 2m(r)}{r(r - 2m(r))}.$$

Therefore the preceding equations define the metric coefficients e^{σ} and e^{λ} in terms of the parameter $m(r)$ and the pressure $P(r)$.

Finally, we may combine the two equations

$$P' + \tfrac{1}{2}(P+\rho)\sigma' = 0 \qquad \frac{d\sigma}{dr} = \frac{8\pi r^3 P(r) + 2m(r)}{r(r-2m(r))}$$

to give

$$\frac{dP}{dr} = -\frac{(P(r)+\varepsilon(r))(4\pi r^3 P(r)+m(r))}{r(r-2m(r))}.$$

Collecting our results, we have obtained the Oppenheimer–Volkov equations for the structure of a static, spherical, gravitating perfect fluid

$$\frac{dP}{dr} = \frac{(P(r)+\varepsilon(r))\left(m(r)+4\pi r^3 P(r)\right)}{r^2\left(1-\dfrac{2m(r)}{r}\right)},$$

$$m(r) = 4\pi \int_0^r \varepsilon(r) r^2\, dr$$

where $m(r)$ is the total mass contained within a radius r

Oppenheimer–Volkov equations:

$$\frac{dP}{dr} = \frac{\left(P(r)+\varepsilon(r)\right)\left(m(r)+4\pi r^3 P(r)\right)}{r^2\left(1-\dfrac{2m(r)}{r}\right)},$$

$$m(r) = 4\pi\int_0^r \varepsilon(r) r^2\, dr$$

- Solution of these equations requires specification of an *equation of state* that relates the density to the pressure.
- They may then be integrated from the origin outward
  - with initial conditions $m(r=0)=0$ and
  - an arbitrary choice for the central density $\varepsilon(r=0)$

  until the pressure $P(r)$ becomes zero.
- This defines the surface of the star $r=R$, with the mass of the star given by $m(R)$.
- For a given equation of state each choice of $\varepsilon(0)$ will give a unique R and $m(R)$ when the equations are integrated.
- This defines a family of stars characterized by
  - a specific equation of state and
  - the value of a single parameter (the central density, or a quantity related to it like central pressure).

These equations represent the general relativistic (covariant) description of hydrostatic equilibrium for a spherical, gravitating perfect fluid.

- The condition of hydrostatic equilibrium was built into the solution through the assumption

$$u^\mu = (e^{-\sigma/2}, 0, 0, 0) = (g_{00}^{-1/2}, 0, 0, 0).$$

- This implies that the fluid is static since the 4-velocity has no non-zero space components.
- They reduce to the Newtonian description of hydrostatic equilibrium in the limit of weak gravitational fields

However, the Oppenheimer–Volkov equations imply *significant deviations from the Newtonian description* in strong gravitational fields.

- To see this clearly, a little algebra allows us to rewrite them in the form (Problem)

$$4\pi r^2 dP(r) = \frac{-m(r)dm(r)}{r^2} \times \left(1 + \frac{P(r)}{\varepsilon(r)}\right)\left(1 + \frac{4\pi r^3 P(r)}{m(r)}\right)\left(1 - \frac{2m(r)}{r}\right)^{-1}$$

$$dm(r) = 4\pi r^2 \varepsilon(r) dr.$$

These equations may be interpreted in the following way:

$$\underbrace{4\pi r^2 dP(r)}_{\text{Force acting on shell}} = \underbrace{\frac{-m(r)dm(r)}{r^2}}_{\text{Newtonian}}$$

$$\times \left(1+\underbrace{\frac{P(r)}{\varepsilon(r)}}_{\text{GR}}\right)\left(1+\underbrace{\frac{4\pi r^3 P(r)}{m(r)}}_{\text{GR}}\right)\left(1-\underbrace{\frac{2m(r)}{r}}_{\text{GR}}\right)^{-1}$$

$$dM(r) = \underbrace{4\pi r^2 \varepsilon(r) dr}_{\text{Mass–energy of shell}} .$$

- The second equation gives the mass–energy of a shell lying between radii r and $r+dr$.
- The left side of the first equation is the net force acting outward on this shell.
- The first factor on the right side of the first equation is the attractive Newtonian gravity acting on the shell because of the mass interior to it.
- The last three factors on the right side of the first equation—the factors on the second line—represent GR effects causing deviation from Newtonian gravitation.
- Since all three factors on the second line of the first equation exceed unity as the star becomes relativistic, *in GR gravity is consistently stronger than in Newtonian gravity*.

Note: *Gravity is enhanced by coupling to pressure* in the GR description, *unlike Newtonian gravity*.

10.3 Interpretation of the Mass Parameter

The parameter $m(r)$ entering the Oppenheimer–Volkov equations was interpreted provisionally as the *total mass–energy enclosed within a radius* r. Let's now justify this interpretation.

- Outside a star of radius R, the mass function $m(r)$ becomes equal to $m(R)$, which is
- the *mass that would be detected through Kepler's laws* for the orbital motion of a well-separated binary system.
- In the *Newtonian limit* it is clear from

$$m(r) = 4\pi \int_0^r \varepsilon(r) r^2 \, dr,$$

  that $m(r)$ is the *mass* contained within the radius r.
- For relativistic stars $m(r)$ may be consistently split into
  - contributions from a *rest mass* $m_0(r)$,
  - an *internal energy* $U(r)$, and
  - a *gravitational energy* $\Omega(r)$:

$$m(r) = m_0(r) + U(r) + \Omega(r).$$

- Formally we can split the energy density ε into a contribution from the rest mass and one from internal energy,

$$\varepsilon = \mu_0 n + (\varepsilon - \mu_0 n),$$

  where the first term is the total rest mass of n particles of mass μ_0, and the second term is the internal energy.

- The *proper volume* for a spherical shell of thickness dr is

$$\begin{aligned} dV &= 4\pi r^2 \sqrt{\det g_{11}}\, dr = 4\pi r^2 \sqrt{e^{\lambda}}\, dr \\ &= 4\pi r^2 (1-2m/r)^{-1/2} dr. \end{aligned}$$

  Thus the *total rest mass* inside the radius r is

$$m_0(r) = \int_0^r \mu_0 n\, dV = 4\pi \int_0^r r^2 (1-2m/r)^{-1/2} \mu_0 n\, dr,$$

  the *total internal energy* inside r is

$$\begin{aligned} U(r) &= \int_0^r (\varepsilon - \mu_0 n)\, dV \\ &= 4\pi \int_0^r r^2 (1-2m/r)^{-1/2} (\varepsilon - \mu_0 n)\, dr, \end{aligned}$$

  and the *total mass–energy* inside r is

$$m(r) = 4\pi \int_0^r \varepsilon(r) r^2\, dr.$$

- Thus, the difference

$$\begin{aligned} \Omega(r) &= m(r) - m_0(r) - U(r) \\ &= -4\pi \int_0^r r^2 \varepsilon \left(1 - (1-2m/r)^{-1/2}\right) dr \end{aligned}$$

  must be the *total gravitational energy* inside r.

This gives us some confidence that $m(r)$ is indeed the *total mass–energy* inside the coordinate r.

10.3.1 Gravitational Mass and Baryonic Mass

The integral

$$m(r) = 4\pi \int_0^r \varepsilon(r) r^2 \, dr$$

is of the same form as that for Newtonian gravity if the mass distribution is given by $\rho(r) = \varepsilon(r)/c^2$.

- However, in general relativity $\varepsilon(r)$ is *not an arbitrary distribution* but rather corresponds to a solution P of

$$\frac{dP}{dr} = \frac{\left(P(r) + \varepsilon(r)\right)\left(m(r) + 4\pi r^3 P(r)\right)}{r^2\left(1 - \dfrac{2m(r)}{r}\right)},$$

  with an equation of state $\varepsilon = \varepsilon(P)$. Despite the form of

$$m(r) = 4\pi \int_0^r \varepsilon(r) r^2 \, dr,$$

- $m(r)$ is the *sum* of the mass and the gravitational energy,
- and *the mass of the star has no well-defined meaning* in isolation from the gravitational energy.
- The mass–energy m is termed the *gravitational mass*.
- The total mass of the nucleons if they were dispersed to infinity is termed the *baryonic mass* of the star.
- The gravitational mass and the baryonic mass are *not the same* (they differ by the *gravitational binding energy*).

Gravitational mass $\sim 20\%$ smaller than *baryonic mass*.

10.4 Pulsars and Tests of General Relativity

Pulsars are *rapidly-spinning neutron stars* that sweep beamed radiation over the Earth periodically, giving the illusion of pulsation.

- This apparent pulsation occurs with *atomic-clock precision*.
- Thus, pulsars offer possibilities for *precise timing measurements*, particularly when they are found as a component of a binary star system.
- The structure and evolution of pulsars is of considerable intrinsic interest.
- However, they also are of great practical importance for general relativity because of their superb timing characteristics.
- These provide some of the *most precise tests available* for the theory.

In the remainder of this chapter examples will be given of the stringent tests of general relativity afforded by *pulsars in close binary systems*.

10.4.1 The Binary Pulsar

The *Binary Pulsar PSR 1913+16* (also known as the *Hulse–Taylor pulsar*) was discovered using the Arecibo 305 meter radio antenna.

- It is about 5 kpc away, near the boundary of the constellations Aquila and Sagitta.
- This pulsar rotates 17 times a second, giving a pulsation period of 59 milliseconds.
- It is in a binary system with another neutron star (not a pulsar), with a 7.75 hour period.
- The precise repetition frequency of the pulsar means that it is basically a very high quality clock
  - orbiting in a binary system that
  - feels very strong, time-varying gravitational effects.

$\longrightarrow$ Precision tests of general relativity

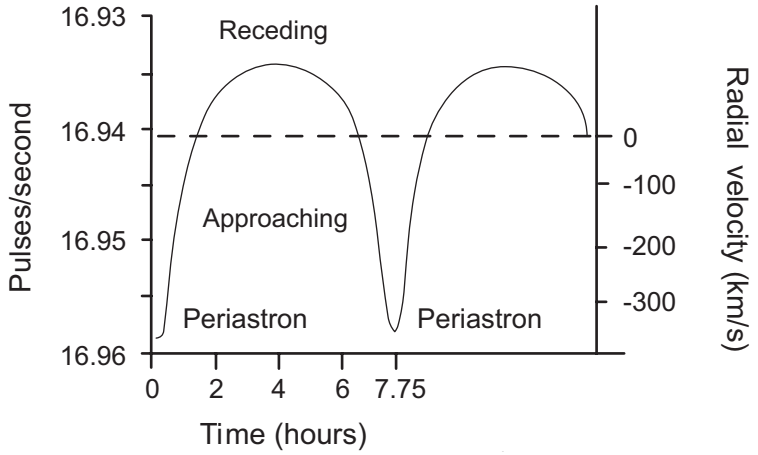


Figure 10.1: Pulse rate and inferred radial velocity as a function of time for the Binary Pulsar.

10.4.2 Periodic Variations

The repetition period for a pulsar is associated with the spin of the pulsar and is atomic-clock-like in its precision. Thus

- Variations in that period as observed from Earth must be associated with orbital motion in the binary.
- These variations can be used to give very precise information about the orbit.
- When the pulsar is moving toward us, the repetition rate of the pulses as observed from Earth will be higher than when the pulsar is moving away (Doppler effect).
- This can be used to measure the radial velocity (Fig. 10.1).

The pulse arrival times vary as the pulsar moves through its orbit

- It takes three seconds longer for the pulses to arrive from the far side of the orbit than from the near side.
- From this, the Binary Pulsar orbit can be inferred to be about a million kilometers (three light seconds) further away from Earth when on the far side of its orbit than when on the near side.

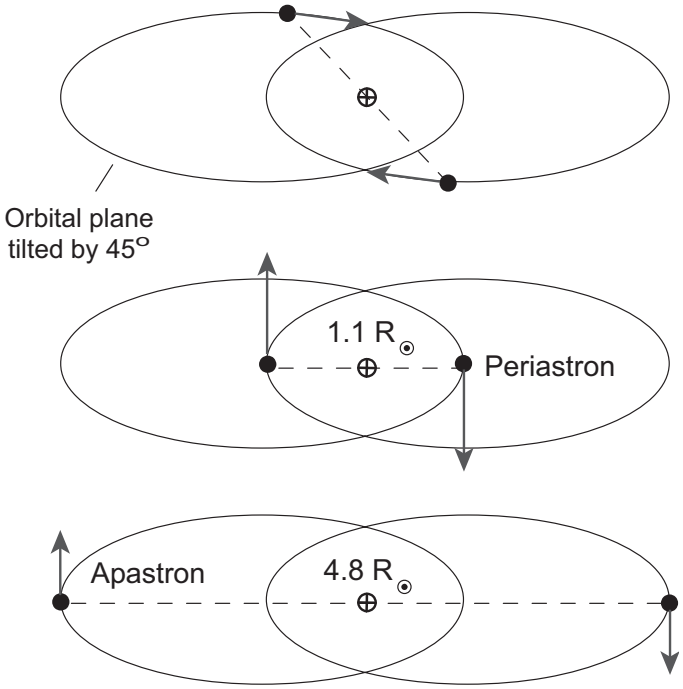


Figure 10.2: Binary Pulsar orbits.

10.4.3 Orbital Characteristics

The orbits determined for the binary are shown in Fig. 10.2.

- Each neutron star has a mass of about $1.4M_\odot$.
- The orbits are very eccentric ($e \sim 0.6$.).
- The minimum separation (periastron) is about $1.1R_\odot$.
- The maximum separation (apastron) is about $4.8R_\odot$
- The orbital plane is inclined by about 45 degrees.

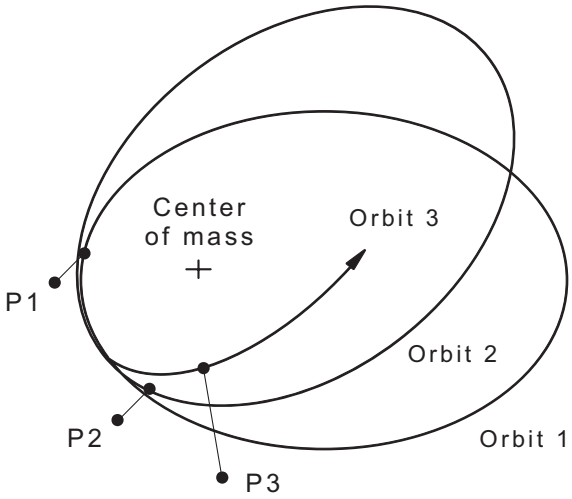


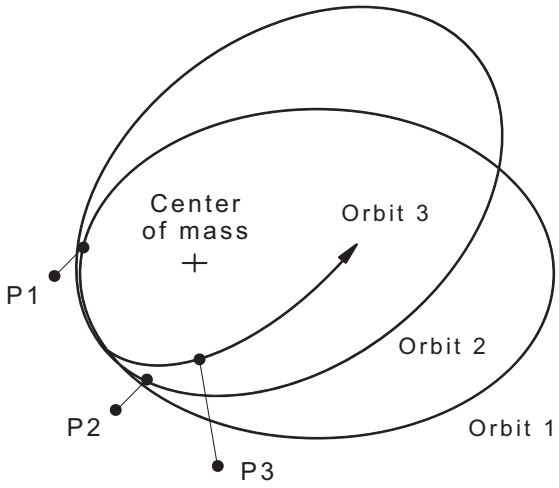
Figure 10.3: Precession of the periastron.

- By Kepler's laws, the radial velocity of the pulsar varies substantially as it moves around its elliptical orbit.
- These orbits are *not quite closed ellipses* because of precession effects associated with general relativity.
  - This causes the *location of the periastron* to shift a small amount for each revolution (Fig. 10.3).
  - The points P1, P2, and P3 are periastrons on three successive orbits (with the amount of precession greatly exaggerated for clarity

10.5 Precision Tests of General Relativity

The discovery and study of the Binary Pulsar was of such fundamental importance that *Taylor and Hulse* were awarded the *1993 Nobel Prize in Physics* for their work

- This was the only Nobel ever given for relativity before the 2017 prize for discovery of gravitational waves.
- Chief among the reasons for this importance is that the Binary Pulsar provided the most stringent tests of GR available before
  - the discovery of the *Double Pulsar* and
  - the direct *observation of gravitational waves*.



Precession of Orbits

Because spacetime is warped by the gravitational field in the vicinity of the pulsar, the *orbit will precess with time*.

- This is the same effect as the precession of the perihelion of Mercury, but it is much larger for the present case.
- The Binary Pulsar's periastron advances by *4.2 degrees per year*, in accord with the predictions of GR.
- In a single day the orbit of the Binary Pulsar advances by as much as the orbit of Mercury advances in a century!

Time Dilation

- When near periastron, gravity is stronger and its velocity is higher, so time should run slower.
- Conversely, near apastron the field is weaker and the velocity lower, so time should run faster.
- It does both, in the amount predicted by GR.

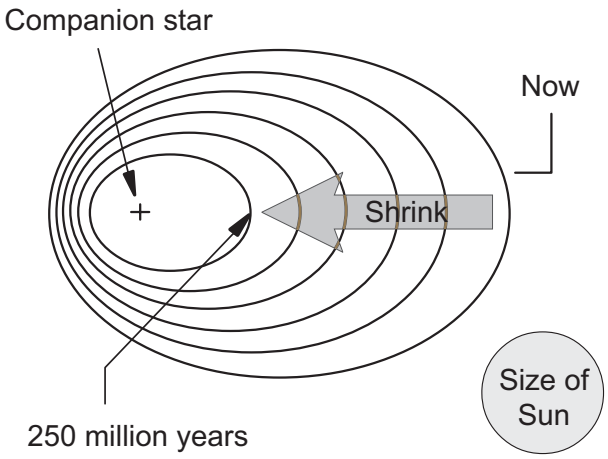
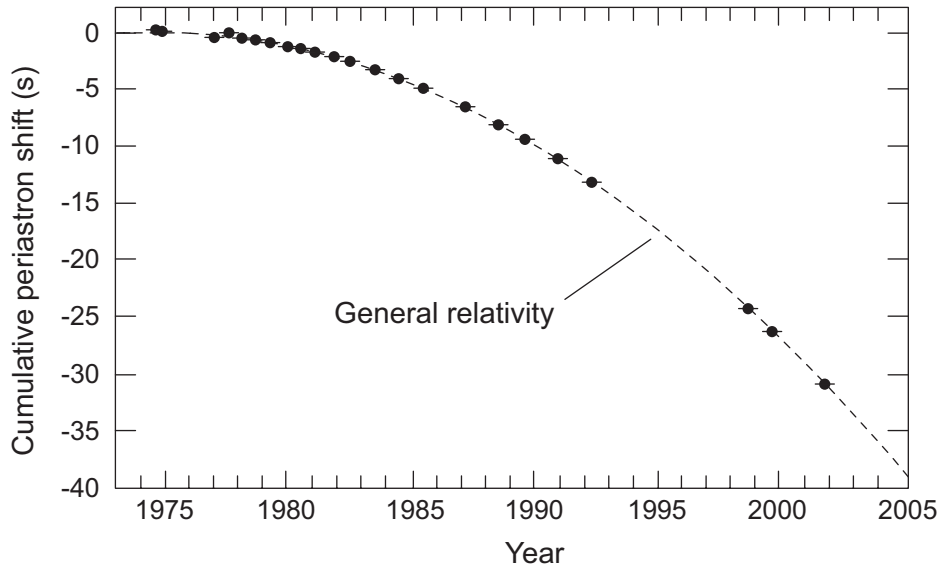


Figure 10.4: Shrinkage of the orbit of the Binary Pulsar because of gravitational wave emission.

10.5.1 Emission of Gravitational Waves

The revolving pair of masses is predicted by general relativity to radiate gravitational waves, causing the orbit to shrink (Fig. 10.4).

- The time of periastron can be measured very precisely and is found to be shifting.
- This shift corresponds to a decrease in the orbital period by *76 millionths of a second per year*.
- The corresponding decrease in the size of the orbit by about *3.3 millimeters per revolution*.



The quantitative decrease in periastron time is illustrated by the data points in the above figure.

- Because the orbital period is short, the shift in periastron arrival time has accumulated to more than 30 seconds (earlier) since discovery.
- This decay of the size of the orbit is in agreement with the amount of energy that general relativity predicts should be leaving the system in the form of gravitational waves (dashed line in figure)
- Precision measurements on the Binary Pulsar gave strong indirect evidence for the correctness of this key prediction of general relavity even before gravitational waves were detected directly.

10.5.2 Origin and Fate of the Binary Pulsar

Formation of a neutron star binary is not easy. One of two things must happen

- A binary must form with two stars massive enough to become supernovae and produce neutron stars, and the neutron stars thus formed must remain bound to each other through the two supernova explosions.
- The neutron star binary must result from gravitational capture of one neutron star by another.

These are improbable events, but not impossible, and the existence of the Binary Pulsar (and several similar systems) demonstrates empirically that mechanisms exist for it to happen.

Once a neutron star binary is formed its orbital motion radiates energy as gravitational waves, the orbits must shrink, and eventually the two neutron stars *must merge.*

- Because of the gravitational wave radiation and the corresponding shrinkage of the Binary Pulsar orbit (3.3 millimeters per revolution), merger is predicted in about 300 million years.
- The sum of the masses of the two neutron stars is likely above the critical mass to form a black hole. Therefore,the probable fate of the Binary Pulsar is merger and collapse to a rotating (Kerr) black hole.
- As two neutron stars in a binary approach each other they will revolve faster (Kepler's third law).
- This will cause them to emit gravitational radiation more rapidly, which will in turn cause the orbit to shrink even faster.
- Thus, near merger of two neutron stars will proceed rapidly in a positive-feedback runaway and will emit very strong gravitational waves that may be detectable with current-generation gravitational wave detectors.

These considerations are valid for any binary, not just the Binary Pulsar, but the gravitational wave effects are much more pronounced for binaries involving highly compact objects.

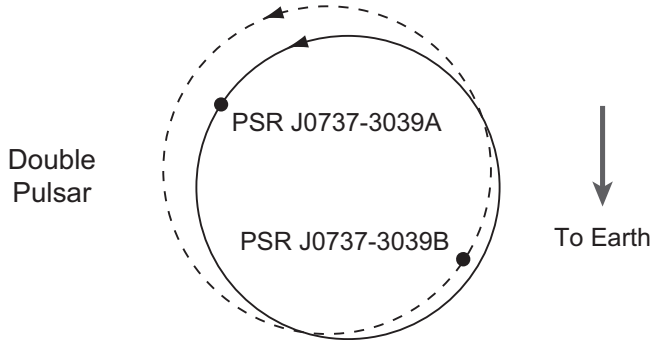


Figure 10.5: Orbital configuration of the Double Pulsar.

10.5.3 The Double Pulsar

In 2003 a binary neutron star system (the Double Pulsar) was discovered in which *both* neutron stars were observed as pulsars in a very tight, partially eclipsing orbit (Fig. 10.5).

- The two neutron stars have masses of $1.3381 \pm 0.0007 M_\odot$ (component A), and $1.2489 \pm 0.0007 M_\odot$ (component B).
- They have spin periods of 22.7 ms and 2.77 s.
- The orbit is slightly eccentric ($e = 0.088$).
- The orbit has a mean radius of about $1.25 R_\odot$.
- Thus the orbital period is only 147 minutes, with a mean orbital velocity of about 10^6 km hr^{-1}.
- The fast orbital period and the exquisite timing from the pulsar clocks has allowed the Double Pulsar to give the most precise tests of general relativity to date.

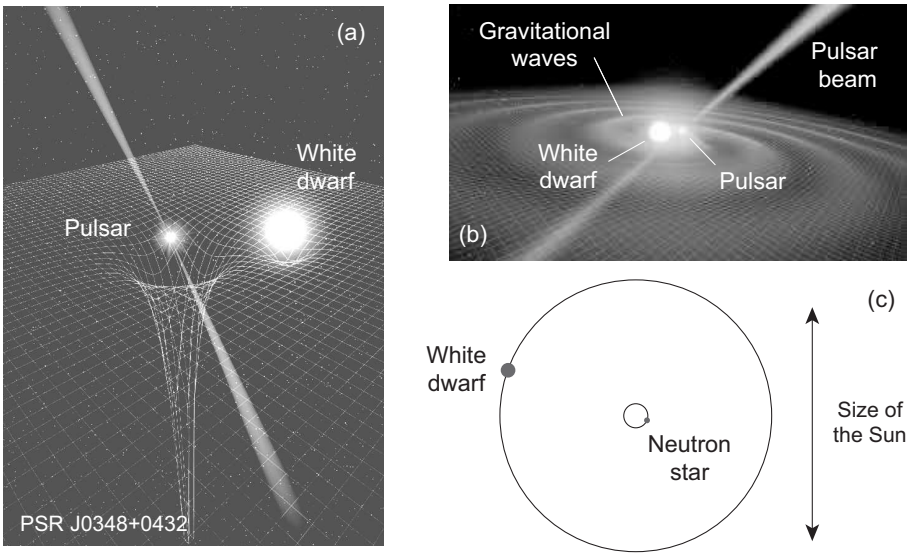
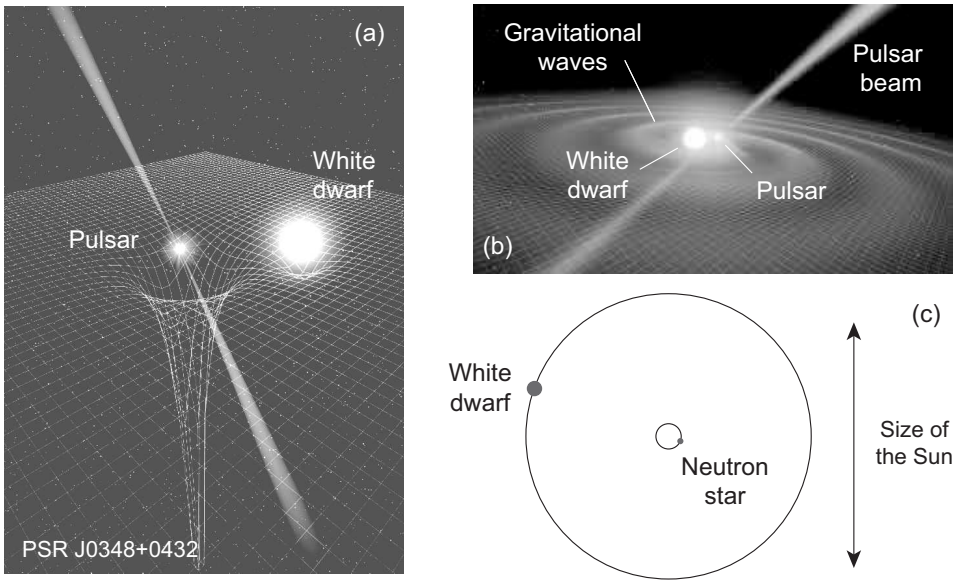


Figure 10.6: (a) (Highly-exaggerated) distortion of spacetime by the pulsar–white dwarf binary PSR J0348+0432. The large, compact mass of the neutron star distorts spacetime much more than the smaller, less-compact mass of the white dwarf. (b) Gravitational wave emission from PSR_J0348+0432. (c) Binary system PSR J0348+0432. Orbits to scale but object sizes are schematic.

10.5.4 The Pulsar–White Dwarf Binary PSR J0348+0432

The binary system PSR J0348+0432

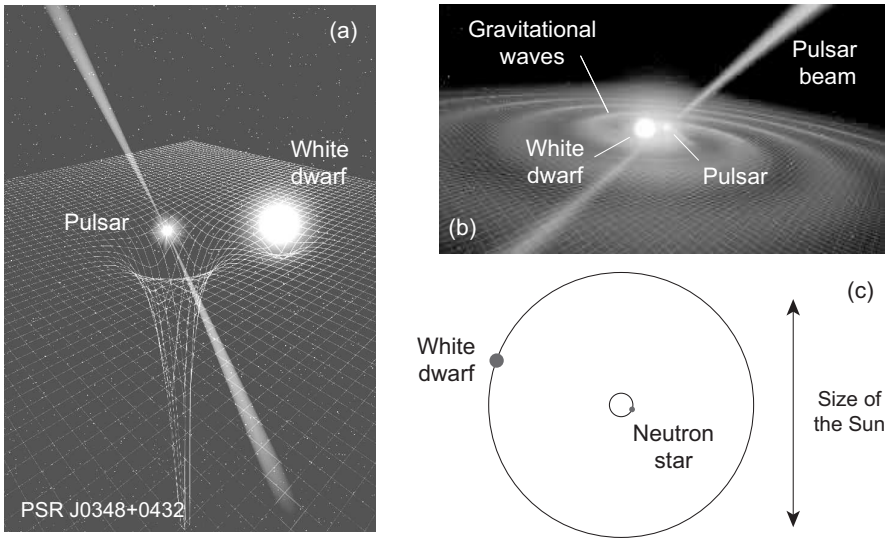
- is about 2.1 kpc away and
- contains a 39 ms pulsar of mass $2.01\,M_\odot$ and a white dwarf of mass $0.172\,M_\odot$, with orbital period 2.46 hours.
- An artist's impression of the distorted spacetime produced by PSR J0348+0432 is shown in Fig. 10.6(a),
- Emission of gravitational waves illustrated in Fig. 10.6(b).
- The geometry of the orbit is illustrated in Fig. 10.6(c).



The very compact orbit illustrated above (comparable in width to the diameter of the Sun)

- suggests that there should be a *rapid loss of orbital energy to gravitational waves*.
- Precise radio timing indicates that the orbital period is decreasing by $8.6\,\mu\text{sec}\,\text{year}^{-1}$, in accord with the prediction of general relativity

This implies a time until merger of $\sim$600 million years.



The *computed rate of decrease in orbital period* for PSR J0348+0432 because of gravitational wave emission is

$$\dot{P}_{\rm GR} = -2.58 \times 10^{-13}\,{\rm s\,s^{-1}}$$

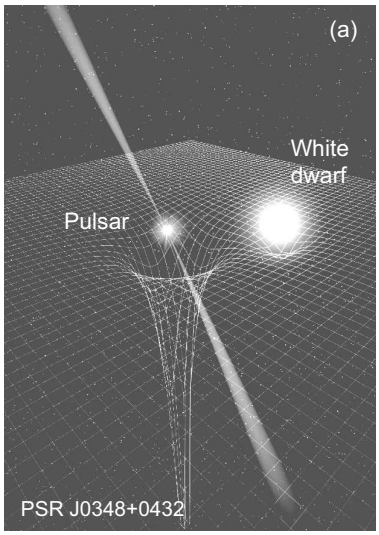
and the *measured rate of decrease* in the orbital period is

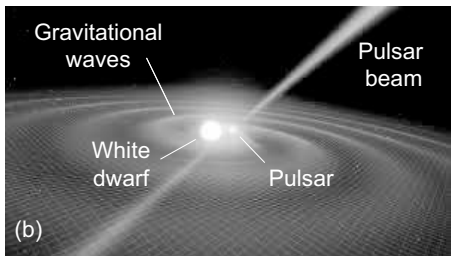
$$\begin{aligned}\dot{P} &= -8.6\,\mu{\rm s\,yr^{-1}} \left(\frac{1\,{\rm s}}{10^6\,\mu{\rm s}}\right)\left(\frac{1\,{\rm yr}}{3.1557\times 10^7\,{\rm s}}\right) \\ &= -2.73 \times 10^{-13}\,{\rm s\,s^{-1}}.\end{aligned}$$

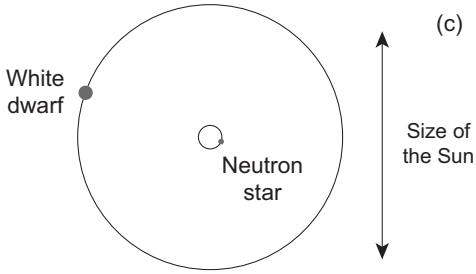
Therefore, the ratio of measured to predicted rate of decrease is

$$\frac{\dot{P}}{\dot{P}_{\rm GR}} = 1.05 \pm 0.18.$$

The orbital period of PSR J0348+0432 is decreasing at a rate accounted for by the rate of gravitational-wave emission required by GR.







The pulsar PSR J0348+0432

- contains one of the *most massive neutron stars known* ($M \sim 2M_\odot$), with a gravitational binding energy that is
- 60% larger than that of neutron stars in any other binary where gravitational-wave damping has been measured.
- Thus, PSR J0348+0432 *tests general relativity under stronger gravity* than for other binary pulsar tests.

Strong-field effects in general relativity

- depend *nonlinearly on the gravitational binding energy*.
- Thus *larger binding energy* entails a substantially more *stringent test of general relativity*.

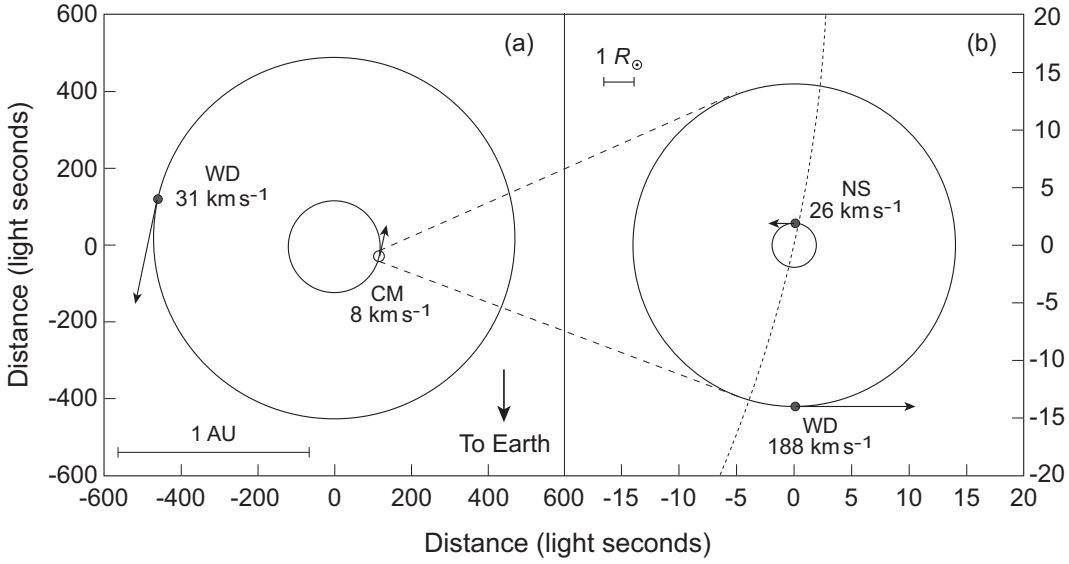
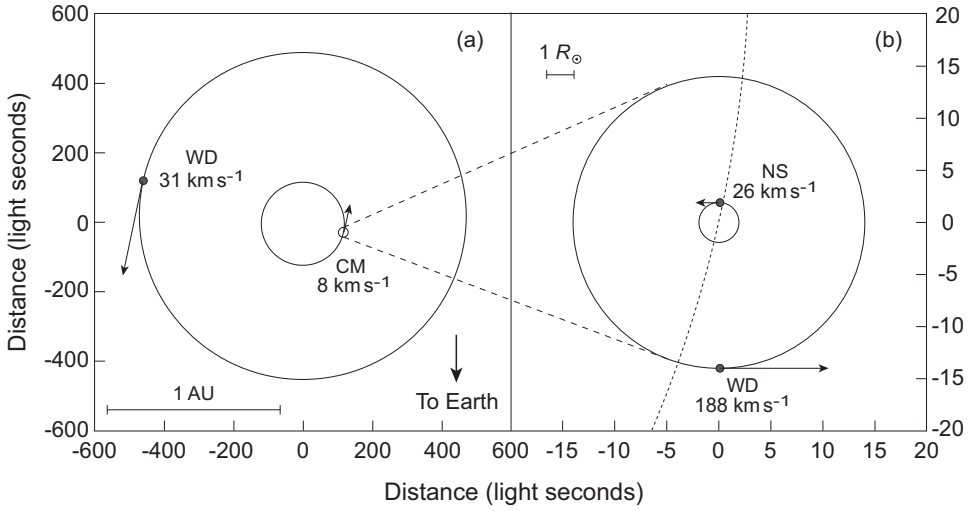


Figure 10.7: Orbits of PSR J0337+1715. (a) Orbits of the outer white dwarf (WD) and the center of mass (CM) for the inner white dwarf and neutron star pair. (b) As for left side but scaled up by a factor of 30 to show the orbits for the inner white dwarf and neutron star (NS). Arrows indicate orbital velocities for the center of mass of the inner binary and the individual white dwarfs and neutron star. All orbits lie almost in the same plane, are nearly circular, and have a tilt angle $i \sim 39^\circ$ relative to the line of sight.

10.5.5 The Pulsar–WD–WD Triplet PSR J0337+1715

The triple star system PSR J0337+1715, which contains a *millisecond pulsar* and *two white dwarfs*, is illustrated in Fig. 10.7.

- Pulsar timing allows a precise determination of masses and orbital parameters.
- The *neutron star* is found to have a mass of $1.4378 M_\odot$,
- the *inner white dwarf* has a mass of $0.1975 M_\odot$, and
- the *outer white dwarf* has a mass of $0.4101 M_\odot$.



According to the *strong equivalence principle*,

- objects with *different gravitational binding energies* should *follow the same orbits* in a gravitational field.
- This can be tested by *tracking the Earth–Moon system* as it *falls gravitationally toward the Sun* in its orbit.

In PSR J0337+1715 a similar test is possible as the outer white dwarf strongly accelerates the inner binary.

- Gravitational binding energies for the neutron star and white dwarf in the inner binary differ from each other by 4–5 orders of magnitude, and
- the neutron star binding energy is roughly 10^9 times larger than that of planets or moons in our Solar System.

Hence violations of the strong equivalence principle should be greatly amplified in J0337+1715.

Part II

Black Holes

Chapter 11

Spherical Black Holes

One of the most spectacular consequences of general relativity is the prediction that gravitational fields can become so strong that they can *effectively trap even light*.

- Space becomes so curved that there are no paths for light to follow from an interior to exterior region.
- Such objects are called *black holes*, and there is extremely strong circumstantial evidence that they exist.
- In this chapter we apply the Einstein theory of gravity to the idea of black holes using the Schwarzschild solution.
- In the next chapter we shall take a first step in considering how gravitational physics is altered if the principles of quantum mechanics come into play (*Hawking black holes*),
- In the chapter after that we shall consider how the Schwarzschild solution is modified if a black hole is assumed to possess angular momentum (*Kerr black holes*.)

11.1 Schwarzschild Black Holes

We shall now show that

- there is an *event horizon* in the Schwarzschild spacetime at $r_S = 2M$,
- which implies that there is a *black hole* inside the event horizon,

where the *escape velocity exceeds* c.

11.1.1 Event Horizons

Imagine an attempt to escape a gravitational field generated by some spherical object of mass M.

- The *condition for escape* to a radius r is that the kinetic energy exceed the gravitational potential energy:

$$\tfrac{1}{2}mv^2 \geq \frac{GMm}{r}.$$

- But the maximal physical velocity for any object is $v = c$; substituting c for v and solving for r,

$$r = \frac{2GMm}{mc^2} = \frac{2GM}{c^2} = r_S,$$

where we've used $r_S = 2M$ with c and G restored.

> Therefore, r_S is the radius at which *the escape velocity equals the velocity of light*.

This is just a suggestive result from Newtonian physics

- supplemented by concepts from special relativity and
- dubious assumptions about light in Newtonian gravity.

But a more rigorous analysis comes to the same conclusion:

> The gravitational curvature is so strong inside r_S that *even light cannot escape*.

Thus, the *Schwarzschild radius* r_S

- may also be termed the *event horizon* of the Schwarzschild solution and, since light is
- *fundamentally trapped* inside this radius,

the region *interior to* r_S is termed a *black hole*.

Before proceeding we clear up a potential source of confusion:

- It has been argued that the Schwarzschild solution is approximately valid outside the Sun or Earth, and that
- $r_s = 2M$ defines an event horizon for a black hole.
- *So why aren't the Sun and Earth black holes* with corresponding event horizons?

The answer: The Schwarzschild solution is a vacuum solution *valid only outside the mass distribution* producing the gravitational field.

In the case of the Sun and Earth it may be verified easily that

- the Schwarzschild radius *lies deep inside both objects*,
- where the Schwarzschild solution is *not valid.*

Thus (you will be happy to know!), the Earth and Sun are not black holes because they are *not nearly compact enough*.

An event horizon and the associated black-hole properties of the Schwarzschild solution

- manifest themselves physically only if the mass M responsible for the Schwarzschild gravitational field
- is *contained entirely within the Schwarzschild radius*,
- which implies *extremely compact objects* of much higher density than planets or normal stars.

Thus it is important to distinguish

- the *Schwarzschild spacetime,* which is approximately valid outside any static spherical mass, and a
- *Schwarzschild black hole,* which forms in a Schwarzschild spacetime only if the mass M responsible for the gravitational field is *contained entirely within its Schwarzschild radius*.

The preceding qualitative discussion can be placed on firmer ground by considering a spacecraft approaching an event horizon in free fall (engines off).

- For simplicity, we assume the *trajectory to be radial*.
- and consider two points of view:
  1. From a point at constant large distance from the black hole *(professors sipping martinis)*.
  2. From a point inside the spacecraft *(the students)*.
- We shall use the *Schwarzschild solution (metric)* for analysis.

11.1.2 Approaching the Event Horizon: Outside View

We consider *only radial motion*. Setting $d\theta = d\varphi = 0$ in the line element

$$ds^2 = -d\tau^2 = -\left(1 - \frac{2M}{r}\right) dt^2 + \left(1 - \frac{2M}{r}\right)^{-1} dr^2$$
$$= -\left(1 - \frac{r_S}{r}\right) dt^2 + \left(1 - \frac{r_S}{r}\right)^{-1} dr^2.$$

- As the spacecraft approaches the event horizon its velocity as *viewed from the outside in a fixed frame* is $v = dr/dt$.
- Light signals from spacecraft travel on the light cone ($ds^2 = 0$) and thus from the line element

$$v = \frac{dr}{dt} = \left(1 - \frac{r_S}{r}\right).$$

- As viewed from a distance outside r_S, *the spacecraft appears to slow as it approaches* r_S and eventually stops as $r \to r_S$.

The distant observers *(professors sipping martinis)* will never see the spacecraft cross r_S: *its image will remain frozen at* $r = r_S$ for all eternity.

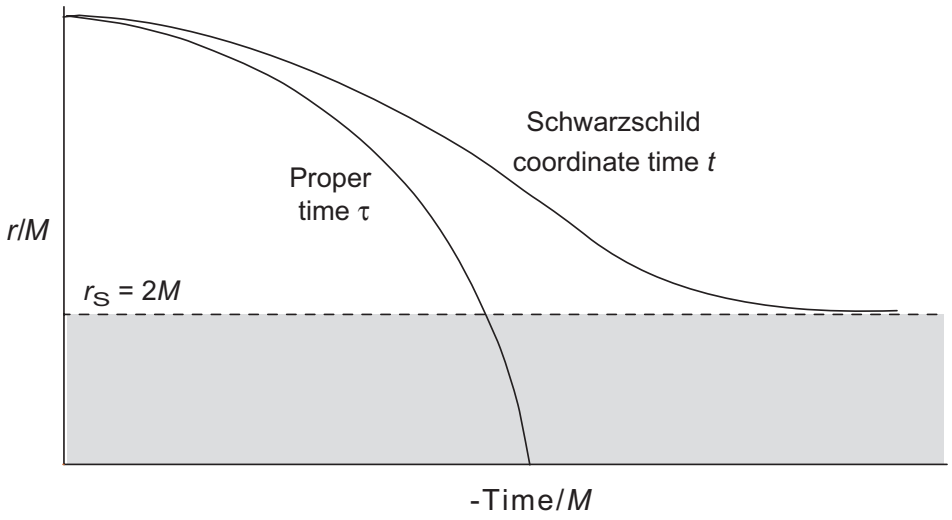
But let's examine what this means more carefully. Rewrite

$$\frac{dr}{dt} = \left(1 - \frac{r_S}{r}\right) \quad \longrightarrow \quad dt = \frac{dr}{1 - r_S/r}.$$

- As $r \to r_S$, time between successive wave crests for light coming from the spacecraft tends to infinity and therefore

$$\lambda \to \infty \qquad \nu \to 0 \qquad E \to 0.$$

- The external observer sees the spacecraft slow rapidly as it nears r_S, but the spacecraft image is seen to *strongly redshift* at the same time.
- This behavior is just that of the *Schwarzschild coordinate time* seen earlier for a test particle in radial free fall:



- Therefore, the actual external observation is that the spacecraft *rapidly slows and redshifts* until the
- *image fades from view* before the spacecraft reaches r_S.

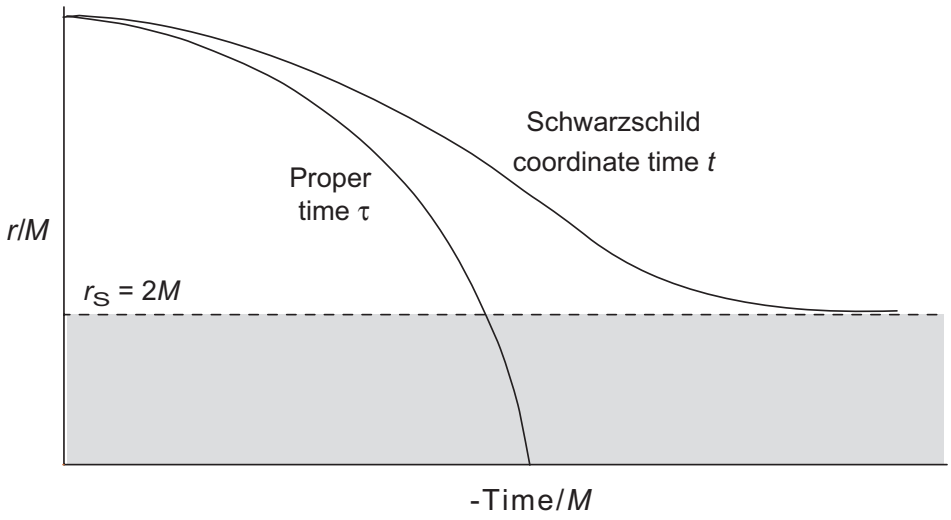
11.1.3 Approaching the Event Horizon: Spacecraft View

Things are very different as viewed by the *(quite doomed) students* from the interior of the spacecraft.

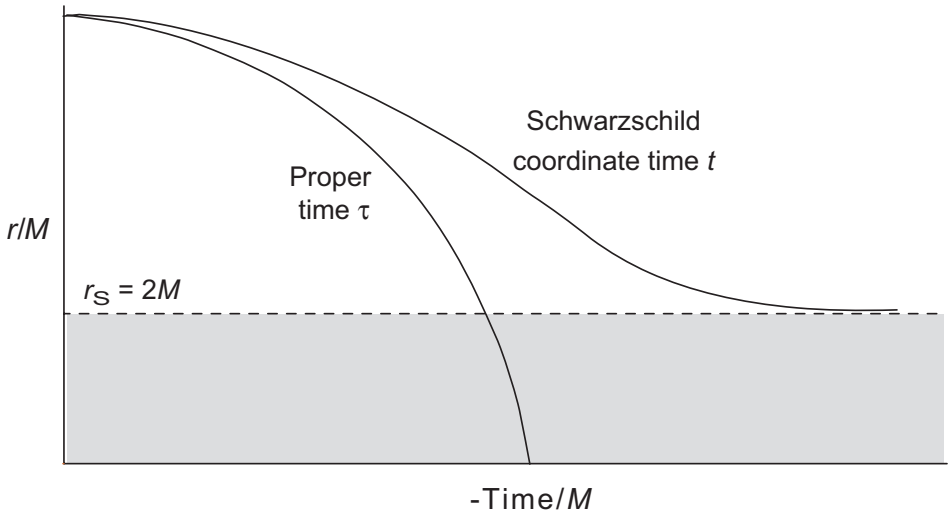
- The occupants will use their own clocks *(measuring proper time)* to gauge the passage of time.
- Starting from a radial position r_0 outside the event horizon, the spacecraft will *reach the origin in a proper time*

$$\tau = -\frac{2}{3}\frac{r_0^{3/2}}{(2M)^{1/2}},$$

as indicated by the *Schwarzschild proper time* in the following plot:



- The spacecraft occupants will generally notice *no spacetime singularity at the horizon*.



- Any tidal forces at the horizon may be very large but will remain finite (*Riemann curvature is finite at* r_S).
- The spacecraft crosses r_S and reaches the *(real) singularity* at $r = 0$ in a finite amount of time,
- where it would encounter infinite tidal forces (*Riemann curvature has components that become infinite at* $r = 0$).
- The trip from r_S to the singularity is very fast (Problem):
  1. $\sim 10^{-4}$ seconds for stellar-mass black holes.
  2. ~ 10 minutes for a billion solar mass black hole.

For $10^9 M_\odot$ black holes the *tidal forces at the horizon* are *small and quite survivable,* but will grow quickly to infinity as the singularity is approached.

11.2 Lightcone Description of a Trip to a Black Hole

Consider a *lightcone description* of a trip into a Schwarzschild black hole.

- Assuming *radial light rays,*

$$\underbrace{d\theta = d\varphi = 0}_{\text{radial}} \qquad \underbrace{ds^2 = 0}_{\text{light rays}}$$

  the line element reduces to

$$ds^2 = -\left(1 - \frac{2M}{r}\right) dt^2 + \left(1 - \frac{2M}{r}\right)^{-1} dr^2 = 0.$$

- Thus the *equation for the lightcone at a local coordinate* r can be read directly from the metric

$$\frac{dt}{dr} = \pm \left(1 - \frac{2M}{r}\right)^{-1}.$$

  - The plus sign corresponds to *outgoing photons* (r increasing with time for $r > 2M$)
  - The minus sign to *ingoing photons* (r decreasing with time for $r > 2M$)

- For large r

$$\frac{dt}{dr} = \pm \left(1 - \frac{2M}{r}\right)^{-1}.$$

  becomes equal to ± 1, *as for flat spacetime*.

- However as $r \to r_S$ the forward lightcone *opening angle tends to zero* as $dt/dr \to \infty$.

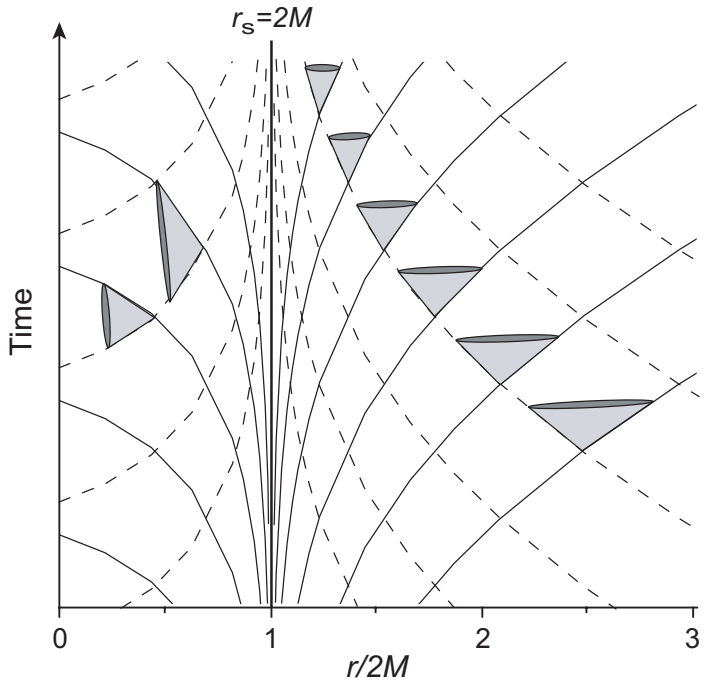


Figure 11.1: Lightcone structure of Schwarzschild spacetime.

Integrating

$$\frac{dt}{dr} = \pm\left(1 - \frac{2M}{r}\right)^{-1}.$$

gives

$$t = \begin{cases} -r - 2M\ln|r/2M - 1| + \text{constant} & \text{(Ingoing)} \\ r + 2M\ln|r/2M - 1| + \text{constant} & \text{(Outgoing)} \end{cases}$$

- Null geodesics defined by this are plotted in Fig. 11.1.
- Tangents at the intersections of the dashed and solid lines *define local lightcones* corresponding to dt/dr, which are *sketched at some spacetime points*.

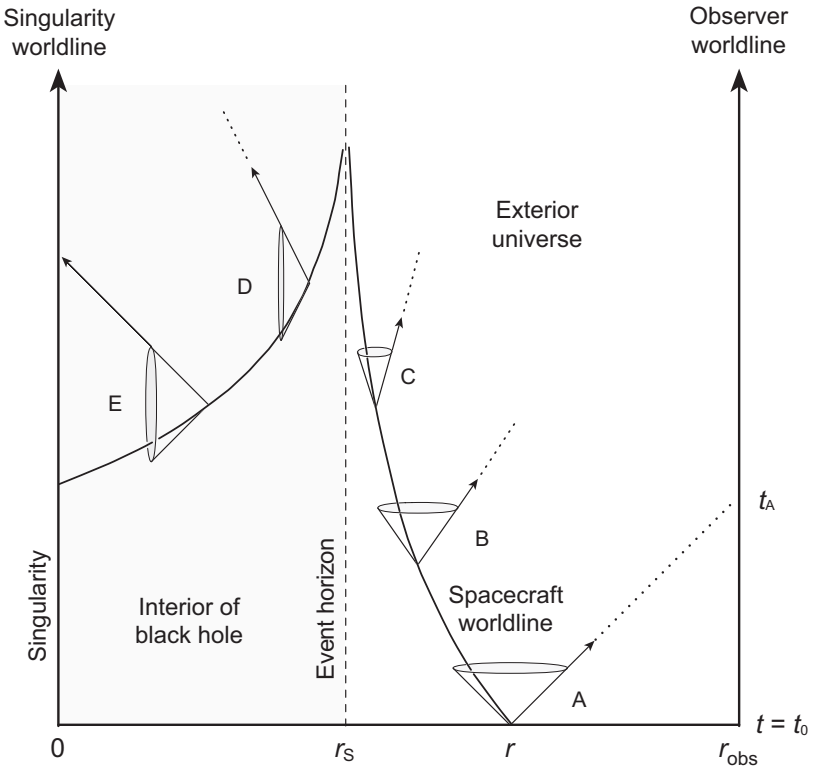
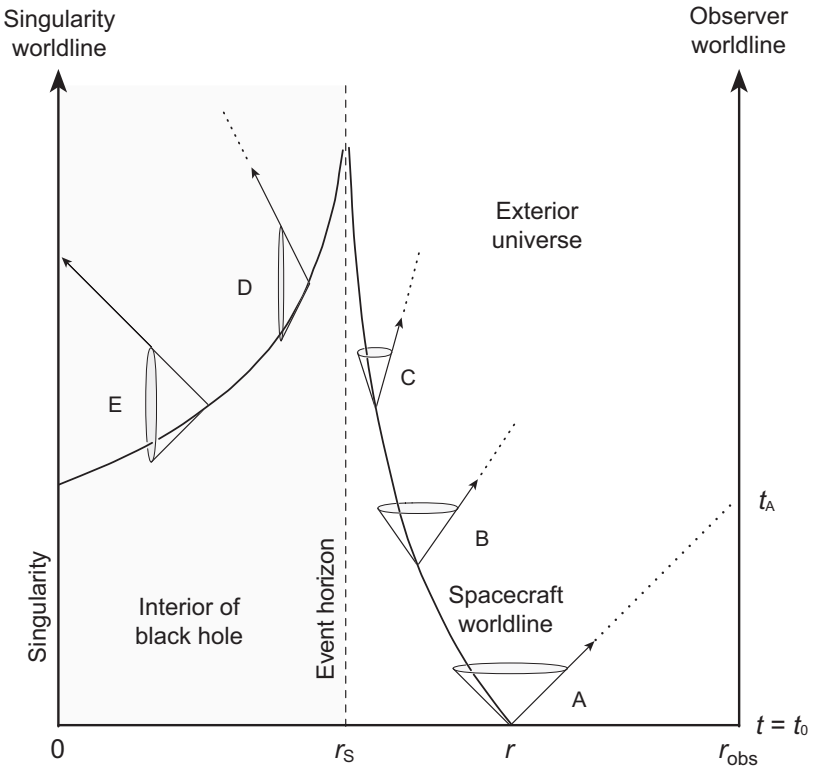


Figure 11.2: Light cone description of a trip into a black hole.

- The worldline of a spacecraft is illustrated in Fig. 11.2, starting well exterior to the black hole. Gravity is weak there and the light cone has the usual appearance.
- As illustrated by the dotted line from A, a light signal emitted from the spacecraft can intersect the worldline of an observer at constant distance $r_{\rm obs}$ in a finite time $t_{\rm A} > t_0$.
- As the spacecraft falls toward the black hole on its worldline the *forward light cone narrows,* since

$$\frac{dt}{dr} = \pm\left(1 - \frac{2M}{r}\right)^{-1}.$$



- Now, at B a light signal can intersect the external observer worldline only at a distant point in the future (arrow on light cone B).
- As the spacecraft approaches r_S,
  - the opening angle of the forward light cone tends to zero and
  - a signal emitted from the spacecraft tends toward *infinite time* to reach the external observer's worldline at r_{obs} (arrow on light cone C).
  - The external observer sees *infinite redshift.*

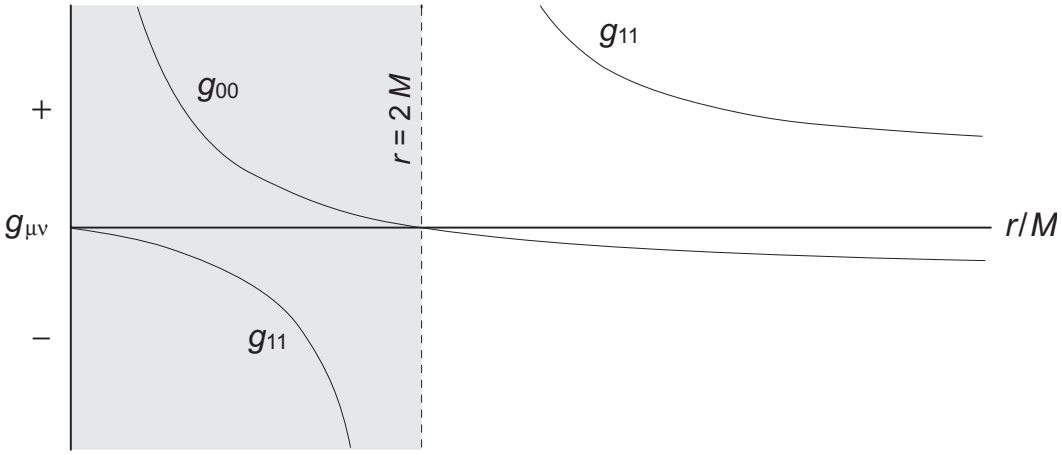


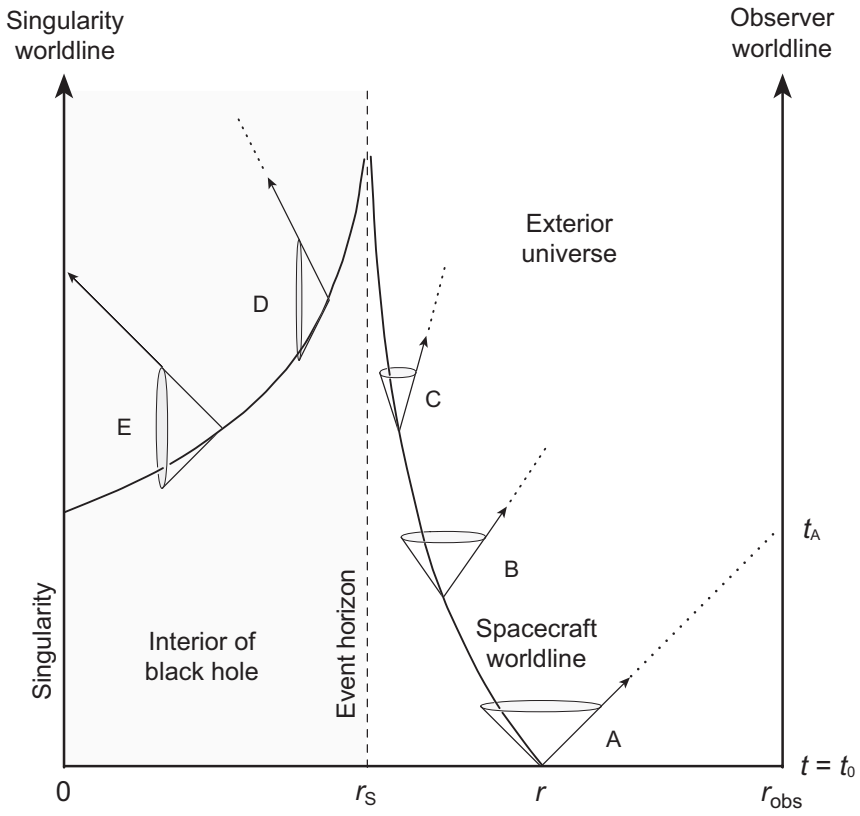
Figure 11.3: Spacelike and timelike regions for g_{00} and g_{11}.

Now consider light cones *interior to the event horizon*.

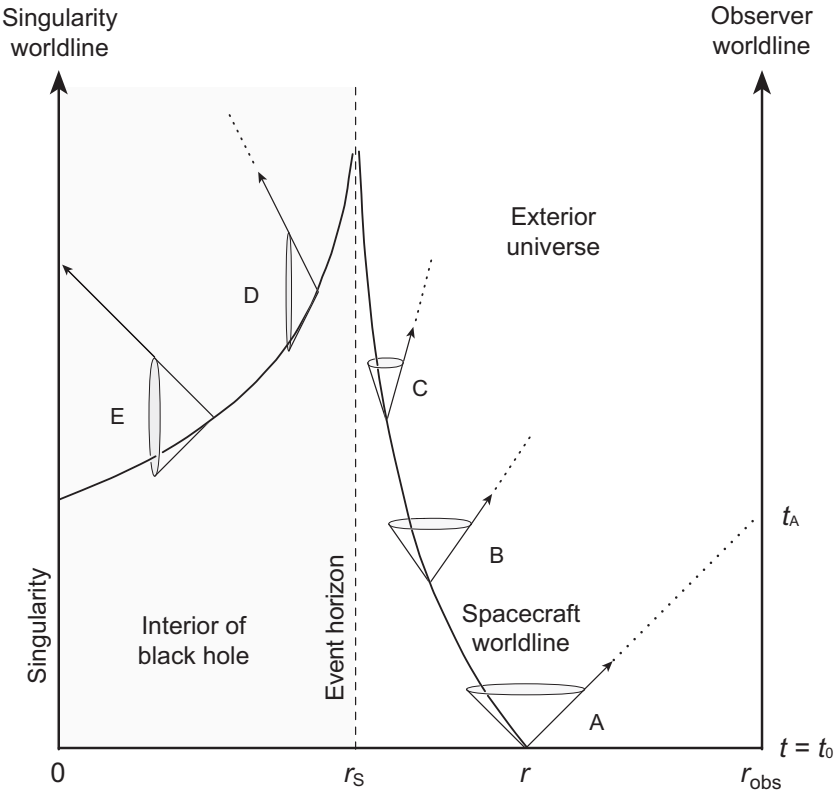
- From the radial and time parts of the Schwarzschild metric illustrated in Fig. 11.3, we observe that *dr and dt reverse their character at the horizon* ($r = 2M$).

  1. This is because the metric coefficients g_{00} and g_{11} *switch signs* at that point.
  2. Outside the event horizon the t direction, $\partial/\partial t$, is timelike ($g_{00} < 0$) and the r direction, $\partial/\partial r$, is spacelike ($g_{11} > 0$).
  3. Inside the event horizon, $\partial/\partial t$ is spacelike ($g_{00} > 0$) and $\partial/\partial r$ is timelike ($g_{11} < 0$).

- Thus inside the event horizon the lightcones get rotated by $\frac{\pi}{2}$ relative to outside (space $\leftrightarrow$ time coordinates).



- The worldline of the spacecraft descends inside r_S because the coordinate time decreases (it is now behaving like r) and the decrease in r represents the passage of time, but the *proper time is continuously increasing* in this region.
- *Outside the horizon* r *is spacelike* and sufficient rocket power can reverse the infall and make r increase.
- *Inside the horizon* r *is timelike* and no application of rocket power can reverse the direction of time.
- The radial coordinate of the spacecraft must decrease inside the horizon, *for the same reason that time flows into the future* in normal experience (whatever that reason is!).



- Inside the horizon there are *no paths* in the forward light-cone of the spacecraft that can reach the external observer at r_0 (the right vertical axis)—see lightcones D and E.

  All timelike and null paths are *bounded by the horizon* and *must encounter the* $r = 0$ *singularity*.

- Here is the real reason that nothing can escape. *Dynamics (building a better rocket) are irrelevant:* once inside r_S
  - the geometry of spacetime permits no forward light cones that intersect exterior regions, and
  - no forward light cones that don't contain the origin.

Thus, there is *no escape from the classical Schwarzschild black hole* once inside the event horizon because

1. There are literally *no paths in spacetime* that go from the interior to the exterior.
2. *All timelike or null paths within the horizon* lead to the singularity at $r = 0$.

But notice the adjective *"classical"* ... More later.

11.3 Eddington–Finkelstein Coordinates

The preceding discussion is illuminating but the interpretation of the results is *complicated by the behavior near the coordinate singularity* at $r = 2M$.

- In this section and the next we discuss two alternative coordinate systems that *remove the coordinate singularity at the horizon*.
  - *Eddington–Finkelstein* coordinates
  - *Kruskal–Szekeres* coordinates
- These coordinate systems have advantages for interpreting the interior behavior of the Schwarzschild geometry.
- However, the standard coordinates remain useful for describing the exterior behavior because of their simple asymptotic behavior.

In the *Eddington–Finkelstein coordinate system* a new variable v is introduced through

$$t = \underbrace{v}_{\text{new}} - r - 2M \ln \left| \frac{r}{2M} - 1 \right|,$$

where the variables

- r, t, and M have their usual meanings in the Schwarzschild metric, and
- θ and φ are unchanged by the transformation.

For either $r > 2M$ or $r < 2M$, insertion into the standard Schwarzschild line element gives the *equivalent line element*

$$ds^2 = -\left(1 - \frac{2M}{r}\right) dv^2 + 2dvdr + r^2 d\theta^2 + r^2 \sin^2\theta d\varphi^2.$$

This is *still Schwarzschild geometry*, but now expressed in *new coordinates*.

- The Schwarzschild metric expressed in these new coordinates is *manifestly non-singular* at $r = 2M$
- The singularity at $r = 0$ remains; it is *physical*.

The singularity at r_S is a *coordinate singularity* that can be removed by *choosing appropriate new coordinates*.

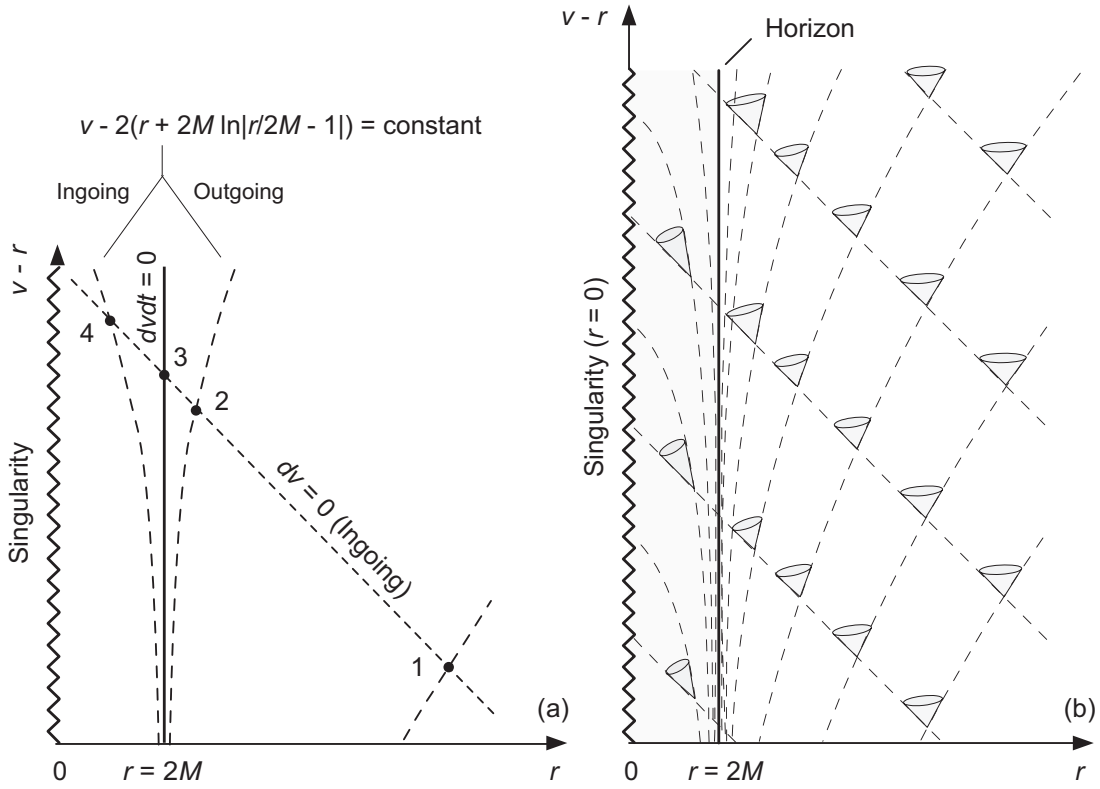


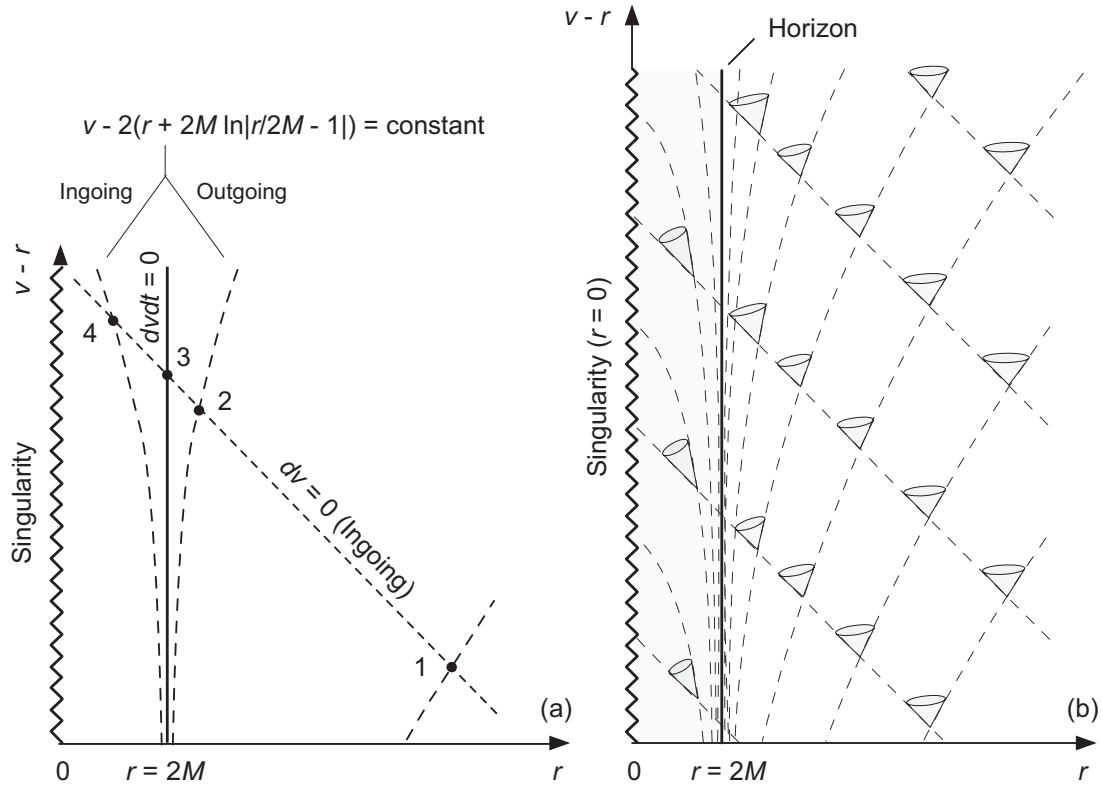
Figure 11.4: (a) Eddington–Finkelstein coordinates for the Schwarzschild black hole with r on the horizontal axis and $v - r$ on the vertical axis. Only two coordinates are plotted, so each point corresponds to a 2-sphere of angular coordinates. (b) Light cones in Eddington–Finkelstein coordinates.

Consider the behavior of radial light rays in these coordinates.

- Set $d\theta = d\varphi = 0$ (radial motion)
- Set $ds^2 = 0$ (light rays).

Then the *Eddington–Finkelstein line element* leads to

$$-\left(1 - \frac{2M}{r}\right) dv^2 + 2dvdr = 0.$$

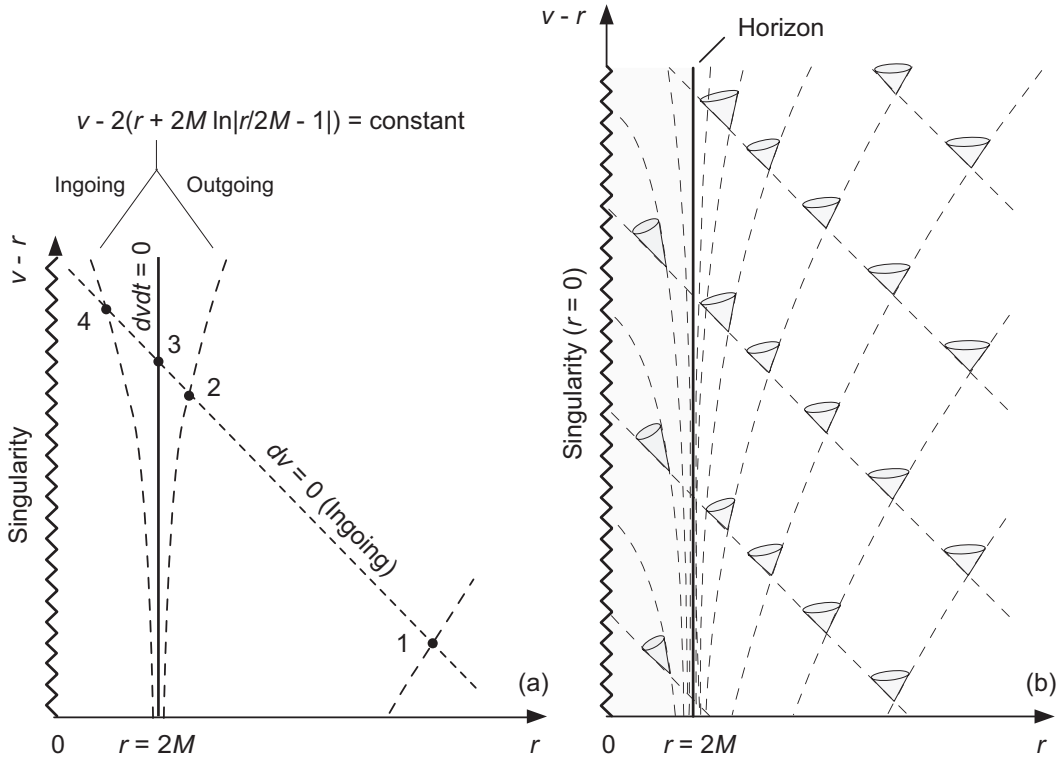


This equation

$$-\left(1-\frac{2M}{r}\right)dv^2+2dvdr=0.$$

has *two general solutions* and *one special solution* [see Fig. (a) above]:

- *General Solution 1:* $dv=0$, so $v=$ constant. $\rightarrow$ *Ingoing light rays* on trajectories of constant v (short-dashed line in Fig. 11.4(a)) .



- *General Solution 2:* If $dv \neq 0$, then divide by dv^2 to give

$$-\left(1-\frac{2M}{r}\right)dv^2 + 2dvdr = 0 \quad \rightarrow \quad \frac{dv}{dr} = 2\left(1-\frac{2M}{r}\right)^{-1},$$

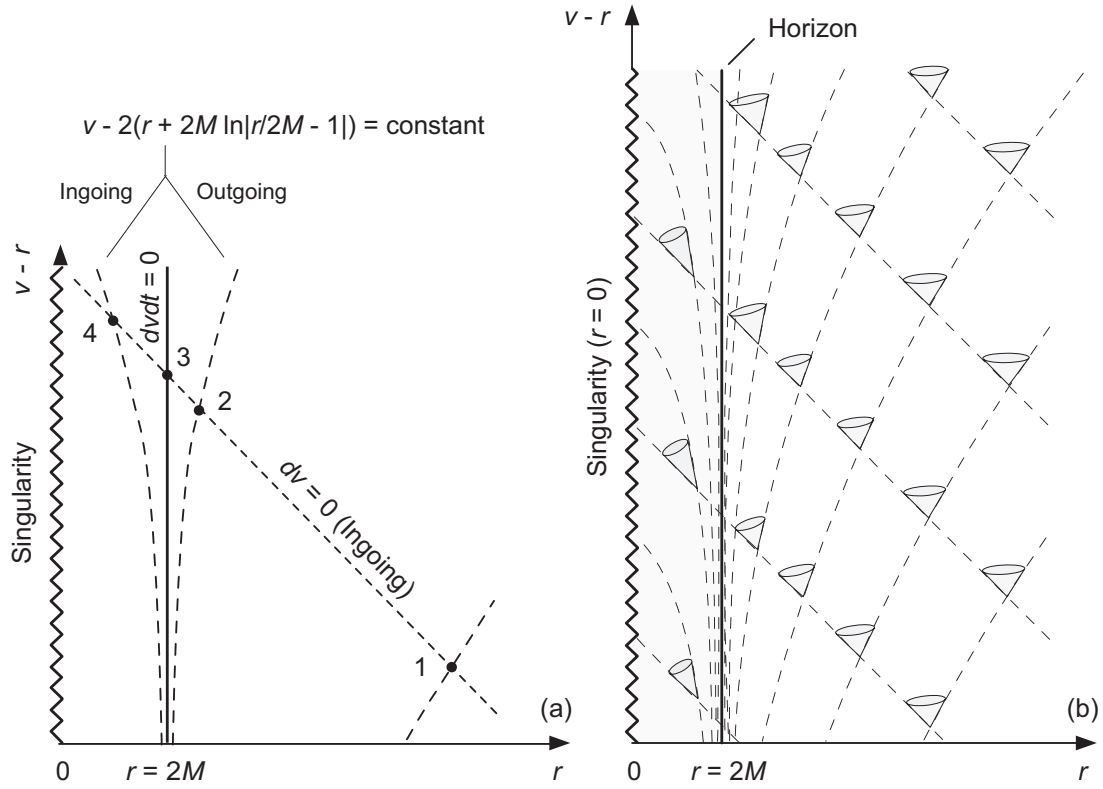
which yields upon integration

$$v - 2\left(r + 2M\ln\left|\frac{r}{2M}-1\right|\right) = \text{constant}.$$

This solution changes behavior at $r = 2M$:

1. *Outgoing* for $r > 2M$.
2. *Ingoing* for $r < 2M$ (r decreases as v increases).

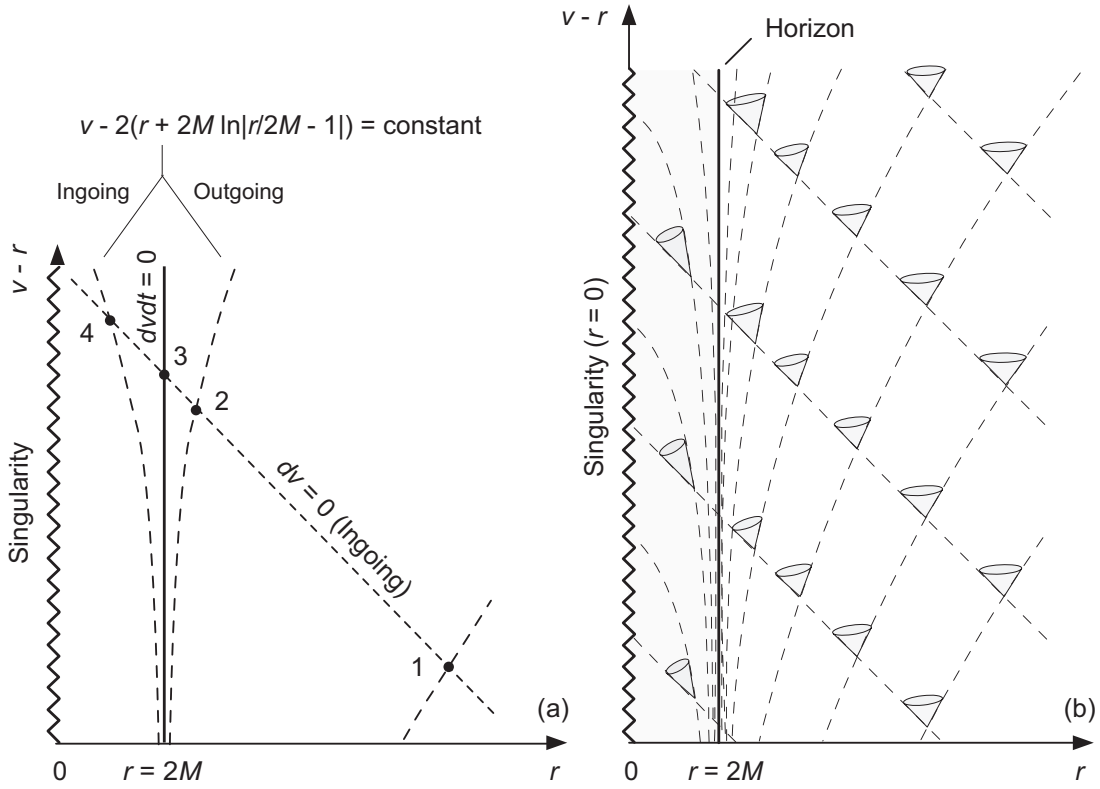
Long-dashed curves in Fig. (a) above illustrate *ingoing and outgoing worldlines* corresponding to this solution.



- *Special Solution:* In the special case that $r = 2M$, the differential equation reduces to

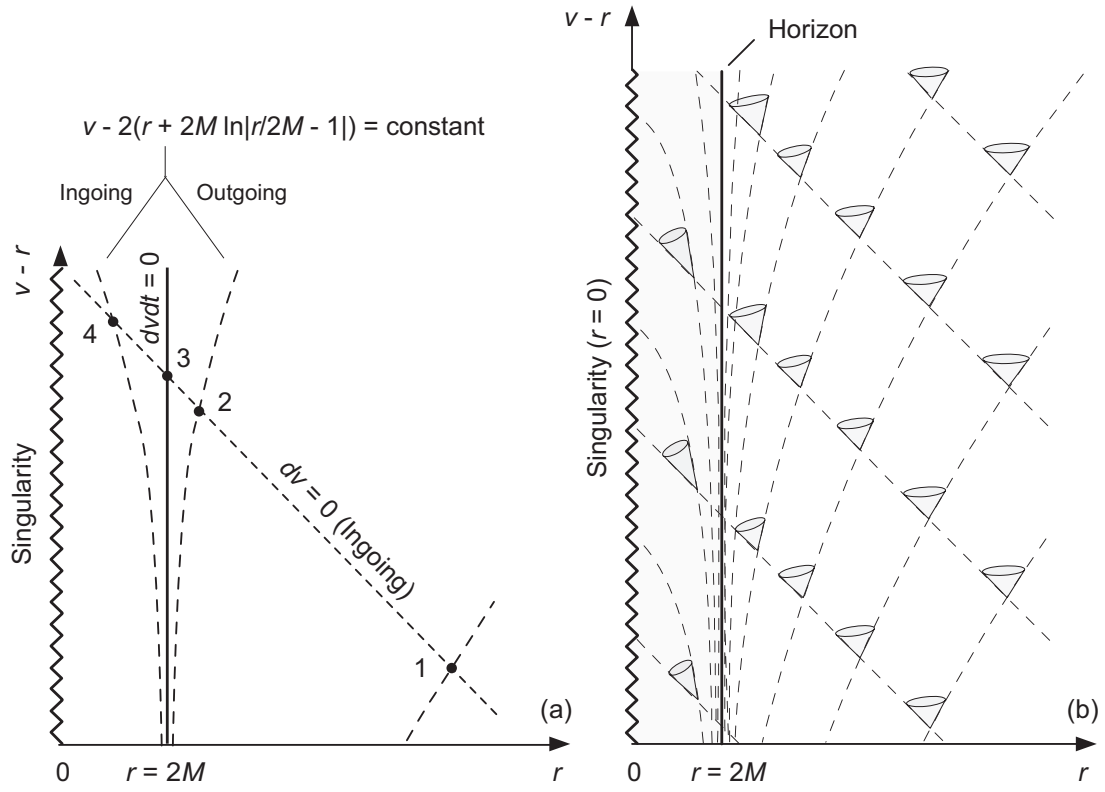
$$-\left(1-\frac{2M}{r}\right)dv^2+2dvdr=0 \quad \rightarrow \quad dvdr=0,$$

which corresponds to *light trapped at the Schwarzschild radius*. The vertical solid line at $r = 2M$ in Fig. (a) above represents this solution.



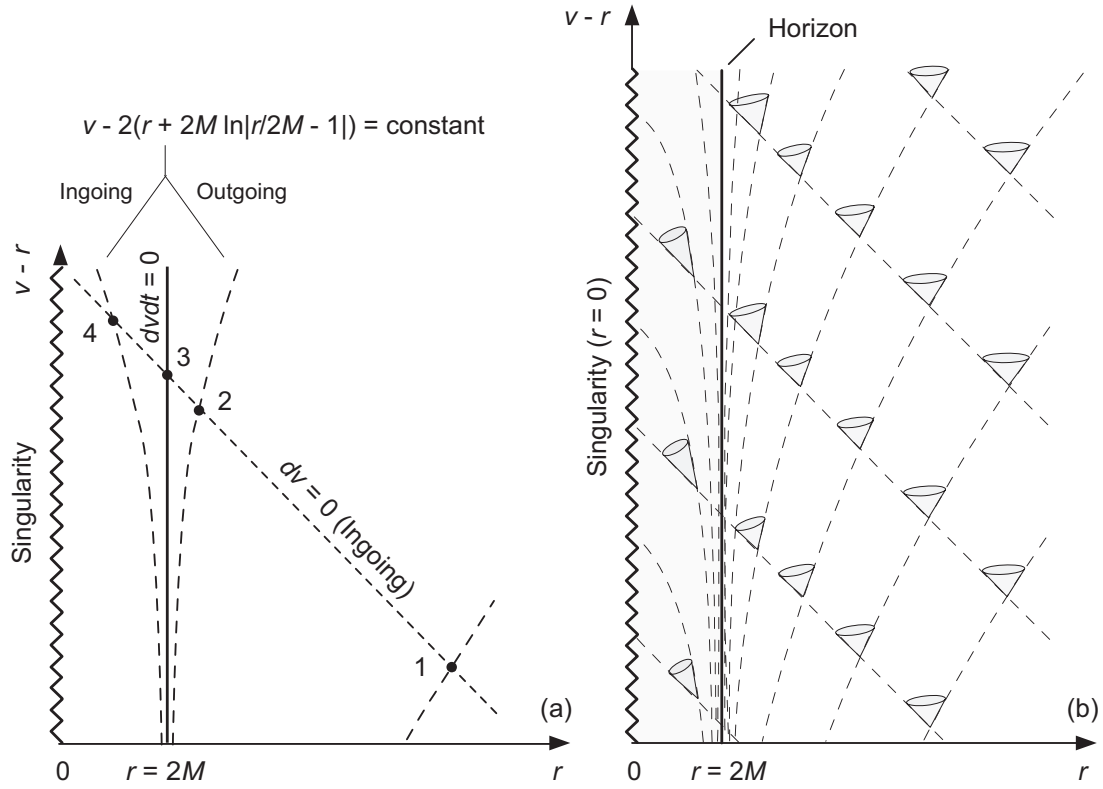
For every spacetime point in Fig. (a) there are *two solutions:*

- For the points labeled 1 and 2 these correspond to
  - one *ingoing* solution and
  - one *outgoing* solution.
- For point 3
  - one solution is *ingoing* and
  - one corresponds to *light trapped at* $r = r_S$.
- For point 4 *both solutions are ingoing*.



The two solutions passing through a point determine the light cone structure at that point [Fig. (b) above].

- The light cones at various points are bounded by the two solutions, so they *tilt "inward" as* **r** *decreases*.
- The radial light ray that defines the left side of the light cone is ingoing (general solution 1).
- If $r \neq 2M$, the radial light ray defining the right side of the light cone corresponds to general solution 2.
    1. These propagate *outward* if $r > 2M$.
    2. For $r < 2M$ they propagate *inward.*



- For $r < 2M$ the light cone is *tilted sufficiently* that *no light ray can escape* the singularity at $r = 0$.
- At $r = 2M$, one light ray moves inward; one is trapped at $r = 2M$.

The horizon may be viewed as a null surface generated by the radial light rays that can neither escape to infinity nor fall in to the singularity.

11.4 Kruskal–Szekeres Coordinates

There is another set of coordinates exhibiting no singularity at $r = 2M$: *Kruskal–Szekeres coordinates.*

- Introduce variables (v, u, θ, φ), where θ and φ have their usual meaning and new variables u and v are defined by

$$\begin{aligned} u &= \left(\frac{r}{2M} - 1\right)^{1/2} e^{r/4M} \cosh\left(\frac{t}{4M}\right) \qquad (r > 2M) \\ &= \left(1 - \frac{r}{2M}\right)^{1/2} e^{r/4M} \sinh\left(\frac{t}{4M}\right) \qquad (r < 2M) \\ v &= \left(\frac{r}{2M} - 1\right)^{1/2} e^{r/4M} \sinh\left(\frac{t}{4M}\right) \qquad (r > 2M) \\ &= \left(1 - \frac{r}{2M}\right)^{1/2} e^{r/4M} \cosh\left(\frac{t}{4M}\right) \qquad (r < 2M) \end{aligned}$$

- The corresponding line element is

$$ds^2 = \frac{32M^3}{r} e^{-r/2M}(-dv^2 + du^2) + r^2 d\theta^2 + r^2 \sin^2\theta d\varphi^2,$$

  where $r = r(u, v)$ is defined through

$$\left(\frac{r}{2M} - 1\right) e^{r/2M} = u^2 - v^2.$$

- This metric is *manifestly non-singular at* $r = 2M$, but singular at $r = 0$.

> This is *still Schwarzschild spacetime*, but now expressed in a new set of coordinates.

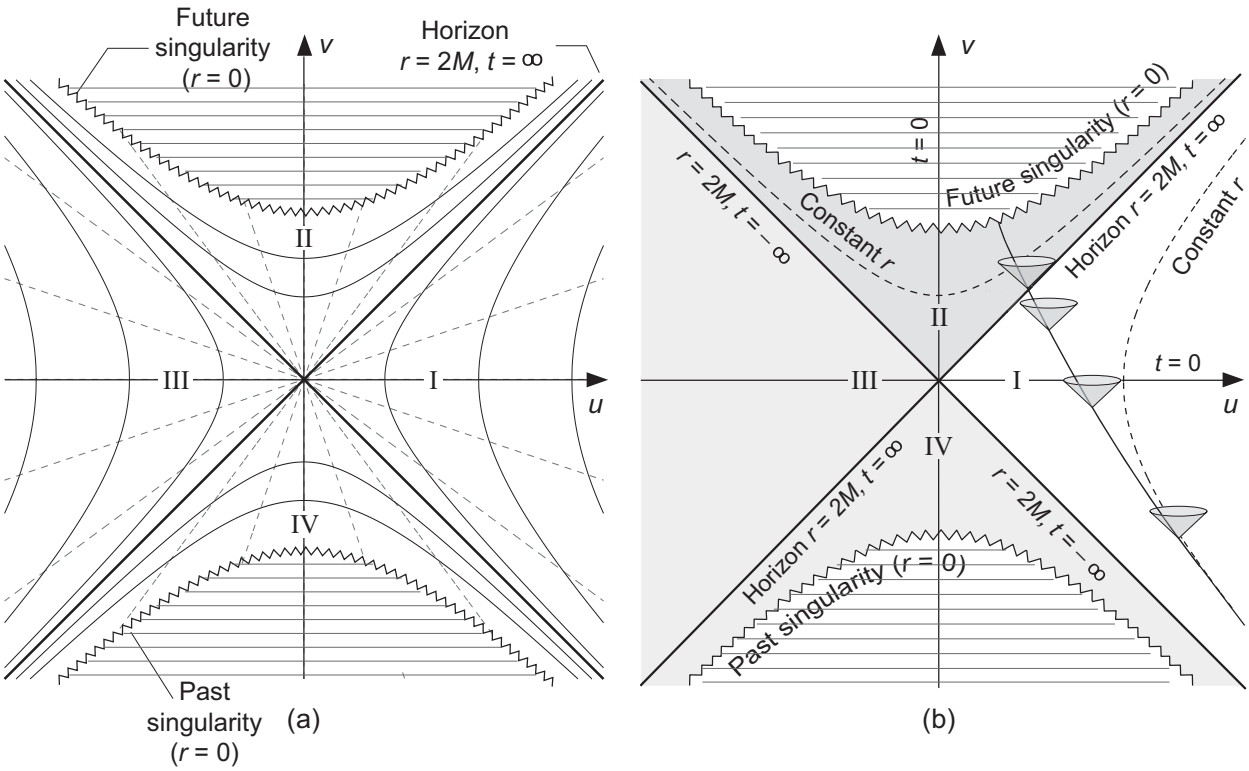
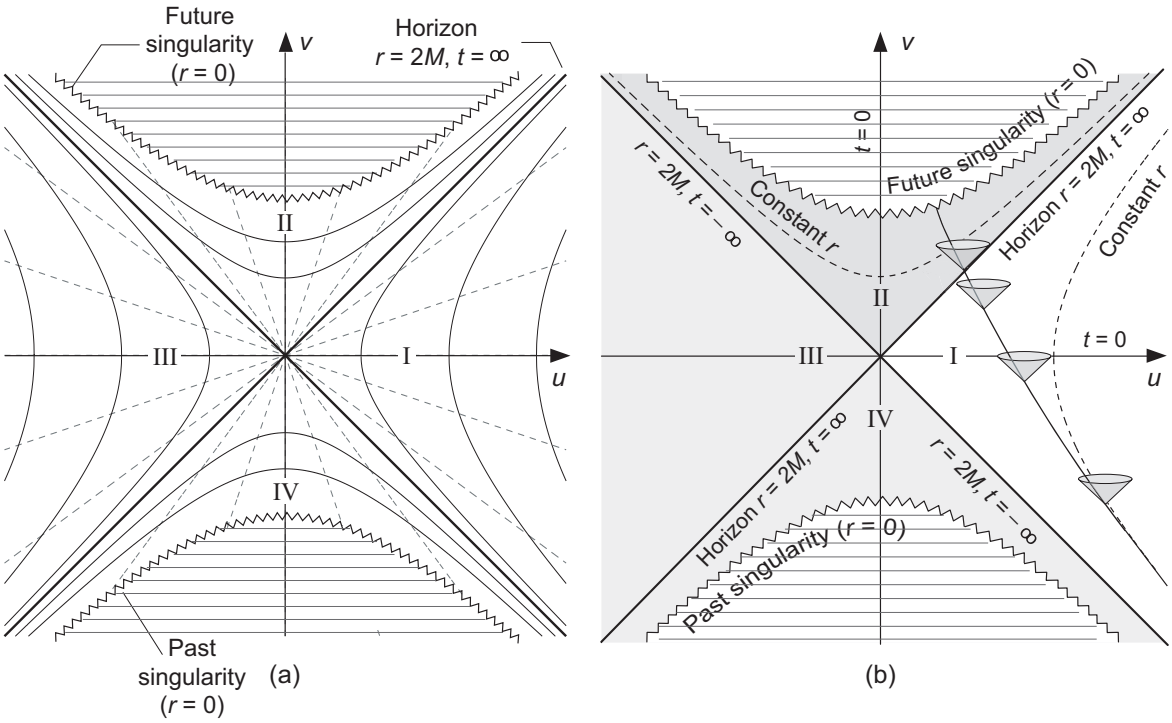


Figure 11.5: (a) Schwarzschild spacetime in Kruskal–Szekeres coordinates. Only the two coordinates u and v are displayed, so each point is really a 2-sphere corresponding to the variables θ and φ. Spacetime singularities are indicated by jagged curves. The hatched regions above and below the $r = 0$ singularities are not a part of the spacetime. Curves of constant r are hyperbolas and the dashed straight lines are lines of constant t. (b) Worldline of a particle falling into a Schwarzschild black hole in Kruskal–Szekeres coordinates.

Kruskal diagram: lines of constant r and t plotted on a u and v grid. Figure 11.5 illustrates.



- From the form of

$$\left(\frac{r}{2M}-1\right)e^{r/2M}=u^2-v^2.$$

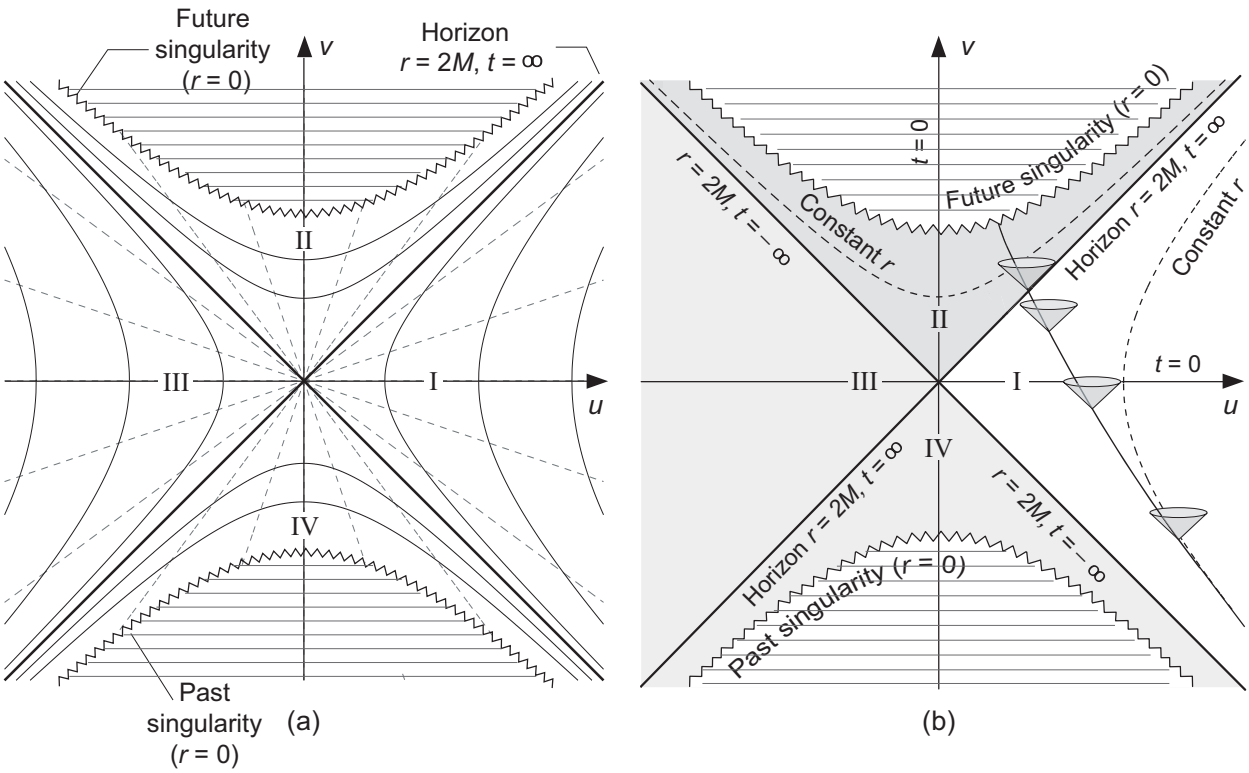
lines of constant r *are hyperbolae* of constant u^2-v^2.

- From the definitions of u and v

$$v=u\tanh\left(\frac{t}{4M}\right) \qquad (r>2M)$$
$$=\frac{u}{\tanh\left(t/4M\right)} \qquad (r<2M).$$

Thus, *lines of constant* t *are straight lines* with slope

- $\tanh(t/4M)$ for $r>2M$
- $1/\tanh(t/4M)$ for $r<2M$.

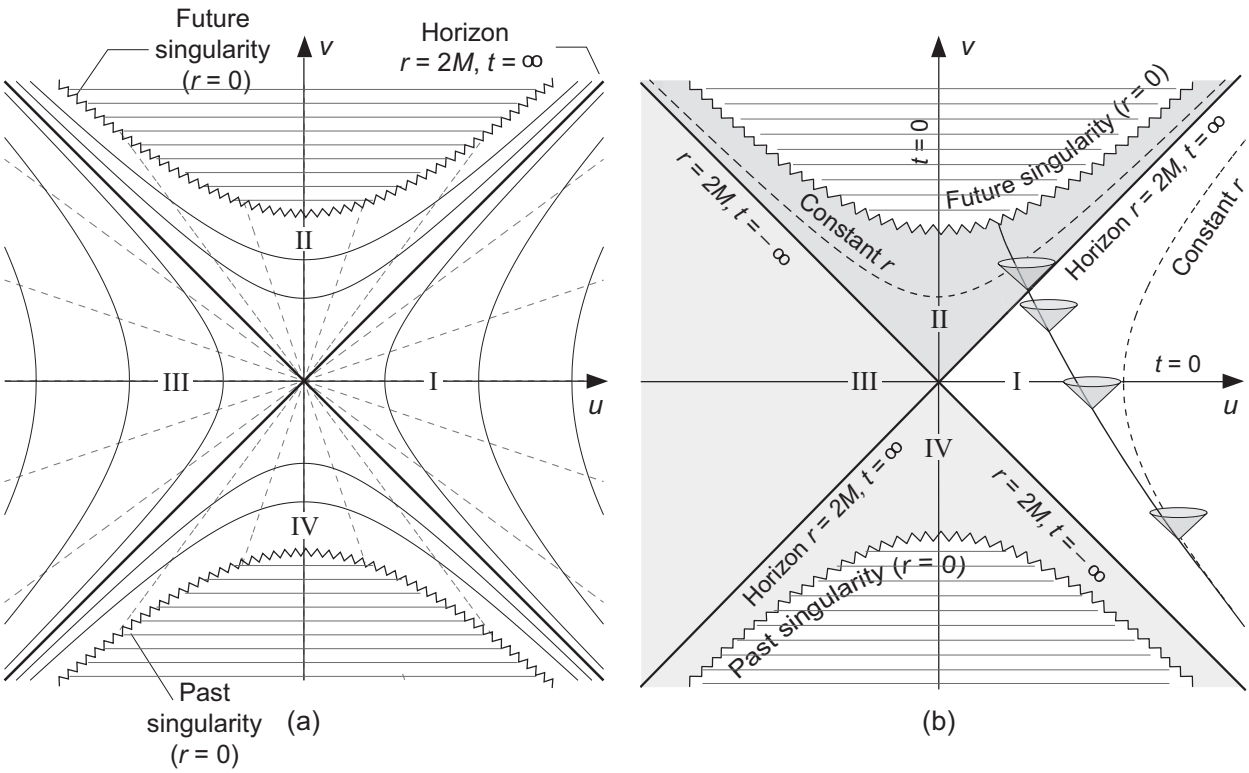


- For radial light rays in Kruskal–Szekeres coordinates ($d\theta = d\varphi = ds^2 = 0$), and the line element

$$ds^2 = \frac{32M^3}{r} e^{-r/2M}(-dv^2 + du^2) + r^2 d\theta^2 + r^2 \sin^2\theta d\varphi^2$$

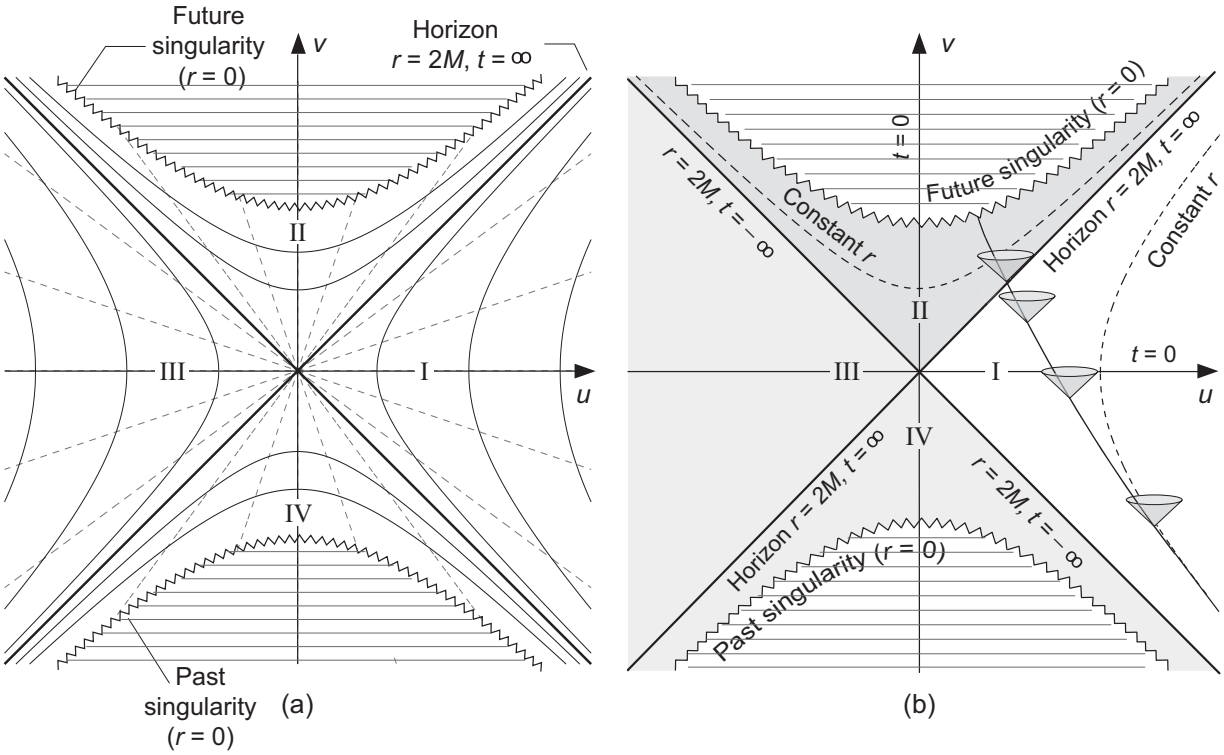
yields $dv = \pm du$:

> Lightcones in the Kruskal–Szekeres coordinates *always open at 45-degree angles*, like in Minkowski space.



- Over the full range of coordinates (v, u, θ, φ),
  - the metric component $g_{00} = g_{vv}$ *remains negative* and
  - $g_{11} = g_{uu}$, $g_{22} = g_{\theta\theta}$, and $g_{33} = g_{\varphi\varphi}$ *remain positive*.
- Therefore, in Kruskal–Szekeres coordinates
  - the v direction is *always timelike* and
  - the u direction is *always spacelike*,

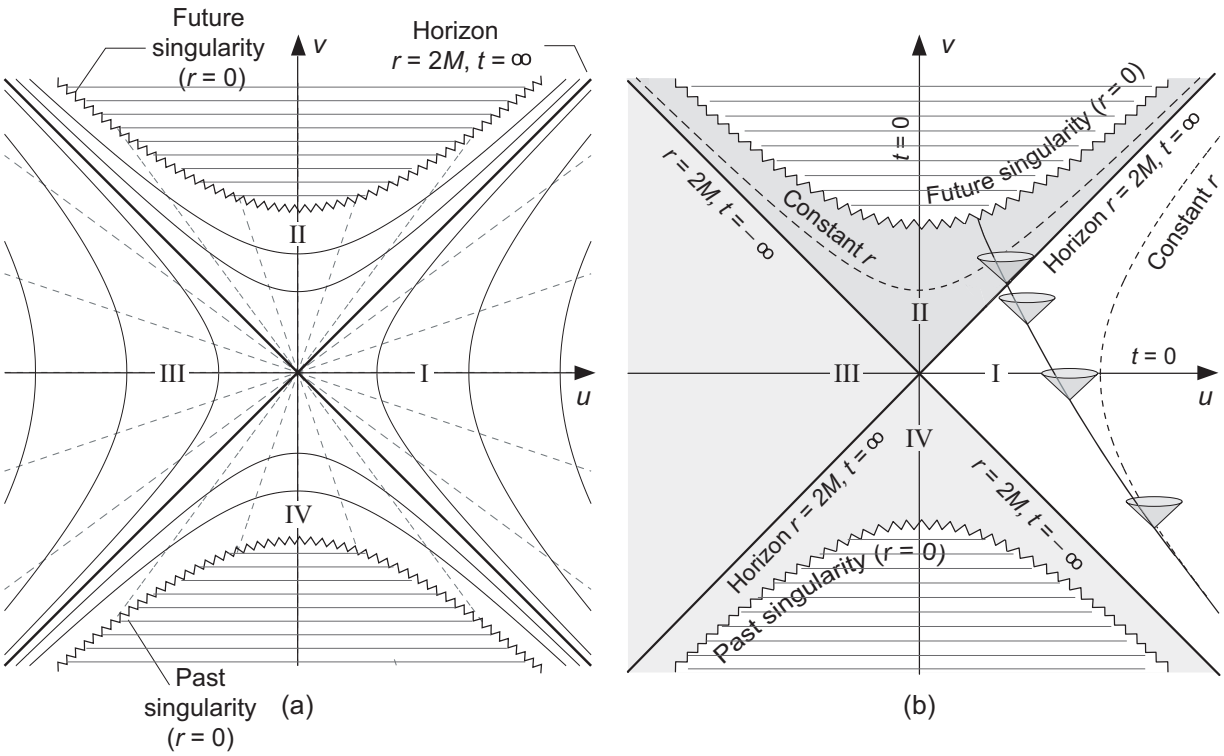
  in contrast to the normal Schwarzschild coordinates where r and t *switch their character at the horizon*.



Lines corresponding to $r = 2M$ separate spacetime into *four quadrants*, labeled I, II, III, and IV. See the figure above.

- In *quadrant I*, $r > 2M$ so this is the Schwarzschild spacetime *outside the horizon*.
- In *quadrant II*, $r < 2M$ so
  - this is the *black hole interior to the horizon* and
  - worldlines there must encounter the $r = 0$ singularity in their future since *the singularity is spacelike*.
- Thus regions I and II are of *direct physical interest*.

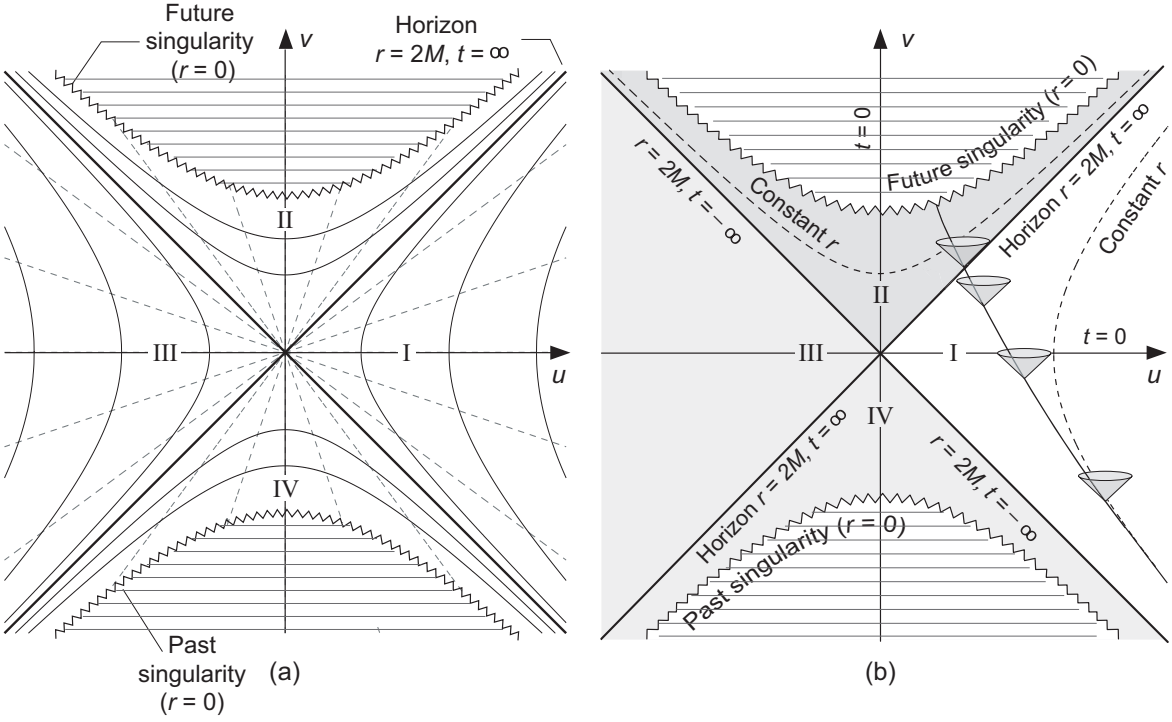
Regions III and IV are of interest mathematically but it is *unclear if they have physical implications*.



Region IV is equivalent to region II but with time inverted and the *singularity in the past*.

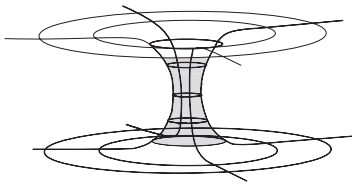
- This is called a *white hole,* which is the opposite of a black hole:
  - Matter *spews into the Universe* from a singularity rather than disappearing into a singularity, and
  - The *horizon forbids entry* rather than exit.
- There is *no astrophysical evidence* for white holes.

Region III is an exterior region (outside the horizon) as for region I, but it corresponds to a *different asymptotically flat spacetime* than region I.



Move along the horizontal axis ($t = 0$) in Fig. (a), beginning in region I at $u = +\infty$ and continuing to $u = -\infty$ in region III.

- The coordinate r is marked by timelike hyperbolas cutting the axis, so r decreases to a minimum $r = 2M$ at the origin and then increases again after passing through the origin.
- In terms of an embedding diagram, a geometry similar to



is obtained, with a narrow throat *connecting two different asymptotically flat spacetimes*.

This is called an *Einstein–Rosen bridge* (or a *wormhole*).

The Einstein–Rosen bridge

- is *interesting mathematically* but
- is *likely not relevant for physics* of the Schwarzschild metric, for reasons discussed more extensively in the book chapter.
- Nevertheless, the possibility of *forming wormholes under more exotic circumstances* has received substantial attention.

Our presentation will henceforth concentrate on the *physically interesting regions I and II* of the Schwarzschild spacetime.

How are the different coordinate systems we have used to parameterize the Schwarzschild metric related mathematically?

- A manifold is *geodesically complete* if every geodesic beginning from an arbitrary point can be extended to infinite values of the affine parameter in both directions.
- A manifold is said to be *maximal* if every geodesic
  - can be *extended infinitely* in both directions, or
  - *terminates* in an intrinsic singularity.
- Hence a geodesically complete manifold is also maximal, but the converse need not be true.
- Minkowski space is *geodesically complete* and *maximal*.
- The Schwarzschild spacetime parameterized with Schwarzschild coordinates or with Eddington–Finkelstein coordinates *is not maximal*.
- The Kruskal parameterization is the *unique maximal extension* of Schwarzschild spacetime.
- However, The Kruskal extension is
  - *not geodesically complete* because
  - it *contains intrinsic singularities*.

See also the later discussion of *trapped surfaces* in black hole spacetimes.

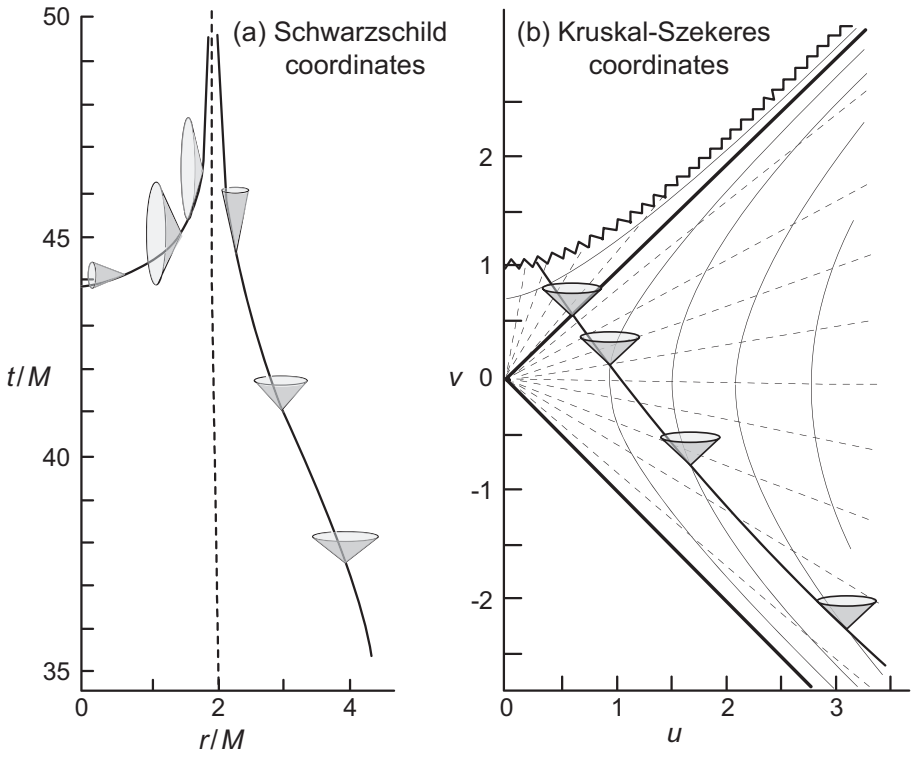
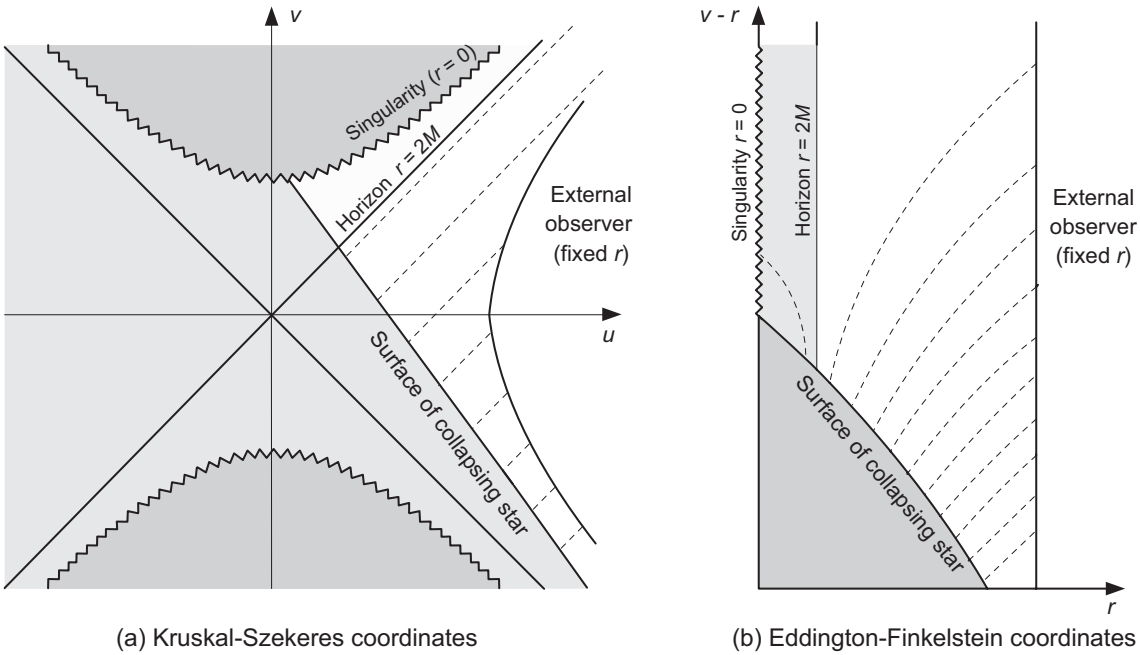


Figure 11.6: A trip to the center of a black hole in standard Schwarzschild coordinates and in Kruskal–Szekeres coordinates.

The identification of $r = 2M$ as an *event horizon of the Schwarzschild spacetime* is particularly clear in Kruskal–Szekeres coordinates [Fig. 11.6(b)].

- All lightcones and the horizon make *45-degree angles with the vertical*.
- Thus, for any point within the horizon, its
  - forward worldline *must contain* the $r = 0$ singularity
  - and *cannot contain* the $r = 2M$ horizon.



(a) Kruskal-Szekeres coordinates (b) Eddington-Finkelstein coordinates

Figure 11.7: Collapse to a Schwarzschild black hole.

Fig. 11.7 illustrates a *spherical mass collapsing to a black hole* in Kruskal–Szekeres and Eddington–Finkelstein coordinates.

- A distant observer at fixed r observes light signals sent periodically from the surface of the collapsing star.
- Light pulses, propagating on the dashed lines, arrive at increasingly longer intervals for the outside observer.
- At the horizon, light signals take an infinite length of time to reach the external observer.
- Once the surface is inside the horizon, no signals can reach the outside observer as the star collapses to a singularity.
- *Note:* the Schwarzschild solution is *valid outside the star*. Inside GR applies but the *solution is not Schwarzschild*.

11.5 Black Hole Theorems and Conjectures

In this section we summarize (in a non-rigorous way) a set of *theorems and conjectures concerning black holes*. Some we have already used in various contexts.

- *Singularity theorems:* Loosely, any gravitational collapse that proceeds far enough results in a *spacetime singularity*.
- *Cosmic censorship conjecture:* All spacetime singularities are hidden by event horizons (*no naked singularities*).
- *(Classical) area increase theorem:* In all classical processes involving horizons, the *area of the horizons can never decrease*.
- *Second law of black hole thermodynamics:* Where quantum mechanics is important the classical area increase theorem is replaced by

  1. The entropy of a black hole is proportional to the surface area of its horizon.
  2. The total entropy of the Universe can never decrease an any process.

- *The no-hair theorem/conjecture:* If gravitational collapse to a black hole is nearly spherical,
  - All non-spherical parts of the mass distribution (quadrupole moments, ...) except angular momentum are *radiated away as gravitational waves*.
  - Horizons eventually become stationary.
  - A stationary black hole is characterized by:
    - the mass M,
    - the angular momentum J, and
    - the charge Q.
  - M, J, and Q are all determined by fields *outside* the horizon, not by integrals over the interior.

> *"No Hair Theorem":* black holes destroy all detail (*the hair*) about the matter that formed them, leaving only *global mass, angular momentum, and charge* as observable external characteristics.

The most general solution characterized by M, J, and Q is termed a *Kerr–Newman black hole.* However,

- The astrophysical processes that could form a black hole would likely *neutralize any excess charge*.
- Thus *astrophysical black holes are Kerr black holes* (characterized by M and J (the Schwarzschild solution being a special case of the Kerr solution for vanishing J).

- Birkhoff's theorem:
  - The Schwarzschild solution is the *only spherically symmetric solution* of the vacuum Einstein equations.
  - The assumptions that we made of
    - no time dependence and
    - spherical symmetry

    in deriving the Schwarzschild solution are in fact *not independent*.
  - Even if the source is changing with time, the *Schwarzschild metric remains the only spherically symmetric solution* of the vacuum Einstein equations.

- These theorems and conjectures place the *mathematics of black holes* on reasonably firm ground.
- To place the *physics of black holes* on firm ground, these ideas must be tested by observation, which we take up in later chapters.

Chapter 12

Quantum Black Holes

Classically, the fundamental structure of curved spacetime ensures that nothing can escape from within the Schwarzschild event horizon.

- That is an emphatically *deterministic* statement.
- But what about quantum mechanics, which is fundamentally *indeterminate*?

The uncertainty principle and quantum fluctuations of the vacuum play a central role in quantum mechanics.

- As we now explain, because of quantum mechanics it is possible for a black hole to emit mass.
- Therefore, as Stephen Hawking discovered, from a quantum point of view *black holes are not really black!*

12.1 Geodesics and Quantum Uncertainty

- Our discussion to this point has been *classical* in that it assumes that free particles follow *geodesics appropriate for the spacetime*.
- But the uncertainty principle implies that
  1. Microscopic particles *cannot be completely localized on classical trajectories* because they are subject to an uncertainty of the form
  $$\Delta p_i \Delta x_i \geq \hbar$$
     - where Δp_i is the uncertainty in momentum and
     - Δx_i is the uncertainty in position.
  2. *Neither can energy conservation be imposed* except with an uncertainty
  $$\Delta E \, \Delta t \geq \hbar,$$
     - where ΔE is an energy uncertainty and
     - Δt is the corresponding time period during which this energy uncertainty exists.

This implies an *inherent quantum fuzziness* in the 4-momenta associated with our description of spacetime at the quantum level.

For the *Killing vector*

$$K \equiv K_t = (1,0,0,0)$$

in the *Schwarzschild metric*,

$$K \cdot K = g_{\mu\nu} K^\mu K^\nu = g_{00} K^0 K^0 = -\left(1 - \frac{2M}{r}\right),$$

from which we conclude that

K is $\begin{cases} \textit{Timelike} \text{ outside the horizon, since then } K \cdot K < 0, \\ \textit{Spacelike} \text{ inside the horizon, since then } K \cdot K > 0. \end{cases}$

As we now make plausible, this property of the Killing vector K

- permits a virtual quantum fluctuation of the vacuum to be converted into real particles and
- these are detectable at infinity as *emission of mass from the black hole*.

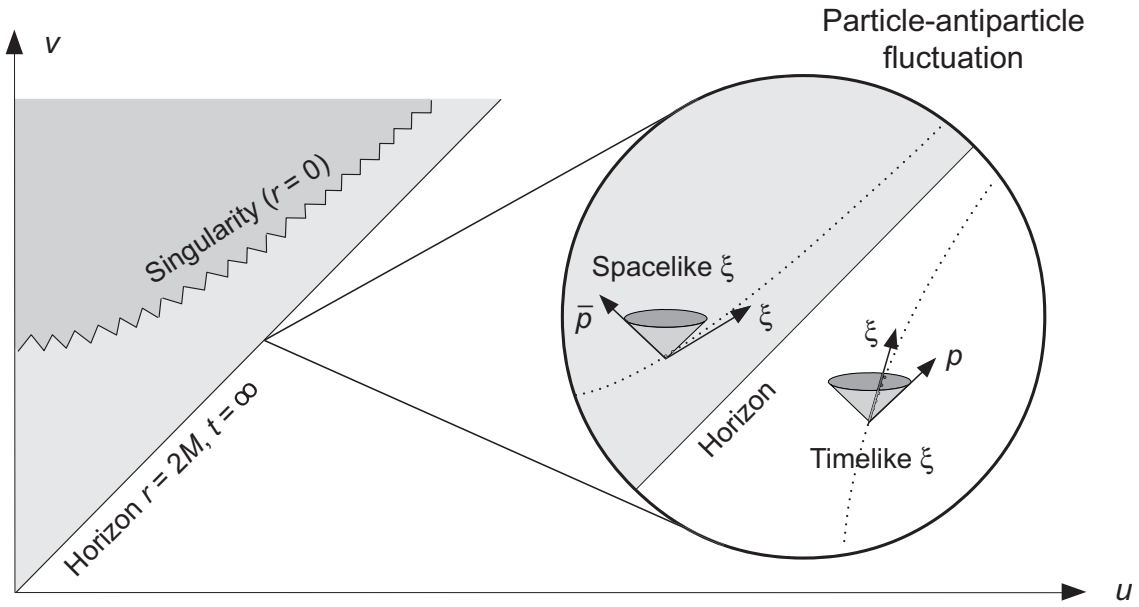
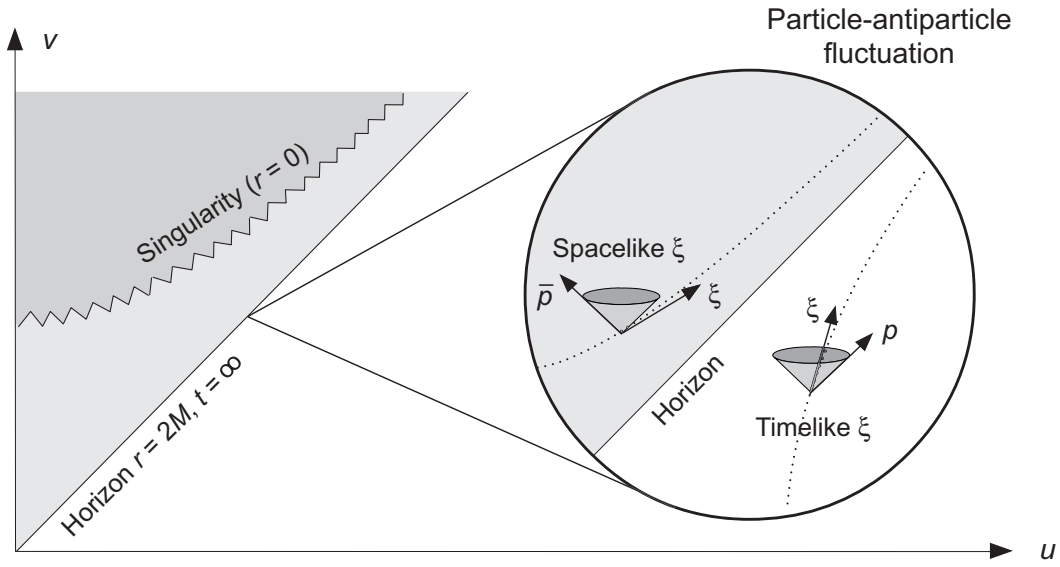


Figure 12.1: Hawking radiation in Kruskal–Szekeres coordinates.

12.2 Hawking Radiation

Assume a *particle–antiparticle pair* created by vacuum fluctuation near the horizon of a black hole, such that the particle and antiparticle end up on opposite sides of the horizon (Fig. 12.1).

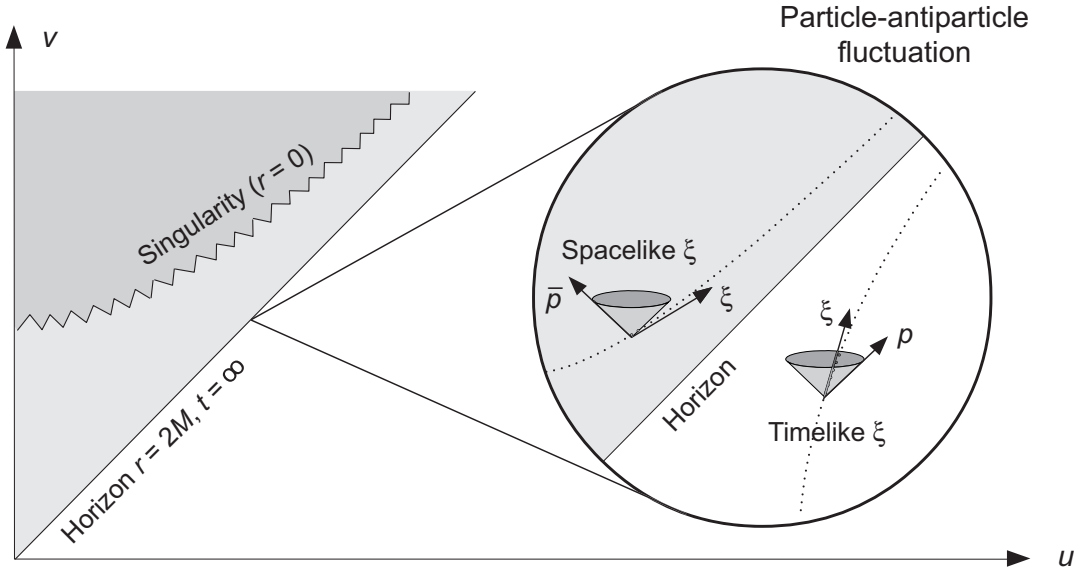
- If the particle–antiparticle pair is created in a small enough region of spacetime, there is nothing special implied by this region lying at the event horizon:
- *the local spacetime is indistinguishable from Minkowski space* because of the *equivalence principle*.
- Therefore, the *normal principles of (special) relativistic quantum field theory will be applicable* to the pair creation process in a local inertial frame at the horizon.



- The conserved quantity analogous to total energy is the scalar product of $K = (1,0,0,0)$ with the 4-momentum p.
- Therefore, if a particle–antiparticle pair is produced near the horizon with 4-momenta p and $\bar{p}$, respectively,

$$K \cdot p + K \cdot \bar{p} = 0,$$

  must be satisfied (to preserve vacuum quantum numbers).
- *For the particle outside the horizon*, $-K \cdot p > 0$ (it is proportional to an energy that is measureable externally).
- *If the antiparticle were also outside the horizon,* it too *must* have $-K \cdot \bar{p} > 0$, in which case
  1. The condition $K \cdot p + K \cdot \bar{p} = 0$, *cannot be satisfied* and
  2. The particle–antiparticle pair can have only a *fleeting existence* of duration $\Delta t \sim \hbar / \Delta E$ (Heisenberg).
- So far, no surprises; just Quantum Mechanics 101 ...



HOWEVER ...

If instead (as in the figure) the antiparticle is *inside the horizon,*

- The Killing vector K is *spacelike* ($K \cdot K > 0$)
- The scalar product $-K \cdot \bar{p}$ is *not an energy*. In fact,
- the product $-K \cdot \bar{p}$ can be a *3-momentum component*.
- Therefore $-K \cdot \bar{p}$ can be *positive or negative (!!).*

Thus, there is magic afoot:

- If $-K \cdot \bar{p}$ is negative $K \cdot p + K \cdot \bar{p} = 0$ *can be satisfied*.
- Then the virtual particle created outside the horizon can propagate to infinity as a *real, detectable particle,* while
- the antiparticle remains trapped inside the event horizon.
- The black hole *emits its mass as a stream of particles (!).*

Therefore, quantum effects imply that a black hole

- can emit its mass as a flux of particles and antiparticles created through vacuum fluctuations.
- The emitted particles are termed *Hawking radiation*.
- Therefore, black holes are *not really black*!

The loss of mass by Hawking radiation for a black hole is called *black hole evaporation*.

It should be noted that the actual theory underlying these ideas is more subtle than the cartoon presented here. However, the cartoon makes the basic idea plausible.

12.3 Mass Emission Rates and Black Hole Temperature

Methods for obtaining results from the Hawking theory are beyond our scope, but the results can be stated simply:

- The distribution of energies emitted by the black hole is
  - equivalent to that of a *blackbody* with a
  - *temperature proportional to the surface gravity* of the black hole.
- (The *surface gravity* is a renormalized acceleration that would be experienced by an observer hear the horizon.)
- Specifically, the *black hole temperature* is

$$T = \frac{\kappa}{2\pi} = \frac{\hbar}{8\pi k_B M} = \frac{\hbar c^3}{8\pi k_B G M}$$
$$= 6.2 \times 10^{-8} \left(\frac{M_\odot}{M}\right) \text{K}.$$

  where $\kappa = (4M)^{-1}$ is the *surface gravity* of a Schwarzschild black hole.

Notice that the *black hole temperature formula*

$$T = \frac{\hbar c^3}{8\pi k_{\mathrm{B}} G M}$$

combines in a single equation fundamental constants associated with

- special relativity (c)
- gravity (G)
- quantum mechanics ($\hbar$) and
- statistical mechanics (k_{B}).

Furthermore, since $T \propto \hbar$,

- Hawking radiation is a *quantum effect*
- that vanishes in the classical limit $\hbar \to 0$.

As shown in a Problem,

- the *radiated power* is given by

$$P = \frac{\hbar c^6}{15360\pi G^2 M^2},$$

- from which the *rate of mass emission* is

$$\frac{dM}{dt} = -\lambda \frac{\hbar}{M^2},$$

  where λ is a dimensionless constant.

- Integrating $dM/dt = -\lambda\hbar/M^2$ for a black hole assumed to emit all of its mass by Hawking radiation in a time $t_{\rm H}$, we obtain

$$M(t) = (3\lambda\hbar(t_{\rm H} - t))^{1/3}.$$

  for its *mass as a function of time*.

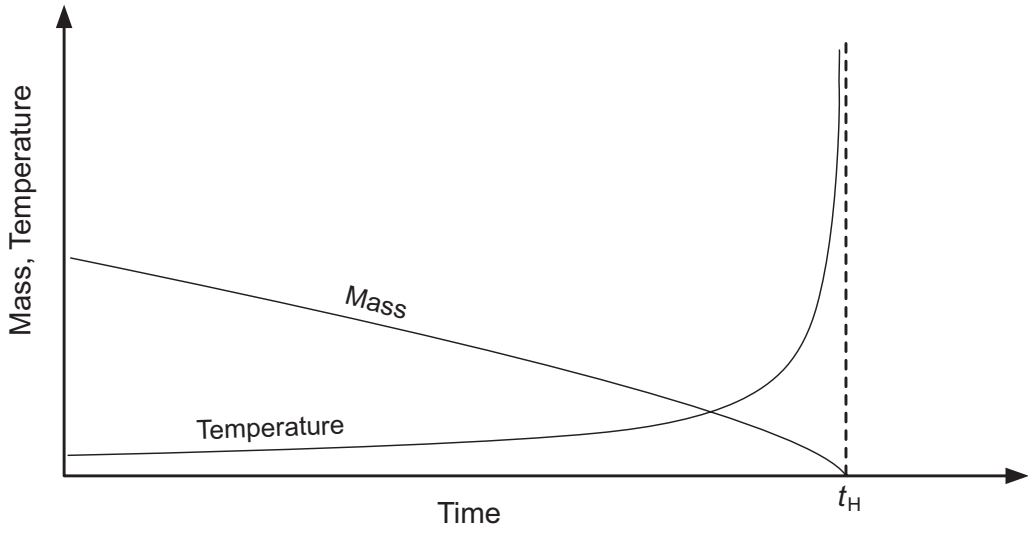


Figure 12.2: Evaporation of a Hawking black hole.

From the results

$$T = \frac{\hbar}{8\pi k_B M} \qquad \frac{dM}{dt} = -\lambda \frac{\hbar}{M^2} \qquad M(t) = (3\lambda\hbar(t_H - t))^{1/3}$$

the mass and temperature behave as in Fig. 12.2. Therefore,

- The black hole *evaporates at an accelerating rate* as
- temperature and the emission rate tend to infinity.
- We may expect a *final burst* of very *high energy radiation*.

Note that for the black hole (as for normal stars)

- temperature increases as
- energy (mass) is emitted!

The black hole has a *negative heat capacity.*

12.4 Miniature Black Holes

How long does it take a Hawking black hole to evaporate?

- From the mass emission formula

$$M(t) = (3\lambda\hbar(t_{\rm H} - t))^{1/3},$$

- we may estimate a *lifetime for complete evaporation* as

$$\begin{aligned} t_{\rm H} &\simeq \frac{M^3}{3\lambda\hbar} = \frac{5120\pi G^2}{\hbar c^4} M^3 \\ &\simeq 6.6 \times 10^{74} \left(\frac{M}{M_\odot}\right)^3 \text{ s}, \end{aligned}$$

  where a *blackbody approximation* of

$$\lambda = (15,360\pi)^{-1}$$

  has been used.

Take 15 $M_\odot$ as a representative mass for a black hole formed from massive star collapse.

- The *temperature* is then

$$T = \frac{\hbar c^3}{8\pi k_B G M} = 6.2 \times 10^{-8} \left(\frac{M_\odot}{M}\right) \text{K} = 4.1 \times 10^{-9}\,\text{K},$$

- the *radiated power* is

$$P = \frac{\hbar c^6}{15360\pi G^2 M^2} = 9.0 \times 10^{-29} \left(\frac{M_\odot}{M}\right)^2 \text{W} = 4 \times 10^{-31}\,\text{W},$$

- and the *time to evaporate* by emitting Hawking radiation is given by

$$t_H \simeq \frac{5120\pi G^2}{\hbar c^4} M^3 \simeq 6.6 \times 10^{74} \left(\frac{M}{M_\odot}\right)^3 \text{s} = 2.2 \times 10^{78}\,\text{s}.$$

- This is $\sim 10^{60}$ *times larger than the age of the Universe!*

Hawking radiation *may be ignored for stellar black holes*.

> The temperature of stellar black holes is so low that the surrounding *cosmic microwave background* is much hotter, so stellar black holes *absorb rather than emit radiation*.

- However, black holes of initial mass $\sim 10^{14}$ g would have lifetimes *comparable to the age of the Universe*.
- Their demise might be detectable through a characteristic *burst of high-energy radiation*.
- The *Schwarzschild radius* for a 10^{14} g black hole is

$$r_S \sim 1.5 \times 10^{-14}\,\text{cm},$$

  which is about $\frac{1}{5}$ the size of a proton.
- To form such a black hole we must compress 10^{14} grams (a large mountain) into the size of a proton!
- The *early big bang* could have produced such densities.
- Therefore, a population of *miniature black holes*
  - could have *formed in the big bang* and
  - could be *decaying in the present Universe*

  with a detectable signature.

No evidence has been found for miniature black holes and their associated Hawking radiation.

- For a 10^{14} g black hole, $P \sim 10^{10}$ W initially.
- This would grow much larger in the final burst of gamma-rays, but the burst *would have to be nearby* to be seen.

12.5 Black Hole Thermodynamics

The preceding results suggest that

- the *gravitational physics of black holes* and
- *classical thermodynamics*

are *closely related*. Remarkably, this has turned out to be correct.

- It had been noted prior to Hawking's discovery by Jacob Bekenstein that there were *similarities between black holes and blackbody radiators*.
- However, the difficulty with describing a black hole in thermodyanamical terms was that
  - a classical black hole *permits no equilibrium with the surroundings* because
  - it *absorbs but cannot emit* radiation.

Hawking radiation supplies the necessary equilibrium that ultimately allows thermodynamics and a temperature to be ascribed to a black hole.

Hawking Area Theorem: Hawking has proven a theorem that

> The total horizon area A *cannot decrease in any physical process* involving black hole horizons,
>
> $$\frac{dA}{dt} \geq 0.$$

- For a Schwarzschild black hole, the horizon area is

$$A = 16\pi M^2 \quad \rightarrow \quad \frac{dA}{dM} = 32\pi M,$$

which can be written (h = Planck, k_B = Boltzmann).

$$dM = \frac{h}{8\pi k_B M} d\left(\frac{k_B A}{4h}\right)$$

- But $dE = dM$ is the change in total energy and the temperature of the black hole is $T = \hbar/8\pi k_B M$.
- Therefore, we may write the preceding as

$$\underbrace{dE = T\,dS}_{\text{1st Law}} \qquad \underbrace{dS \geq 0}_{\text{2nd Law}} \qquad \underbrace{S \equiv \frac{k_B}{4h} A}_{\text{entropy}}$$

which are just the *1st and 2nd laws of thermodynamics*, provided that

$$S \propto \text{(surface area of black hole)}$$

is *interpreted as entropy!*

Evaporation of a black hole through Hawking radiation manifestly *appears to violate the Hawking area theorem*.

- However, the area theorem is based on energy assumptions that are thought to be correct at the classical level but may break down in quantum processes.
- The correct quantum interpretation of Hawking radiation and the area theorem is that
  1. The entropy of the evaporating, isolated black hole *decreases* with time (because it is proportional to the area of the horizon).
  2. The total entropy of the Universe *increases* because of the entropy associated with the Hawking radiation itself.
  3. That is, the area theorem is replaced by the

> *Generalized 2nd Law:* The total entropy of the *black hole plus exterior Universe* may never decrease in any physical process.

12.6 The Four Laws of Black Hole Dynamics

We may formulate *four laws of black hole dynamics* that are analogous to the *four laws of classical thermodynamics:*

- *Zeroth Law: The surface gravity κ of a stationary black hole is constant over its event horizon.*

  This is analogous to the *zeroth law of thermodynamics:* temperature T is constant for a system in thermal equilibrium.

- *First Law: Energy is conserved* because of a relation

$$\delta M = \frac{1}{8\pi}\kappa\delta A + \Omega\delta J + \Phi\delta Q,$$

  - where M is the *mass*,
  - κ is the *surface gravity*,
  - A is the *horizon area*,
  - Ω is the *angular velocity*,
  - J is the *angular momentum*, and
  - Φ is the *electrostatic potential*

  for the black hole.

  This is analogous to the *first law of thermodynamics:* energy is conserved.

- *Second Law:* *The area theorem* or its *quantum generalization*

$$\frac{dA}{dt} \geq 0 \qquad dS \geq 0 \qquad S \equiv \frac{k_B}{4h}A$$

  This is the analog of the *second law of thermodynamics:* total entropy cannot decrease.

- *Third Law:* *The surface gravity* κ *of a black hole* cannot be reduced to zero by a finite series of operations.

  Analogous to the *Nernst form of the third law of thermodynamics:* temperature cannot be reduced to zero in a finite series of operations.

Thus the *four laws of black hole dynamics* are analogous to the *four laws of thermodynamics* if an identification is made between

- *Temperature* T and some multiple of the *surface gravity* κ.
- *Entropy* S and some multiple of the *event horizon area* A.

The four laws of black hole dynamics are more *well-grounded theoretical conjecture* than established law at this point.

12.7 Gravity and Quantum Mechanics: the Planck Scale

The preceding results for Hawking radiation are derived assuming

- that the spacetime in which the quantum calculations are done is
  - a fixed background that is
  - not influenced by the propagation of the Hawking radiation.
- This approximation is expected to be valid as long as $E << M$, where E is the average energy of the Hawking radiation and M is the mass of the black hole.
- It breaks down on the *Planck scale,* defined by

$$\textit{Planck mass}: \quad M_{\rm P} \equiv \left(\frac{\hbar c}{G}\right)^{1/2} = 2.18 \times 10^{-5}\,{\rm g},$$

$$\textit{Planck length}: \quad \ell_{\rm P} \equiv \left(\frac{\hbar G}{c^3}\right)^{1/2} = \frac{\hbar c}{E_{\rm P}} = 1.62 \times 10^{-33}\,{\rm cm},$$

$$\textit{Planck time}: \quad t_{\rm P} \equiv \left(\frac{\hbar G}{c^5}\right)^{1/2} = \frac{\ell_{\rm P}}{c} = 5.39 \times 10^{-44}\,{\rm s},$$

$$\textit{Planck energy}: \quad E_{\rm P} \equiv M_{\rm P}c^2 = 1.22 \times 10^{19}\,{\rm GeV},$$

$$\textit{Planck temperature}: \quad T_{\rm P} \equiv \frac{E_{\rm P}}{k_{\rm B}} = 1.44 \times 10^{32}\,{\rm K}.$$

- For a *black hole of Planck mass*,
  - the effects of gravity become important even on a quantum ($\hbar$) scale,
  - requiring a theory of *quantum gravity*.
- We don't yet have an adequate theory of quantum gravity.
- Note that *one approaches the Planck scale* near the endpoint of Hawking black hole evaporation.

Therefore, *we do not actually know yet what happens* at the conclusion of Hawking evaporation for a black hole.

12.8 Black Holes and Information

Entropy is related to *information content:*

- It is equal to the logarithm of the number of microscopic configurations that leave the macroscopic description of an object unchanged.
- Black hole evaporation by the Hawking mechanism leads to *apparent paradoxes* associated with this relationship.
- This may be illustrated by noting that
  - Hawking radiation is *produced randomly* by vacuum fluctuations, so
  - in the simplest picture it *contains no information*.
- Thus, the information content of the matter from which the black hole formed
  - appears to be *lost to the Universe* if
  - the black hole then *decays completely away* by Hawking radiation.

This is a complex issue that is *not fully resolved*.

- Some contend that perhaps this means that the Universe does not conserve information.
- Others have conjectured that a (future) quantum gravity treatment of black hole evaporation may resolve the issue.

12.8.1 The Holographic Principle

The entropy relation

$$S \equiv \frac{k_B}{4h} A,$$

implies that

- The entropy of a black hole is proportional to the *surface area* of its horizon.
- But the information content of the black hole is associated with its entropy and
- The information content of a region of space is usually thought of as being proportional to the *volume* of that region, *not to the area of a bounding surface*.
- This has led to a proposed solution of the black hole information paradox called the *holographic principle.*
- The holographic principle asserts that the description of a volume of space can be thought of as being *encoded on a 2-dimensional boundary of that region*.

For a black hole, this implies that

> Surface fluctuations of the event horizon must in some way contain a complete description of all the objects that have ever fallen into the black hole.

12.8.2 The Holographic Universe

Even more speculatively, the holographic principle has been extended to a cosmological statement that

> Perhaps the entire universe should be thought of as a two-dimensional information structure painted on the cosmological horizon.

Extension of the holographic principle to a cosmological statement may be motivated by noting that

- A universe with a cosmological horizon resembles in some ways the interior of a black hole.
- In this view the entire Universe is a kind of gigantic hologram and
- Our perception that it has three (rather than two) spatial dimensions is an illusion rooted in an effective description of the actual Universe that is valid only at low energies.

> The *AdS/CFT correspondence* (more generally *gauge/gravity duality*) described in later chapters is a specific implementation of such ideas.

Chapter 13

Rotating Black Holes

The *Schwarzschild solution* appropriate outside a spherical, non-spinning mass distribution, was discovered in 1916.

- It was *not until 1963* that a solution corresponding to spinning black holes was discovered.
- This solution leads to the possible existence of a family of
  - rotating,
  - deformed

  metrics called *Kerr black holes.*

Angular momentum is complicated:

- In Newtonian gravity rotation
  - produces centrifugal effects but
  - does not influence the gravitational field directly.
- But angular momentum implies rotational energy, so in general relativity
  - *rotation* of a gravitational field is itself
  - a *source for the field*.
- In addition, strongly-rotating solutions imply *deviation from spherical symmetry*, which *complicates the mathematics* immensely.
- For very small angular momentum the spacetime may be approximated by a perturbation of the Schwarzschild metric but
- This is no longer a valid solution for large angular momentum.

Hence the long delay in finding a solution corresponding to black holes with significant angular momentum.

13.1 The Kerr Solution

The Schwarzschild solution corresponds to the *simplest black hole*, characterized by a single parameter, the mass M.

- Other solutions to the field equations of general relativity permit black holes with more degrees of freedom.

  The most general black hole can possess

  1. *Mass,*
  2. *Charge,* and
  3. *Angular momentum*

  as distinguishing quantities.

- Of particular interest are those solutions where we
  - relax the restriction to spherical symmetry and thus
  - permit the black hole to be characterized by angular momentum in addition to mass

  These are *Kerr black holes*.

- Rotating black holes are *important in astrophysics* because they power phenomena like quasars and other active galaxies, X-ray binaries, and gamma-ray bursts.
- Unlike for Schwarzschild black holes, it is possible to devise mechanisms that permit energy and angular momentum to be extracted from a (classical) rotating black hole

13.1.1 The Kerr Metric

The Kerr metric corresponds to the line element

$$ds^2 = -\left(1 - \frac{2Mr}{\rho^2}\right) dt^2 - \frac{4Mra\sin^2\theta}{\rho^2} d\varphi dt + \frac{\rho^2}{\Delta} dr^2 + \rho^2 d\theta^2 + \left(r^2 + a^2 + \frac{2Mra^2\sin^2\theta}{\rho^2}\right) \sin^2\theta d\varphi^2$$

with the definitions

$$a \equiv J/M \qquad \rho^2 \equiv r^2 + a^2\cos^2\theta \qquad \Delta \equiv r^2 - 2Mr + a^2.$$

- The coordinates (t, r, θ, φ) are called *Boyer–Lindquist coordinates*.
- The parameter a, termed the *Kerr parameter,* has units of length in geometrized units.
- The parameter J will be interpreted as *angular momentum* and
- the parameter M will be interpreted as the *mass* for the black hole.

The Kerr solution

$$ds^2 = -\left(1 - \frac{2Mr}{\rho^2}\right) dt^2 - \frac{4Mra\sin^2\theta}{\rho^2} d\varphi dt + \frac{\rho^2}{\Delta} dr^2 + \rho^2 d\theta^2 + \left(r^2 + a^2 + \frac{2Mra^2\sin^2\theta}{\rho^2}\right) \sin^2\theta d\varphi^2$$

$$a \equiv J/M \qquad \rho^2 \equiv r^2 + a^2\cos^2\theta \qquad \Delta \equiv r^2 - 2Mr + a^2.$$

has the following properties.

- *Vacuum solution:* the Kerr metric is a *vacuum solution of the Einstein equations*, valid in the absence of matter.
- *Reduction to Schwarzschild metric:* If the black hole is not rotating ($a = J/M = 0$), the Kerr line element reduces to the Schwarzschild line element.
- *Asymptotically flat:* The Kerr metric becomes asymptotically flat for $r >> M$ and $r >> a$.
- *Symmetries:* The Kerr metric is independent of t and φ, implying the existence of Killing vectors

  $K_t = (1,0,0,0)$ (stationary metric)
  $K_\varphi = (0,0,0,1)$ (axially symmetric metric).

  The Kerr metric has *only axial (not spherical)* symmetry.
- The metric has *off-diagonal terms*

  $$g_{03} = g_{30} = -\frac{2Mra\sin^2\theta}{\rho^2}$$ *(inertial frame dragging).*

For the Kerr metric

$$ds^2 = -\left(1 - \frac{2Mr}{\rho^2}\right) dt^2 - \frac{4Mra\sin^2\theta}{\rho^2} d\varphi dt + \frac{\rho^2}{\Delta} dr^2 + \rho^2 d\theta^2 + \left(r^2 + a^2 + \frac{2Mra^2\sin^2\theta}{\rho^2}\right) \sin^2\theta d\varphi^2$$

$$a \equiv J/M \qquad \rho^2 \equiv r^2 + a^2\cos^2\theta \qquad \Delta \equiv r^2 - 2Mr + a^2.$$

- Surfaces of constant Boyer–Lindquist coordinates r and t *do not have the metric of a 2-sphere*.
- *Singularity and horizon structure:* $\Delta \to 0$ at

$$r_\pm = M \pm \sqrt{M^2 - a^2}$$

and $ds \to \infty$, assuming $a \leq M$.

1. This is a *coordinate singularity*.
2. As $a \to 0$ we find that $r_+ \to 2M$, which *coincides with the Schwarzschild coordinate singularity*.
3. Thus r_+ *corresponds to the horizon* that makes the Kerr solution a black hole.

On the other hand, the limit $\rho \to 0$ corresponds to

- a *physical singularity* with associated
- components of *infinite spacetime curvature*,

similar to the case for the Schwarzschild solution.

13.1.2 Extremal Kerr Black Holes

- The horizon radius

$$r_{\pm} = M \pm \sqrt{M^2 - a^2}$$

exists only for $a \leq M$.

- Thus, there is a *maximum angular momentum* $J_{\max}$ *for a Kerr black hole* since $a = J/M$ and $a_{\max} = M$,

$$J_{\max} = a_{\max} M = M^2.$$

- Black holes for which $J = M^2$ are termed *extreme Kerr black holes.*

- Near-extreme black holes may develop in many astrophysical situations:

  1. Angular momentum transfer through accretion disks in either
     - binary star systems or
     - around supermassive black holes in galaxy cores

     tends to spin up the central object.
  2. Massive stars collapsing to black holes may have significant initial angular momentum

13.1.3 Cosmic Censorship

That the horizon

- exists for a Kerr black hole
- only under restricted conditions

raises the question of whether a singularity could exist in the absence of a horizon (*"naked singularity"*).

> *Cosmic Censorship Hypothesis:* Nature conspires to "censor" spacetime singularities in that all such singularities come with event horizons that render them invisible to the outside universe.

- No known violations (observationally or theoretically).
- It cannot at this point be derived from any more fundamental concept and must be viewed as only an hypothesis.

> All realistic theoretical attempts to add angular momentum to a Kerr black hole beyond the extremal limit have failed.

13.1.4 The Kerr Horizon

The *area of the horizon* for a black hole is of considerable importance because of the Hawking area theorem:

> The horizon area of a classical black hole can never decrease in any physical process.

- The Kerr horizon corresponds to a constant value of $r = r_+$.
- Since the metric is stationary, the horizon is also a surface of constant t.
- Setting $dr = dt = 0$ in the Kerr line element gives the line element for the 2-dimensional horizon,

$$d\sigma^2 = \rho_+^2 d\theta^2 + \left(\frac{2Mr_+}{\rho_+}\right)^2 \sin^2\theta d\varphi^2,$$

  where ρ_+^2 is defined by

$$\rho^2 \equiv r_+^2 + a^2\cos^2\theta.$$

- This is *not the line interval of a 2-sphere*.

> The Kerr horizon has constant Boyer–Lindquist coordinate $r = r_+$ but it is *not spherical*.

The metric tensor for the Kerr horizon is

$$g = \begin{pmatrix} \rho_+^2 & 0 \\ 0 & \left(\dfrac{2Mr_+}{\rho_+}\right)^2 \sin^2\theta \end{pmatrix}$$

The area of the Kerr horizon is then

$$\begin{aligned} A_{\mathrm{K}} &= \int_0^{2\pi} d\varphi \int_0^{\pi} \sqrt{\det g}\, d\theta \\ &= 2Mr_+ \int_0^{2\pi} d\varphi \int_0^{\pi} \sin\theta\, d\theta \\ &= 8\pi M r_+ = 8\pi M \left(M + \sqrt{M^2 - a^2}\right). \end{aligned}$$

The horizon area for a Schwarzschild black hole

- is obtained by setting $a = 0$ in this expression,
- corresponding to *vanishing angular momentum*,
- in which case $r_+ = 2M$ and

$$A_{\mathrm{S}} = 16\pi M^2,$$

as expected for a *spherical horizon of radius* $2M$.

13.2 Orbits in the Kerr Metric

We may take a similar approach as in the Schwarzschild metric to determine the orbits of particles and photons.

- In the Kerr metric are not confined to a plane (only axial, not full spherical symmetry).
- Nevertheless, we shall consider *only motion in the equatorial plane* ($\theta = \frac{\pi}{2}$) to illustrate concepts.

The Kerr line element with that restriction is then

$$ds^2 = -\left(1 - \frac{2M}{r}\right) dt^2 - \frac{4Ma}{r} d\varphi dt + \frac{r^2}{\Delta} dr^2 + \left(r^2 + a^2 + \frac{2Ma^2}{r}\right) d\varphi^2.$$

There are two *Killing vectors*,

$$K_t = (1,0,0,0) \qquad K_\varphi = (0,0,0,1),$$

and *two associated conserved quantitities*

$$\varepsilon = -K_t \cdot u \qquad \ell = K_\varphi \cdot u,$$

At large distances from the gravitational source

- ε is the conserved *energy per unit rest mass*.
- ℓ may be interpreted as the *angular momentum component per unit mass along the symmetry axis*.

The norm of the 4-velocity provides the usual constraint

$$u \cdot u = -1.$$

for timelike (massive) particles.

From the Kerr metric

$$-\varepsilon = g_{00}u^0 + g_{03}u^3 \qquad \ell = g_{30}u^0 + g_{33}u^3,$$

and solving for $u^0 = dt/d\tau$ and $u^\varphi = d\varphi/d\tau$ gives

$$\frac{dt}{d\tau} = \frac{1}{\Delta}\left[\left(r^2 + a^2 + \frac{2Ma^2}{r}\right)\varepsilon - \frac{2Ma}{r}\ell\right]$$

$$\frac{d\varphi}{d\tau} = \frac{1}{\Delta}\left[\left(1 - \frac{2M}{r}\right)\ell + \frac{2Ma}{r}\varepsilon\right].$$

By similar steps as for the Schwarzschild metric, this leads to the equations of motion for timelike particles

$$\frac{\varepsilon^2 - 1}{2} = \frac{1}{2}\left(\frac{dr}{d\tau}\right)^2 + V_{\text{eff}}(r, \varepsilon, \ell)$$

$$V_{\text{eff}}(r, \varepsilon, \ell) \equiv -\frac{M}{r} + \frac{\ell^2 - a^2(\varepsilon^2 - 1)}{2r^2} - \frac{M(\ell - a\varepsilon)^2}{r^3}.$$

A similar approach can be used to determine the *motion of photons* in the Kerr metric.

- The primary difference is that the *massless photons* are now *required to satisfy*

$$u \cdot u = 0.$$

- Let's skip the details and just quote the result; for *lightlike particles*,

$$\frac{1}{\ell^2}\left(\frac{dr}{d\lambda}\right)^2 = \frac{1}{b^2} - V(r,b,\sigma),$$
$$V(r,b,\sigma) \equiv \frac{1}{r^2}\left[1 - \frac{a^2}{b^2} - \frac{2M}{r}\left(1 - \frac{\sigma a}{b}\right)^2\right],$$

- where λ is an *affine parameter*,
- $b \equiv |\ell/\varepsilon|$ is the *impact parameter*, and
- $\sigma = \operatorname{sign}\ell$ indicates whether the photon moves
  - *with the rotational motion* of the black hole ($\sigma = +1$)
  - or *against it* ($\sigma = -1$).

13.3 Frame Dragging

A striking feature of the Kerr solution is *frame dragging:* loosely, the black hole *drags spacetime with it* as it rotates.

- This arises ultimately because the *Kerr metric contains off-diagonal components* $g_{03} = g_{30}$.
- One consequence is that *a particle dropped radially* acquires *non-radial components of motion* as it falls freely.

Let's investigate by calculating $d\varphi/dr$ for a particle

- that is dropped from rest ($\varepsilon = 1$), with
- zero angular momentum ($\ell = 0$) onto a Kerr black hole.

For 4-momenta in the Kerr metric

$$p^\varphi \equiv p^3 = g^{3\mu} p_\mu = g^{33} p_3 + g^{30} p_0 \qquad p^0 = mu^0 = m dt/d\tau$$
$$p^t \equiv p^0 = g^{0\mu} p_\mu = g^{00} p_0 + g^{03} p_3 \qquad p^3 = mu^3 = m d\varphi/d\tau$$

and combining these relations gives an expression for $d\varphi/dt$,

$$\frac{d\varphi}{dt} = \frac{p^3}{p^0} = \frac{g^{33} p_3 + g^{30} p_0}{g^{00} p_0 + g^{03} p_3}.$$

We have obtained

$$\frac{d\varphi}{dt} = \frac{p^3}{p^0} = \frac{g^{33}p_3 + g^{30}p_0}{g^{00}p_0 + g^{03}p_3}.$$

But from the Killing vector K_φ and the metric

$$\ell \equiv K_\varphi \cdot u = g_{30}u^0 + g_{33}u^3,$$

we have that for an $\ell = 0$ particle

$$p_3 = mu_3 = g_{30}p^0 + g_{33}p^3 = 0$$

and thus

$$\omega(r,\theta) \equiv \frac{d\varphi}{dt} = \frac{g^{30}}{g^{00}} = \frac{g^{\varphi t}}{g^{tt}}.$$

- The quantity $\omega(r,\theta)$ measures the *amount of frame dragging*.
- It may be viewed as the *angular velocity* of a *zero angular momentum particle*.

> The particle is dragged in angle φ as it falls radially inward, *even though no forces act on it*.

Frame dragging has been exhibited here for a Kerr metric but

- It is expected to occur for any metric having terms depending on angular momentum.
- It produces a *detectable gyroscopic precession* termed the *Lense–Thirring effect* that we discussed in Ch. 9.

NASA's *Gravity Probe B* had 4 gyroscopes aboard a satellite in an orbit almost directly over the poles.

- It *measured gyroscopic precession*,
- giving direct evidence of *frame dragging* by the rotating gravitational field of the Earth.
- The *measured amount of frame dragging* was the *amount predicted by general relativity*.

The Kerr line element

- defines the *covariant components* of the Kerr metric $g_{\mu\nu}$.
- To evaluate $\omega(r,\theta)$ we need the *contravariant components* $g^{\mu\nu}$.

These may be obtained as the *matrix inverse* of $g_{\mu\nu}$. The metric is diagonal in r and θ so

$$g^{rr} = g_{rr}^{-1} = \frac{\Delta}{\rho^2} \qquad g^{\theta\theta} = g_{\theta\theta}^{-1} = \frac{1}{\rho^2},$$

and we need only evaluate

$$g^{-1} = \begin{pmatrix} g_{tt} & g_{t\varphi} \\ g_{\varphi t} & g_{\varphi\varphi} \end{pmatrix}^{-1}$$

to obtain the other non-zero entries for $g^{\mu\nu}$. Letting

$$D = \det g = g_{tt}g_{\varphi\varphi} - (g_{t\varphi})^2,$$

the matrix inverse is

$$g^{-1} = \frac{1}{D}\begin{pmatrix} g_{\varphi\varphi} & -g_{t\varphi} \\ -g_{\varphi t} & g_{tt} \end{pmatrix}.$$

Inserting explicit expressions for the $g_{\mu\nu}$ and carrying out some algebra yields

$$\omega(r,\theta) = \frac{g^{\varphi t}}{g^{tt}} = \frac{2Mra}{(r^2+a^2)^2 - a^2\Delta\sin^2\theta}.$$

Notice that the angular velocity that measures the amount of frame dragging

$$\omega(r,\theta) = \frac{g^{\varphi t}}{g^{tt}} = \frac{2Mra}{(r^2+a^2)^2 - a^2\Delta\sin^2\theta}.$$

- has the same sign as the Kerr parameter $a = J/M$ and
- falls off as r^{-3} for large r.

This indicates that *frame dragging is most pronounced at small* r in the Kerr spacetime.

We will now demonstrate that a region just outside the horizon has quite remarkable properties as a consequence.

13.3.1 The Ergosphere

With capable enough propulsion,

- an observer could remain stationary at any point *outside* the event horizon of a Schwarzschild black hole.
- However, no amount of propulsion can enable an observer to remain stationary *inside* a Schwarzschild horizon (this would imply *causality violation*).

We now demonstrate that, for a rotating black hole, *even outside the horizon* it may be impossible for an observer to remain stationary.

The 4-velocity for a *stationary observer* is

$$u^{\mu}_{\text{obs}} = (u^0_{\text{obs}}, 0, 0, 0) = \left(\frac{dt}{d\tau}, 0, 0, 0\right).$$

Writing out $u_{\text{obs}} \cdot u_{\text{obs}} = -1$ gives

$$u_{\text{obs}} \cdot u_{\text{obs}} = g_{00}(u^0_{\text{obs}})^2 = -1.$$

But for the Kerr metric

$$\begin{aligned} g_{00} &= -\left(1 - \frac{2Mr}{\rho^2}\right) \\ &= -\left(1 - \frac{2Mr}{r^2 + a^2\cos^2\theta}\right) = -\left(\frac{r^2 + a^2\cos^2\theta - 2Mr}{r^2 + a^2\cos^2\theta}\right) \end{aligned}$$

vanishes if

$$r^2 - 2Mr + a^2\cos^2\theta = 0.$$

Generally then, *solving the above quadratic for* r

- $g_{00} = 0$ on the surface defined by

$$r_{\text{e}}(\theta) = M + \sqrt{M^2 - a^2\cos^2\theta},$$

- $g_{00} > 0$ inside this surface.
- $g_{00} < 0$ outside this surface.

Therefore, *since* $(u^0_{\text{obs}})^2$ *must be positive*,

$$u_{\text{obs}} \cdot u_{\text{obs}} = g_{00}(u^0_{\text{obs}})^2 = -1$$

cannot be satisfied for $r \leq r_{\text{e}}(\theta)$.

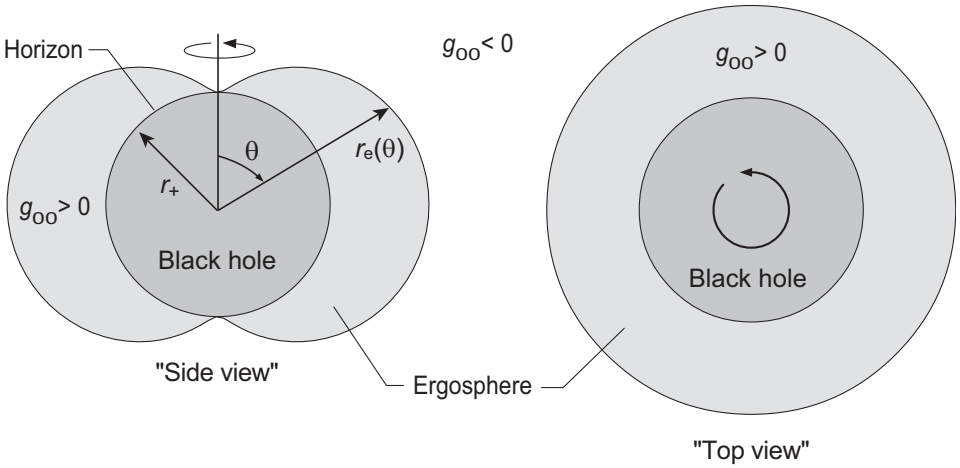


Figure 13.1: The ergosphere (light gray regions) of a Kerr black hole with $a/M = 0.998$ in Boyer–Lindquist coordinates. The geometry is not correctly illustrated. For example, the horizon has a constant coordinate $r = r_+$ but its geometry is not spherical.

Comparing

$$r_\pm = M \pm \sqrt{M^2 - a^2} \qquad r_e(\theta) = M + \sqrt{M^2 - a^2 \cos^2\theta},$$

we see that

- If $a \neq 0$ the surface $r_e(\theta)$ generally lies outside the horizon r_+, except at the poles, where the two surfaces are coincident (Figure 13.1).
- If $a = 0$, the Kerr black hole reduces to a Schwarzschild black hole, where r_+ and $r_e(\theta)$ define the same surface.

The region lying between $r_e(\theta)$ and the horizon r_+ is termed the *ergosphere*, for reasons that will become more apparent shortly.

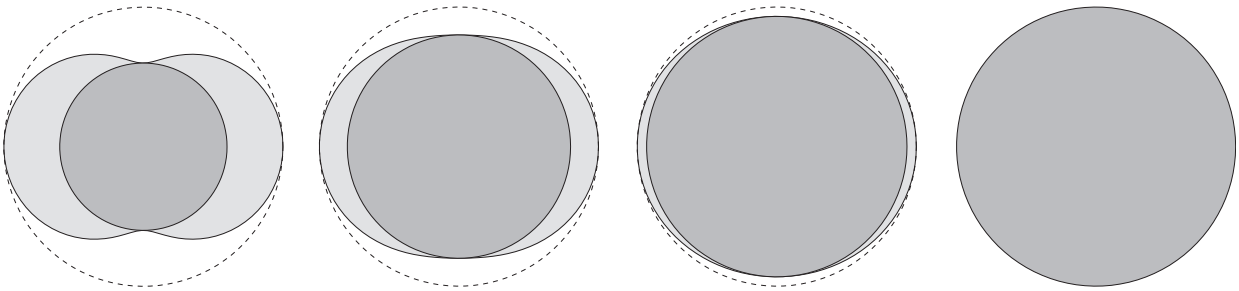


Figure 13.2: Ergospheres of a Kerr black hole for different values of a/M at constant mass M. The region inside the horizon (the black hole) is indicated by dark gray and the ergosphere is indicated in lighter gray. The dashed circle defines the equivalent Schwarzschild radius $r = 2M$. As $a \to 0$ the outer boundary of the ergosphere and the Kerr horizon approach the event horizon of a Schwarzschild black hole of the same mass.

The *relationship of the ergosphere and the event horizon* for Kerr black holes having different values of a is illustrated in Fig. 13.2.

From preceding considerations, there can be *no stationary observers within the ergosphere.*

Further insight comes from considering *motion of photons within the ergosphere*.

- Assume photons within the ergosphere moving tangent to a circle at constant r in the equatorial plane $\theta = \frac{\pi}{2}$, so $dr = d\theta = 0$.
- Since they are photons, $ds^2 = 0$ and from the Kerr line element with $dr = d\theta = 0$,

$$g_{00}dt^2 + 2g_{03}dtd\varphi + g_{33}d\varphi^2 = 0.$$

- Divide by dt^2 to give a quadratic equation in $d\varphi/dt$

$$g_{00} + 2g_{03}\frac{d\varphi}{dt} + g_{33}\left(\frac{d\varphi}{dt}\right)^2 = 0$$

Thus

$$\begin{aligned}\frac{d\varphi}{dt} &= \frac{-2g_{03} \pm \sqrt{4g_{03}^2 - 4g_{33}g_{00}}}{2g_{33}} \\ &= -\frac{g_{03}}{g_{33}} \pm \sqrt{\left(\frac{g_{03}}{g_{33}}\right)^2 - \frac{g_{00}}{g_{33}}},\end{aligned}$$

where

- $+$ is for motion *opposite black hole rotation*
- $-$ is for motion *with black hole rotation*.

- Now g_{00} vanishes at the boundary of the ergosphere and it is positive inside the boundary.
- Setting $g_{00} = 0$ in

$$\frac{d\varphi}{dt} = -\frac{g_{03}}{g_{33}} \pm \sqrt{\left(\frac{g_{03}}{g_{33}}\right)^2 - \frac{g_{00}}{g_{33}}},$$

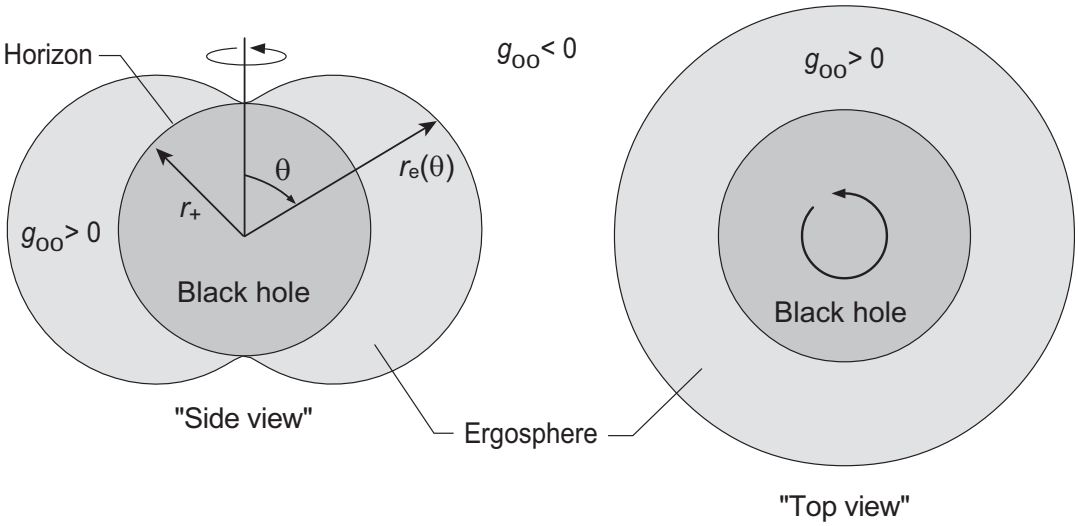
gives two solutions,

$$\underbrace{\frac{d\varphi}{dt} = 0}_{\text{opposite rotation}} \qquad \underbrace{\frac{d\varphi}{dt} = -\frac{2g_{03}}{g_{33}}}_{\text{with rotation}},$$

The photon sent backwards against the rotation at the surface of the ergosphere is *stationary in* φ!

- Obviously a particle, which must have a velocity less than a photon, must rotate with the black hole irrespective of the amount of angular momentum that it has.
- Inside the ergosphere $g_{00} > 0$, so *all photons and particles must rotate with the black hole.*

The frame dragging for $r < r_e(\theta)$ is so severe that $v > c$ *would be required for an observer to remain at rest* with respect to infinity.



- In the limit $a = J/M \to 0$, where the angular momentum of the Kerr black hole vanishes,

$$r_e(\theta) = M + \sqrt{M^2 - a^2 \cos^2\theta} \quad \longrightarrow \quad 2M$$

  and becomes *coincident with the Schwarzschild horizon*.

- Rotation has extended the region $r < 2M$ of the spherical black hole where no stationary observers can exist to a
  - larger region $r < r_e(\theta)$ surrounding the rotating black hole where
  - no observer can remain at rest because of frame dragging effects.

This ergospheric region lies *outside the horizon*, implying that a particle could *enter it and still escape* from the black hole.

13.4 Extracting Rotational Energy from Black Holes

We have seen that

- Quantum vacuum fluctuations allow mass to be extracted from a Schwarzschild black hole as *Hawking radiation*.
- But it is impossible to extract mass from a *classical* Schwarzschild black hole (it all lies within the event horizon).

However, the existence of

- *separate surfaces defining the ergosphere and the horizon* for a Kerr black hole
- implies the possibility of *extracting rotational energy* from a *classical* black hole.

The simplest way to demonstrate the feasibility of extracting rotational energy from a black hole is through a *Penrose process*.

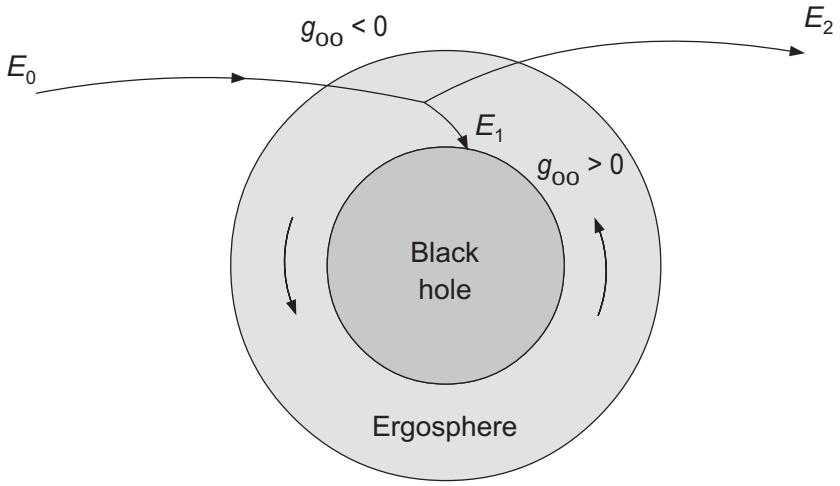


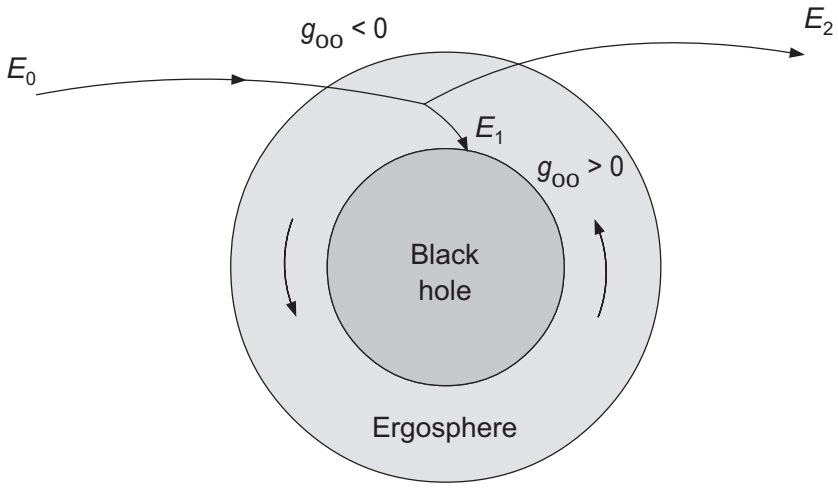
Figure 13.3: A Penrose process.

A Penrose process is illustrated schematically in Fig. 13.3.

- A particle falls into the ergosphere of a Kerr black hole and decays into two particles.
  - *Particle 1* falls through the horizon;
  - *Particle 2* exits the ergosphere and escapes to infinity.
- The decay within the ergosphere is a local process. By equivalence principle arguments, it may be analyzed in a freely falling frame by the usual rules of scattering theory.
- Therefore, energy and momentum are conserved in the decay, implying that in terms of 4-momenta p

$$p_0 = p_2 + p_1$$

(*Note: subscripts label particles, not components*).



- If particle 2 has rest mass m_2, its energy is

$$E_2 = -p_2 \cdot K_t = m_2 \varepsilon \qquad K_t = (1,0,0,0)$$

where $\varepsilon = -K_t \cdot u = -K_t \cdot (p/m)$ has been used.

- Taking the scalar product of $p_0 = p_2 + p_1$ with K_t

$$p_0 \cdot K_t = (p_2 + p_1) \cdot K_t$$

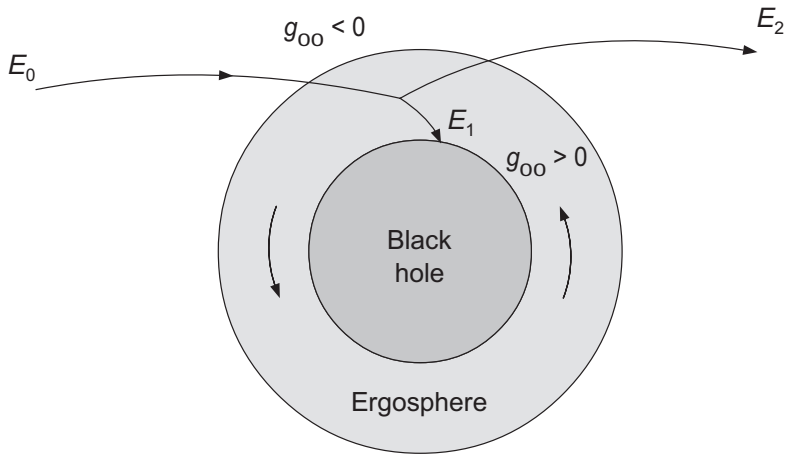
yields the requirement,

$$E_2 = E_0 - E_1,$$

since $E = -K_t \cdot p$.

- If particle 1 were to reach ∞ instead of crossing the event horizon, E_1 would necessarily be positive so

1. $E_2 < E_0$, and
2. Less energy would be emitted than put in.



However . . .

- For the Killing vector $K_t = (1,0,0,0)$,

$$K_t \cdot K_t = g_{\mu\nu} K_t^\mu K_t^\nu$$
$$= g_{00} K_t^0 K_t^0 = g_{00} = -\left(1 - \frac{2Mr}{\rho^2}\right),$$

  and g_{00} vanishes on $r_e(\theta)$ and is positive inside it.

- Therefore, *within the ergosphere K_t is a spacelike vector:*

$$K_t \cdot K_t = g_{00} > 0 \qquad \text{(within the ergosphere)}.$$

- By arguments similar to those for Hawking radiation,

  1. $-E_1$ is *not an energy* within the ergosphere but is a *component of spatial momentum*, which can have either a *positive or negative value (!)*.
  2. For decays where the trajectories are arranged such that $E_1 < 0$, we obtain from $E_2 = E_0 - E_1$ that $E_2 > E_0$ and *net energy is extracted in the Penrose process*.

The energy extracted in the Penrose process comes *at the expense of the rotational energy of the Kerr black hole*.

- For trajectories with $E_1 < 0$,
  - the captured particle adds a negative angular momentum,
  - thus reducing the angular momentum and total energy of the black hole,

  and this just balances the angular momentum and total energy carried away by the escaping particle.
- A series of Penrose events could extract all the angular momentum of a Kerr black hole, leaving a Schwarzschild black hole.
- No further energy can then be extracted from the resulting spherical black hole (except by quantum Hawking processes).

The Penrose mechanism establishes *proof of principle* in a simple model that

> The rotational energy of a Kerr black hole is externally accessible in classical processes.

- Practically, Penrose processes
  - are not likely to be important in astrophysics because
  - the required conditions are not easily realized.
- Instead, the primary sources of emitted energy from black hole systems likely are
  1. Complex *electromagnetic coupling* of rotating black holes to external accretion disks and jets.
  2. Gravitational energy released by *accretion* onto the black hole.

> There is very strong observational evidence for such processes in a variety of astrophysical environments.

Chapter 14

Observational Evidence for Black Holes

By definition an isolated black hole should be *a difficult object to observe*.

- However, if black holes exist they should often be
  - *accreting surrounding matter* and
  - *interacting gravitationally* with their environment,
  - and this is *potentially observable*.

Thus, although the direct observation of a black hole is difficult, we have very strong reasons to believe that black holes exist and are being observed indirectly.

14.1 Gravitational Collapse and Observations

Three major discoveries pushed *gravitational collapse* and *general relativity* to the forefront of observational astronomy:

- The realization in 1963 that *quasars* were enormous energy sources lying at great distance.
- The discovery of *pulsars* in 1967.
- Discoveries beginning in the 1970s of *X-ray binaries* containing *massive unseen companions*.

Attention soon focused on the possibility that gravitationally-collapsed objects:

- *black holes* for quasars,
- *neutron stars* for pulsars, and
- *black holes or neutron stars* for high-mass X-ray binaries

could explain these new discoveries.

> Quasars in particular pushed the envelope because of the proposal that they might be powered by
>
> - rotating,
> - extremely massive
>
> black holes.

14.2 Singularity Theorems and Black Holes

But do black holes *really* exist? After all, both

- *Einstein* (the most famous physicist of his day) and
- *Eddington* (the most famous astronomer of his day)

felt strongly that such solutions of the field equations existed mathematically but *would never be realized in nature*.

Let's play devil's advocate for a moment.

- The Einstein equations can be solved to yield Schwarzschild or Kerr black holes primarily because they are *vacuum solutions with a high degree of symmetry.*
- Take the Kerr solution as representative (since Schwarzschild is a special case of Kerr).
- Because of its high symmetry the Kerr solution is *remarkably simple*, being described by only two parameters:
  - *Mass*
  - *Angular momentum*.
- This led to worries that the gravitational collapse producing a Kerr black hole was an *anomaly* created by *unrealistically high symmetry*.
- Hence the theoretical black holes of general relativity might not exist in the wild because
  - real gravitational collapse would likely *not proceed with such high symmetries*, and
  - more realistic asymmetric solutions might somehow *avoid the formation of singularities*.

Help in this regard emerged from highly-mathematical work that began in the 1960s and is associated with the names of Roger Penrose, Stephen Hawking, Brandon Carter, and others.

14.2.1 Global Methods in General Relativity

The classical formulation of general relativity is in terms of *local differential equations.*

- To build up a global picture of a spacetime one traditionally
  - *performs local computations* around each event in spacetime and then
  - *patches these solutions together* to give the global structure.
- This local formulation of general relativity was almost the only approach until the 1960s,
- Then it began to be realized that general relativity was also subject to powerful *global laws* that were often remarkably simple and elegant.

These techniques permit the global properties of a spacetime to be investigated *directly,* rather than by building up from local properties.

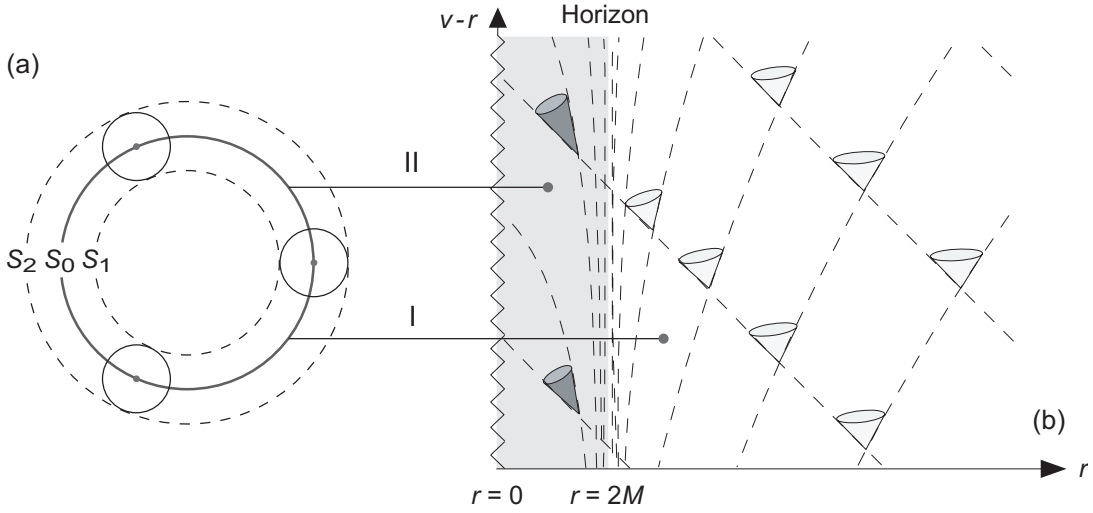


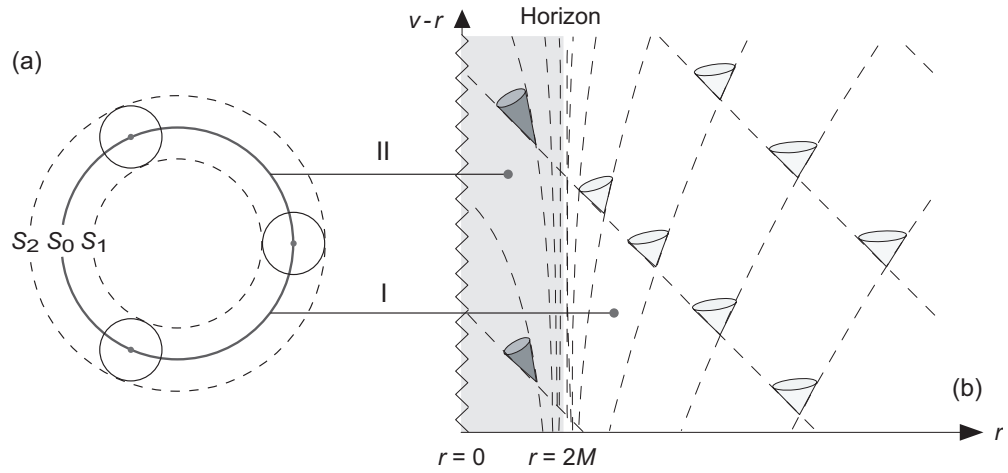
Figure 14.1: (a) Light emission by a 2-sphere S_0 in Schwarzschild spacetime. (b) Schwarzschild spacetime in Eddington–Finkelstein coordinates. There are two solutions at each point. Outside the horizon at $r = 2M$ one solution is ingoing and one is outgoing; inside $r = 2M$ *both* solutions are ingoing. Cases I and II correspond to S_0 located outside and inside the horizon, respectively. For a trapped surface all emitted light moves inward.

14.2.2 Singularities and Trapped Surfaces

In 1965, Roger Penrose introduced *global topological techniques* into the study of spacetimes.

- The details are far too technical for the present discussion but a brief qualitative overview will prove useful.
- A key idea was that of a *trapped surface,* which is a
  - *closed, spacelike 2-surface* from which
  - any light emitted *always moves inward*.

The basic idea is illustrated by Fig. 14.1.



Suppose a 2-sphere S_0, corresponding to a point in the above figure of a Schwarzschild spacetime, emits light.

- Photons will propagate away from S_0 in 2-spheres whose envelopes will form the 2-spheres S_1 and S_2.
- In *case I* the surface S_0 is *outside the horizon* at $r = 2M$ and the areas of the 2-spheres are ordered $S_2 > S_0 > S_1$.
- However, S_0 is *inside the horizon* for *case II* and
  - the orientation of lightcones implies that *both emitted wavefronts will implode* and in general
  - both S_2 and S_1 will have areas *less than* S_0.
- In this latter case S_0 is termed a *closed trapped surface,* since all light emitted from it (even "outgoing" light) travels inward toward the singularity.

The outermost trapped surface for a spacetime is termed the *apparent horizon.*

The preceding illustration is based on the *highly-symmetric Schwarzschild solution*.

- However, Penrose showed that *under very general conditions* (like positivity of energy, so that light is always focused by gravity),

  > *Once a trapped surface forms a singularity must form inside it.*

- Spacetimes admitting trapped surfaces exhibit *geodesic incompleteness*.
- This means that some worldlines for particles or light can be extended only for a finite proper time (finite affine parameter for light) for an observer on the geodesic.
- Penrose interpreted this to be the end of time (no future) at a singularity where physical law breaks down.
- This derivation *invokes minimal assumptions*.
- Hence it indicates that singularities (with attendant horizons if cosmic censorship is valid)
  - are *generic features* of any collapse where gravity is sufficiently strong,
  - *not just for ones with special initial conditions*.

14.2.3 Apparent Horizons and True Horizons

If it exists, the outermost trapped surface defines an *apparent horizon* for a spacetime.

- How are apparent and true event horizons for a black hole related?
- The apparent horizon is a *local* concept because it depends on tests made locally.
- However, the black hole event horizons discussed earlier are *defined globally,* since they
  - depend on constructing null geodesics and
  - determining whether they reach infinity or not.

> Thus determining the true event horizon requires knowing the complete history of the spacetime.

- For the simple case of stationary Schwarzschild spacetime the apparent and true horizons are located at the same radial coordinate, $r = 2M$, so we can speak of just *"the horizon"*,
- In more complex situations this might not be true.
- The next slide gives an example.

An example where the apparent and event horizons *need not coincide* is a black hole *accreting a spherical shell of matter*.

- Imagine a light ray emitted radially from just outside the horizon before the shell accretes.
- When emitted the light is not trapped, but after the shell accretes
  - the gravitational field acting on the light is increased and
  - may trap the light,

  even though it was not trapped when emitted.
- In this case the *apparent horizon is inside the true horizon*.
- Roughly, we may think that
  - the *apparent horizon* defines the light-trapping radius for the black hole *now*, while
  - the *true horizon* is that radius *in the future*.

When computing the dynamical evolution of a spacetime from initial data in *numerical relativity*,

- *the full global history of the spacetime is not available* as the solution is advanced.
- In such a context the local definition of an apparent horizon is useful because
- it permits *determining the presence of a black hole* at any time during the numerical simulation.
- This is so because

If it is true that

- *cosmic censorship* is valid and
- the *null energy condition* is satisfied,

an apparent horizon *signals the existence of a true horizon* that

- either *coincides with* or
- *lies outside*

the apparent horizon.

The overall picture that emerged from such considerations was that

- Asymmetries do not prevent gravitational collapse to a singularity because
- uncharged collapsing matter radiates away its irregularities as gravitational waves,
- leaving finally
  - a *Kerr black hole* with
  - a *singularity clothed by an event horizon*.

Thus it was suggested that the final equilibrium configuration of *any* fully-collapsed, uncharged object is a Kerr black hole (or its $J \to 0$ limit, a Schwarzschild black hole).

- These ideas made it much more plausible that gravitational collapse could form black holes under real-world conditions.
- This in turn spurred *increased interest in observational evidence* for the existence of black holes that will be the subject of this chapter and the next.

14.2.4 Generalized Singularity Theorems

Singularity theorems typically assume

1. a *global structure for spacetime*,
2. a *restriction on allowed energies*, and
3. *gravity strong enough to trap a region of spacetime* in the sense described above.

By making different choices for these assumptions, different singularity theorems may be obtained.

- Such generalizations were used to show that *classically*, the expanding Universe must have a singularity in its past.
- This will be relevant to later discussion of the big bang.
- They also were used by Hawking to prove the *area theorem for black hole horizons.*
- This, along with Hawking's *introduction of quantum mechanics into black hole physics*, was crucial to the development of black hole thermodynamics.

Having given some *theoretical basis* for why black holes might form for realistic gravitational collapse, let's now turn to some of the *observational evidence* suggesting their existence.

14.3 Observing Black Holes

We have very strong reasons to believe that black holes exist and are being observed indirectly because of

1. *Unseen massive companions in binary star systems* that are
   - Strong *sources of X-rays* and
   - Probably *too massive* to be anything other than black holes.
2. The *centers of many galaxies* where
   - Masses inferred by
     - direct measurement of individual star orbits in the center of the Milky Way or
     - virial theorem methods

     are far too large to be accounted for by any simple hypothesis other than a supermassive black hole.
   - In many cases there is direct evidence of an enormous energy source in the center of the galaxy.
3. Observation of *gravitational waves* with properties indicating that they originated in the merger of black holes.

We shall begin by summarizing some of the reasons why *properties of some binary star systems* give us strong confidence that black holes exist.

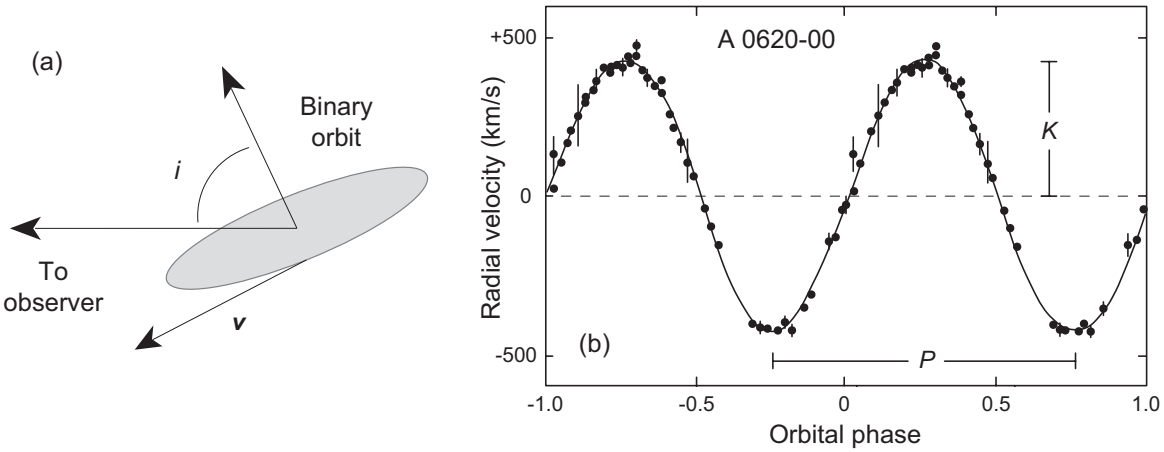


Figure 14.2: (a) Tilt angle i of a binary orbit. (b) Observed radial velocity curve for the black hole binary candidate A 0620–00. The period corresponds to $P = 0.323$ days and the semiamplitude to $K = 433 \pm 3\,\mathrm{km\,s^{-1}}$. The orbital phase corresponds to the fraction of one complete orbit.

14.4 Black Hole Masses in X-ray Binaries

Many close binaries have eccentricity $e \sim 0$. To simply, *assume circular orbits*. From *Kepler's laws*,

- the *mass function* $f(M)$ for the binary may be related to the *radial velocity curve* (Fig. 14.2b) through

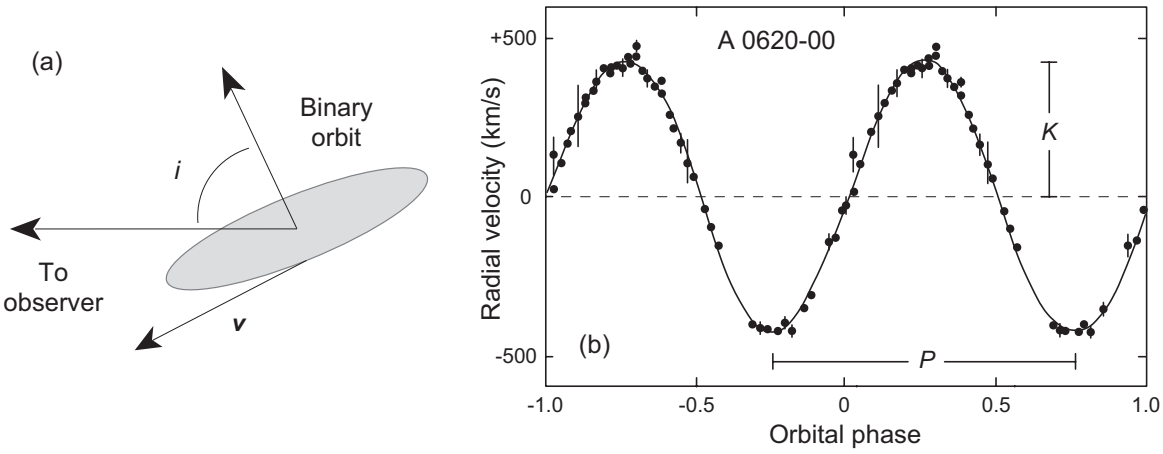
$$f(M) = \underbrace{\frac{(M \sin i)^3}{(M + M_c)^2}}_{\text{function of masses}} = \underbrace{\frac{PK^3}{2\pi G}}_{\text{measureable}}$$

- K is the *semiamplitude* and P the *period* of the curve,
- i is the *tilt angle* of the orbit (Fig. 14.2a)
- M_c is the mass of the visible star.
- M is the *mass of the compact, unseen companion*.

Because the angle i is *usually not known directly*, the measured mass function

- *Places a lower limit* on the mass of the unseen component of the binary.
- The mass of the companion can often be estimated reliably from *systematic spectral features*.

Let's see how this works in practice.



14.4.1 Masses from Mass Functions

- Assume that from *Doppler-shift measurements* of the visible component of a spectroscopic binary,

$$F = F(P,K) = \frac{PK^3}{2\pi G} \qquad \text{(observed)}$$

  is *known from the velocity curve* [Fig. (b) above].

- Then from the *mass function formula*,

$$\frac{M^3 \sin^3 i}{(M+M_c)^2} = F$$

  and the *compact-object mass* M is defined by solution of

$$M^3 + aM^2 + bM + c = 0$$

$$a = -\frac{F}{\sin^3 i} \qquad b = -\frac{2FM_c}{\sin^3 i} \qquad c = -\frac{FM_c^2}{\sin^3 i}.$$

The equation

$$M^3 + aM^2 + bM + c = 0$$

$$a = -\frac{F}{\sin^3 i} \qquad b = -\frac{2FM_c}{\sin^3 i} \qquad c = -\frac{FM_c^2}{\sin^3 i}.$$

has three roots but the solution of interest for this problem is given by the real root

$$M(F, M_c, i) = \left(R + \sqrt{Q^3 + R^2}\right)^{1/3} + \left(R - \sqrt{Q^3 + R^2}\right)^{1/3} - \frac{a}{3}$$

$$R \equiv \frac{1}{54}(9ab - 27c - 2a^3) \qquad Q \equiv \frac{1}{9}(3b - a^2).$$

If F is measured,

- M is a function of two unknowns
  - the *mass of the visible companion* M_c and
  - the *tilt angle* i for the orbit.

If these can be measured or estimated in some way, the above equation provides the mass M of the unseen component, or at least a constraint on it.

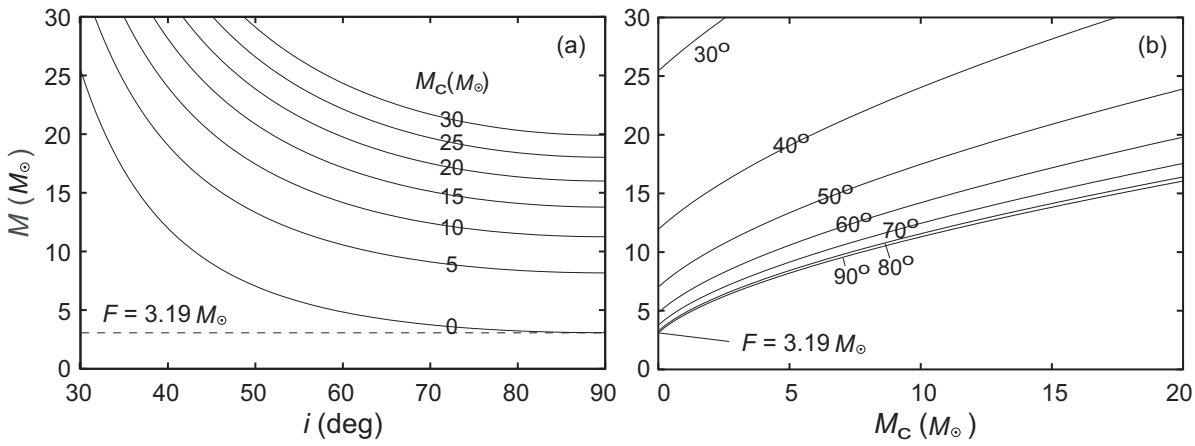


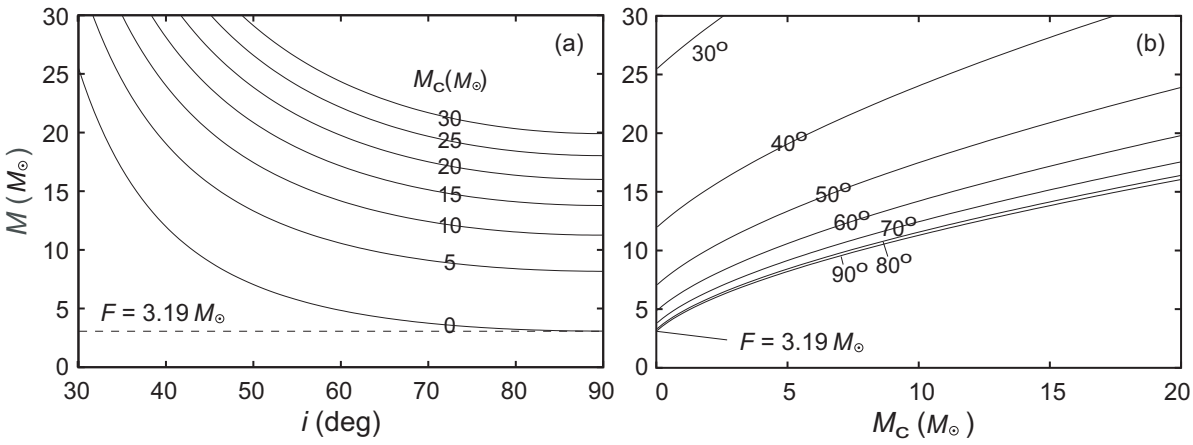
Figure 14.3: Mass plots for A0620–00 with a measured $F = 3.19\,M_\odot$. (a) Mass M of the unseen component versus the tilt angle i for different values of the companion mass M_c. (b) Mass M of the unseen component versus M_c for different values of i.

14.4.2 An Example: A0620–00

The *soft X-ray transient* A0620–00

- is a binary in which a *class K main sequence star* is transferring mass to an accretion disk around an *unseen compact object*.
- A velocity curve was shown earlier.
- The orbital period is $P = 7.75 \pm 0.0001$ hr and the semi-amplitude of the velocity curve is $K = 457 \pm 8\,\mathrm{km\,s^{-1}}$.
- From this the observed value of the mass function for A0620–00 is $F = 3.19\,M_\odot$.

In Fig. 14.3 the solution $M(F, M_c, i)$ from the cubic equation is plotted as a function of i and M_c for $F = 3.19\ M_\odot$.



These figures illustrate clearly

- the *degeneracy of the unknown mass* M in the parameters i and M_c, and
- that the measured value of the mass function $F = 3.19\,M_\odot$ is the *minimum possible mass for the unseen component.*

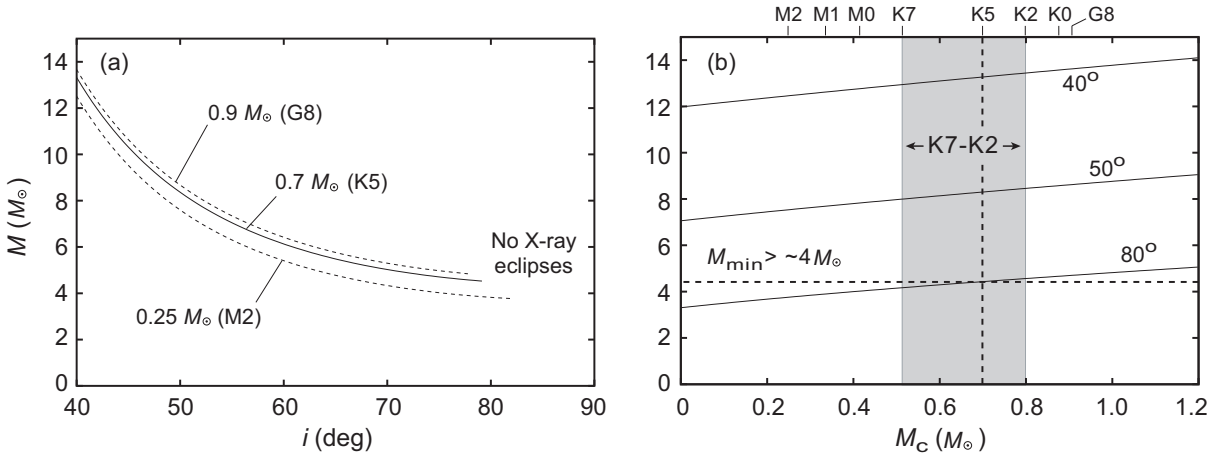
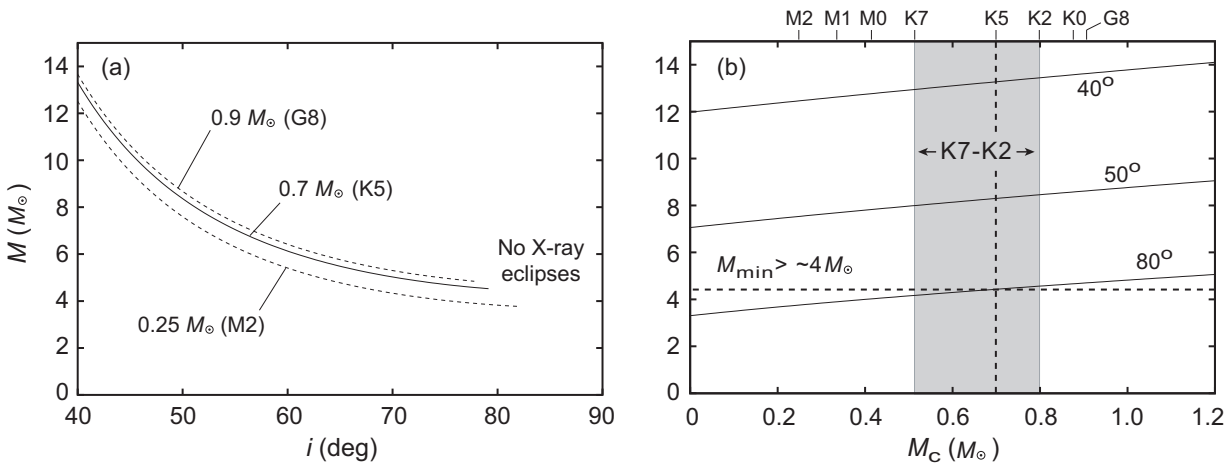


Figure 14.4: Mass plots for A0620–00 with a measured $F = 3.19\,M_\odot$ and constraints on M_c and i. (a) Mass M of the compact object vs. tilt angle i for various companion masses M_c. From systematics, the companion is of spectral class K5, with $M_c \sim 0.7\ M_\odot$. A region on the right is excluded by absence of X-ray eclipses. (b) Mass M vs. M_c for different values of i. Most likely spectral class K5 indicated by the dashed vertical line, the uncertainty range K7–K2 is indicated by the shaded region, and the minimum implied mass $M_{\min}$ for the compact object is indicated by the horizontal dashed line.

This already is a powerful constraint but a more precise statement about M is possible if *further information* can be obtained about M_c and i. Neither is measured directly but

- The *distance to the system* is known to be ~ 770 pc and
- *no eclipses of the X-ray source* are observed, which by geometry excludes values of i for a range too near $\frac{\pi}{2}$.
- The companion is a main sequence star with spectral class lying in the range K7–K2; from stellar systematics this requires the mass to lie in the range $0.5\,M_\odot < M_c < 0.8\,M_\odot$.

In Fig. 14.4 plots are repeated with these constraints displayed.



- In a more extensive analysis employing *additional observational information and systematics*, it was concluded that

$$4.16 \pm 0.01\, M_\odot \leq M \leq 5.15 \pm 0.015\, M_\odot.$$

- Since it is expected that *no neutron star can have a mass larger than* $2-3\, M_\odot$,

We deduce that the unseen companion *must be a black hole*.

Table 14.1: Some black hole candidates in galactic X-ray binaries

X-ray source	Period (days)	$f(M)$	$M_c(M_\odot)$	$M(M_\odot)$
Cygnus X-1	5.6	0.24	24–42	11–21
V404 Cygni	6.5	6.26	$\sim$0.6	10–15
GS 2000+25	0.35	4.97	$\sim$0.7	6–14
H 1705–250	0.52	4.86	0.3–0.6	6.4–6.9
GRO J1655–40	2.4	3.24	2.34	7.02
A 0620–00	0.32	3.18	0.2–0.7	5–10
GS 1124–T68	0.43	3.10	0.5–0.8	4.2–6.5
GRO J0422+32	0.21	1.21	$\sim$0.3	6–14
4U 1543–47	1.12	0.22	$\sim$2.5	2.7–7.5

Table 14.1 illustrates some candidate binary star systems where a *mass function analysis* suggests

- an unseen companion *too massive to be a white dwarf or neutron star*.
- We assume these to be *black holes*.

14.4.3 Another Example: Cygnus X-1

The first-discovered and most-storied stellar black hole candidate is *Cygnus X-1*.

1. In the early 1970s an X-ray source was discovered in Cygnus and designated Cygnus X-1.
2. In 1972, a radio source was found in the same area and identified optically with the blue supergiant HDE226868.
3. Correlations in the
   - radio activity of HDE226868 and
   - X-ray activity of Cygnus X-1

   implied that the two were probably *components of the same binary system*.
4. Doppler measurements of radial velocity for HDE226868 and other data confirmed that Cyg X-1 was a *member of a binary with a period of 5.6 days*.
5. Analysis showed that the X-ray source was *fluctuating in intensity* on timescales as short as 1/1000 of a second.
   - Signals controlling the fluctuation are limited to $v \leq c$,
   - so the source must be *very compact*, probably *no more than hundreds of kilometers in diameter*.

These observations indicate that Cyg X-1 is a *very compact object* (neutron star, or black hole).

5. The mass of the blue supergiant HDE226868 was estimated from known properties of such stars (it is a spectrum and luminosity class O9.7Iab star).

   - This, coupled with a mass function analysis, can be used to estimate the mass of the unseen, compact companion.
   - These estimates are uncertain because the geometry (tilt of the binary orbit) can only partially be inferred from data.
   - Such estimates place a lower limit of about $10M_\odot$ on the unseen companion and more likely indicate a mass near $15M_\odot$.

6. Since we know of no conditions that would permit a neutron star to exist above about $2-3M_\odot$ (or a white dwarf above about $1.4M_\odot$), we conclude that *the unseen companion must be a black hole*.

Although this chain of reasoning is indirect, it builds a very strong case that *Cygnus X-1 contains a black hole*.

14.5 Supermassive Black Holes in the Cores of Galaxies

Many observations of star motion near the centers of galaxies indicate the presence of large, unseen mass concentrations. The most direct is for our own galaxy.

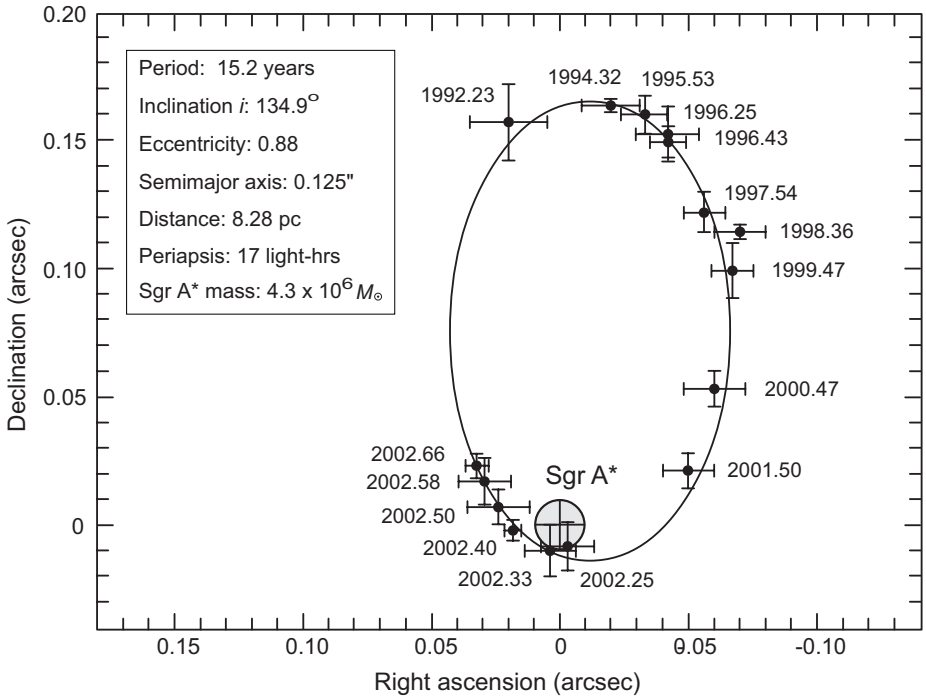
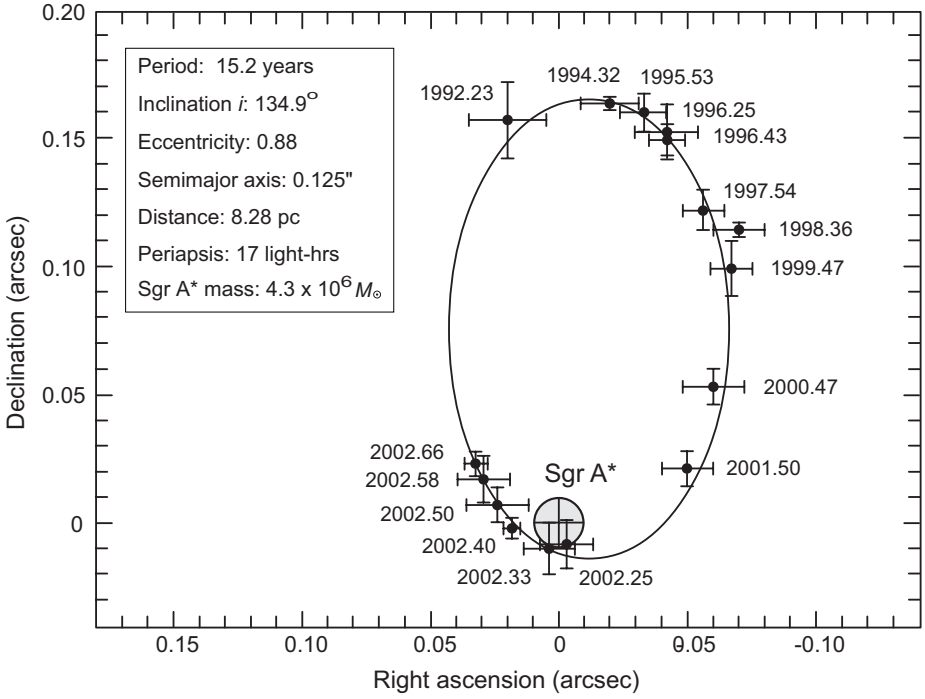


Figure 14.5: Orbit of S0-2 around the radio source Sgr A* through 2002. Filled circle indicates the position uncertainty for Sgr A* as determined from the elliptical orbit assuming a point mass located at the focus to be responsible for the orbital motion. The star completed this orbit in 2008 and the parameters displayed in the box are those obtained from the completed orbit. *Periapsis* is the general term for closest approach of an orbiting body to the center of mass about which it is orbiting.

14.5.1 The Black Hole at the Center of the Milky Way

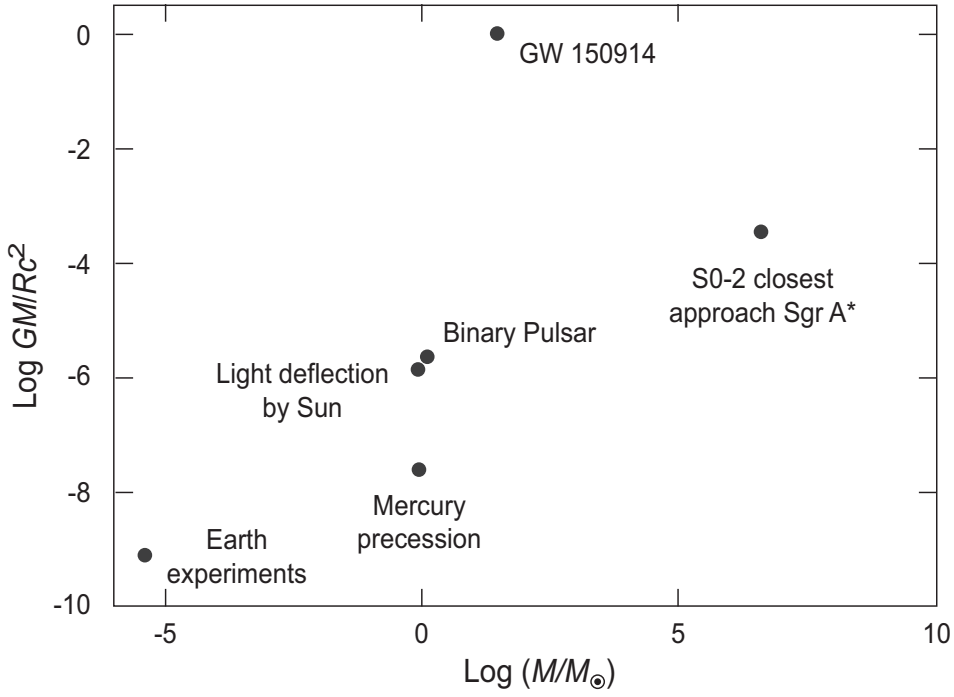
The star *S0-2* (also called S2) orbits in the vicinity of the strong radio source SGR A*, near the center of the Milky Way.

- The motion of S0-2 has been measured systematically.
- The orbit of S2 (the projected ellipse with SGR A* at a focus) is shown in Fig. 14.5.



- Closest approach is about *17 lighthours* from SGR A*.
- The *mass inside the orbit* is estimated to be $4.3 \times 10^6 M_\odot$.
- This mass is concentrated in a region *no larger than the Solar System* that contains little luminous mass.
- The simplest explanation is that the radio source SGR A* coincides with a ~ 4 *million solar mass black hole*.
- SO-2 comes within 17 lighthours of the black hole
  - This distance is still well *outside the tidal distortion radius* for the star (which is about 16 lightminutes).
  - It is about *1500 times larger than the event horizon*.
  - At closest approach the separation is not much larger than the radius of the Solar System.

- Constraints from orbits of other stars near Sgr A* place an upper limit of 45 AU on the size of the $4.3 \times 10^6 M_\odot$ source of the gravitational field.
- *Long-baseline radio interferometry* constrains the region to a size *comparable to the event horizon* of a $4.3 \times 10^6 M_\odot$ black hole.



The gravity experienced by a star like S0-2 is approximately

- *100 times stronger* than the largest gravitational fields found in the Solar System or in binary pulsars.
- Characteristic field strength GM/Rc^2 vs. source mass M for some tests of GR are summarized in the figure above.
- Stars near Sgr A* can test GR in much stronger gravity than in any context other than black hole mergers.
- General effects that can be measured are associated with *curvature (orbital precession)* and *gravitational redshifts*.

Tests of more exotic aspects of GR such as the *no hair theorem* have also been suggested.

14.5.2 The Water Masers of NGC 4258

The galaxy NGC 4258

- is an Sb spiral but its nucleus is moderately active and
- it is also classified as a Seyfert 2 Active Galactic Nucleus (AGN).

It is one of the nearest AGNs.

- A set of *masers* (microwave analog of lasers) has been observed in the central region of the galaxy.
- The maser emission in NGC 4258 is due to clouds of heated water vapor, so these are termed *water masers*.
- Because
  - masers produce *sharp spectral lines* (allowing precise Doppler shifts), and
  - *microwaves are not strongly attenuated by the gas and dust* near the nucleus of the galaxy,
- observation of the water masers has permitted the motion of gas near the center to be *mapped very precisely* using the Very Long Baseline Array (VLBA).

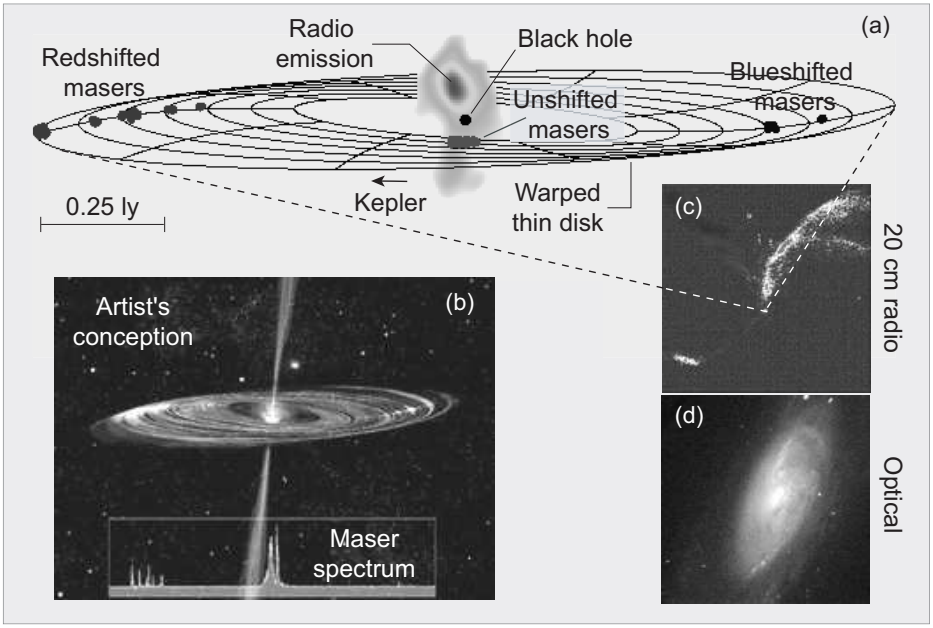
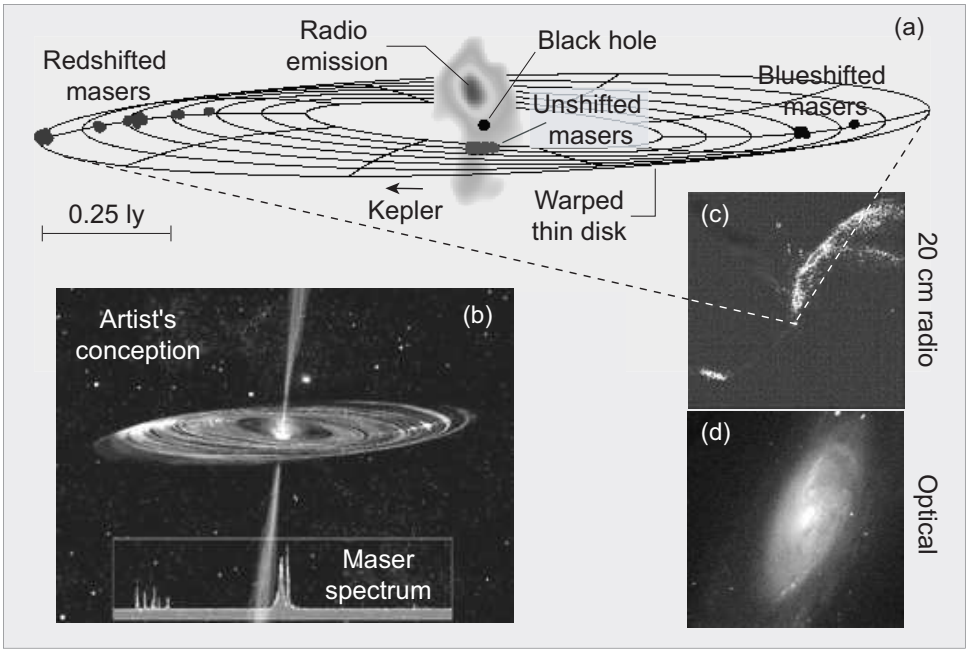


Figure 14.6: Water masers in NGC 4258 and evidence for a black hole.

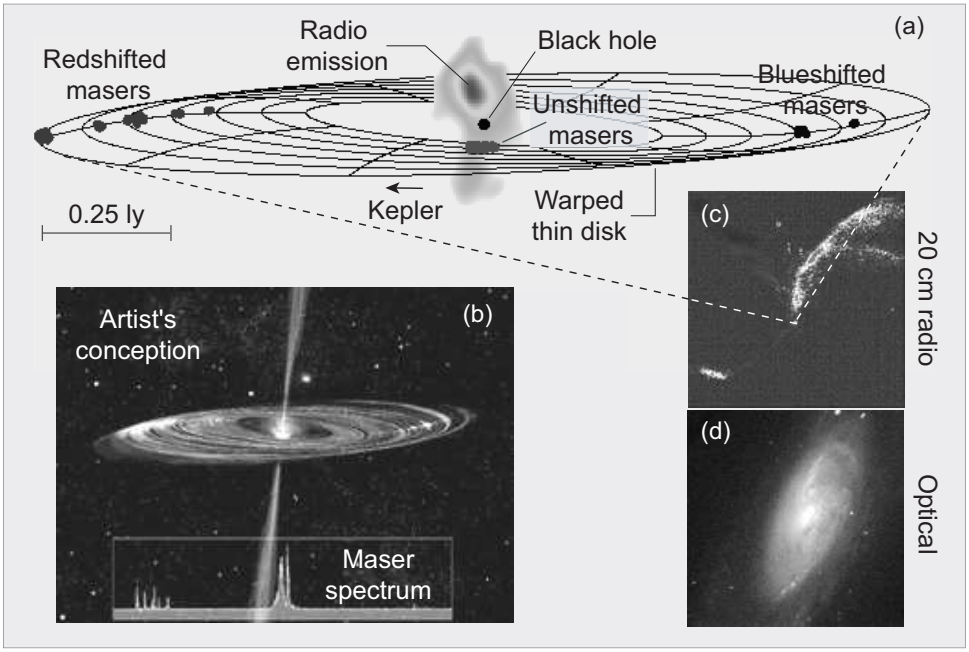
Water maser emission from the central region of NGC 4258 is illustrated in Fig. 14.6.

- The masers are mapped with a precision of better than 1 milliarcsecond and their radial velocities indicate
- bulk motion of gas at velocities near 1000 km s^{-1}.
  - masers approaching us are labeled blueshifted,
  - masers receding from us are labeled redshifted, and
  - those with no net radial velocity are labeled unshifted.

From this we may infer the *general rotation of the disk*, as indicated by the arrow labeled "Kepler".



- The masers are embedded in a thin, warped, dusty, molecular gas disk revolving around the core.
- The Keplerian motion of the masers implies that the masers are in orbit around a large mass completely contained within all their orbits.
- The mass required is approximately $3.9 \times 10^7 M_\odot$.
- Measured mass and size of region enclosed by maser orbits implies minimum density of $10^8 M_\odot$ per cubic ly.
  - 10,000 times more dense than any known star cluster.
  - The only plausible explanation is a *supermassive black hole*.



- The nucleus of the galaxy produces radio jets that appear to come from the dynamical center of the rotating disk and are approximately perpendicular to it (see fig).
- The location of the black hole engine is determined within the uncertainty of the black circle shown in the figure.
  - This black circle denotes the *uncertainty in location* of the black hole, *not its size*.
  - The black circle is about 0.05 ly in diameter, but a supermassive black hole would have an *event horizon hundreds of times smaller than this*.

These results make NGC 4258 one of the strongest cases known for a supermassive black hole engine at the core of an active galactic nucleus.

14.5.3 Virial Methods and Central Masses

For distant galaxies we have insufficient telescopic resolution to track individual stars.

- However, we may still learn something about the mass contained within particular regions by observing the *average velocities of stars* in that region.
- On conceptual grounds, we may expect that
  - the larger the gravitational field that stars feel,
  - the faster they will move.
- This intuitive idea may be quantified by using the virial theorem to show that
- the mass M responsible for the gravitational field in which stars move in some spherical region of radius R is given by

$$M \simeq \frac{5R\sigma_{\rm r}^2}{G},$$

where the radial velocity dispersion

$$\sigma_{\rm r}^2 = \langle v_{\rm r}^2 \rangle \equiv \frac{1}{3N}\sum_{i=1}^{N} v_i^2$$

can be determined by

- averaging over the squares of radial velocity fields determined from Doppler-shift measurements.

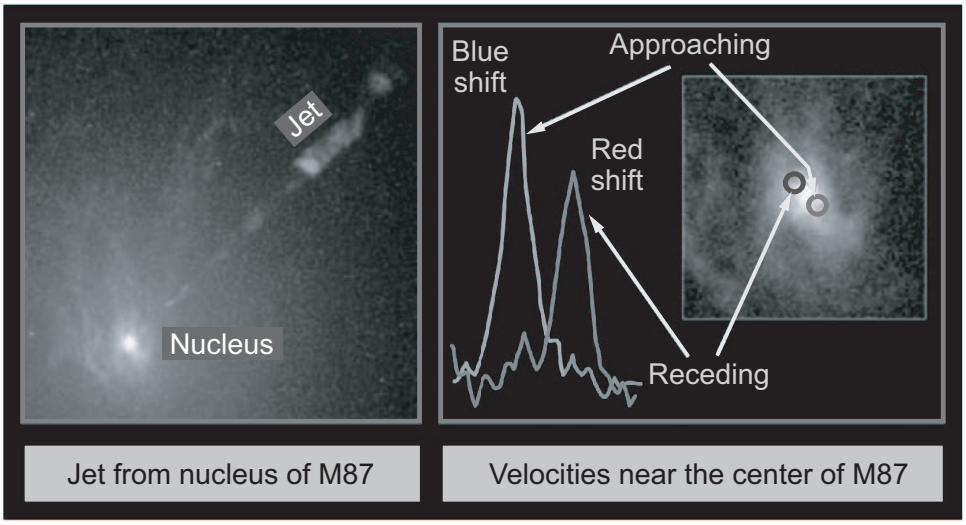
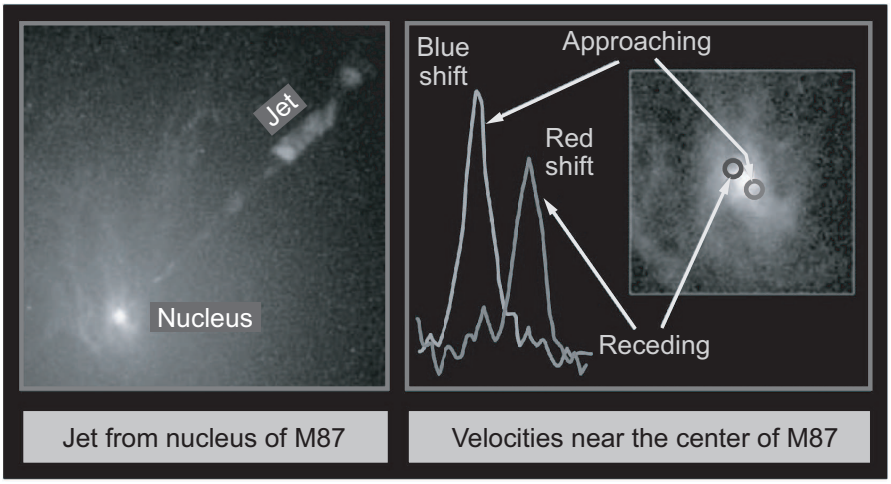


Figure 14.7: Evidence for a central black hole in the galaxy M87.

14.5.4 Evidence for a Supermassive Black Hole in M87

The left portion of Fig. 14.7 is a Hubble Space Telescope image showing the center of the giant elliptical radio galaxy M87.

- The diagonal line emanating from the nucleus in the left image is a jet of high-speed electrons—a synchrotron jet—approximately 6500 light years (2 kpc) long.
- This is what would be expected for matter swirling around the supermassive black hole, with part of it falling forever into the black hole and part of it being ejected in a high-speed jet.
- The right side of Fig. 14.7 illustrates Doppler shift measurements made on the central region of M87 that suggest rapid motion of the matter near the center.



- The measurement studied how the light near the center is *shifted by the Doppler effect*.
- The gas on one side of the disk is moving away from Earth at a speed of about 550 kilometers per second *(redshift)*.
- The gas on the other side of the disk is approaching the Earth at the same speed *(blueshift)*.
- These high velocities suggest a gravitational field produced by a *huge mass concentration* at the center of M87.
- From the *virial theorem* and velocity dispersions,
  - approximately *3 billion solar masses* are concentrated in a region at the galactic core
  - that is only about the *size of the Solar System*.
- This mass is far larger than could be accounted for by the visible matter there.
- The simplest interpretation is that a *3 billion solar mass black hole* lurks in the core of M87.

14.6 Intermediate-Mass Black Holes

Some evidence exists for black holes of mass *intermediate between*

- *stellar black holes* ($\sim 10 - 100 M_\odot$) and
- *galactic black holes* ($\sim 10^6 - 10^9 M_\odot$).

A *pulsar* has been discovered orbiting an *unseen mass concentration* in the globular cluster 47 Tucanae.

- The precise timing of the pulsar indicates that the magnitude of the unseen mass concentration is

$$M = 2200^{+1500}_{-800} M_\odot.$$

- This may be the first solid evidence for an *intermediate-mass black hole*.

There is *no electromagnetic signal* from this object, so if it is a black hole it *must not be accreting*.

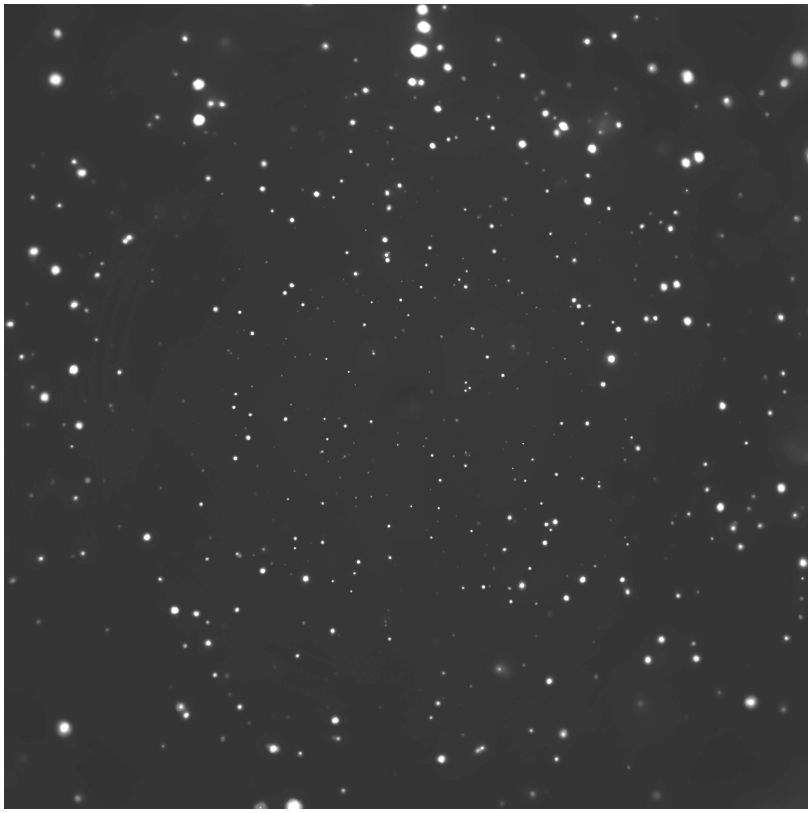
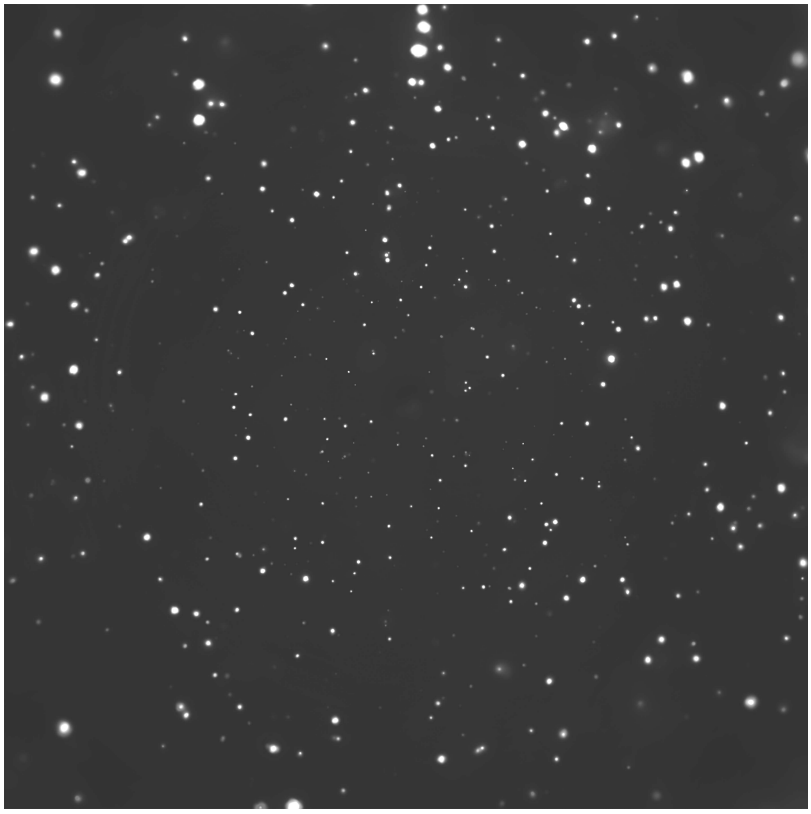


Figure 14.8: Chandra Deep Field–South. Each point lies at a redshift $z \geq 3.5$ and is a strong X-ray source marking a supermassive black hole, or a galaxy containing multiple X-ray binaries with accreting black holes.

14.7 Black Holes in the Early Universe

A comprehensive survey of black holes at large redshift is displayed in Fig. 14.8.

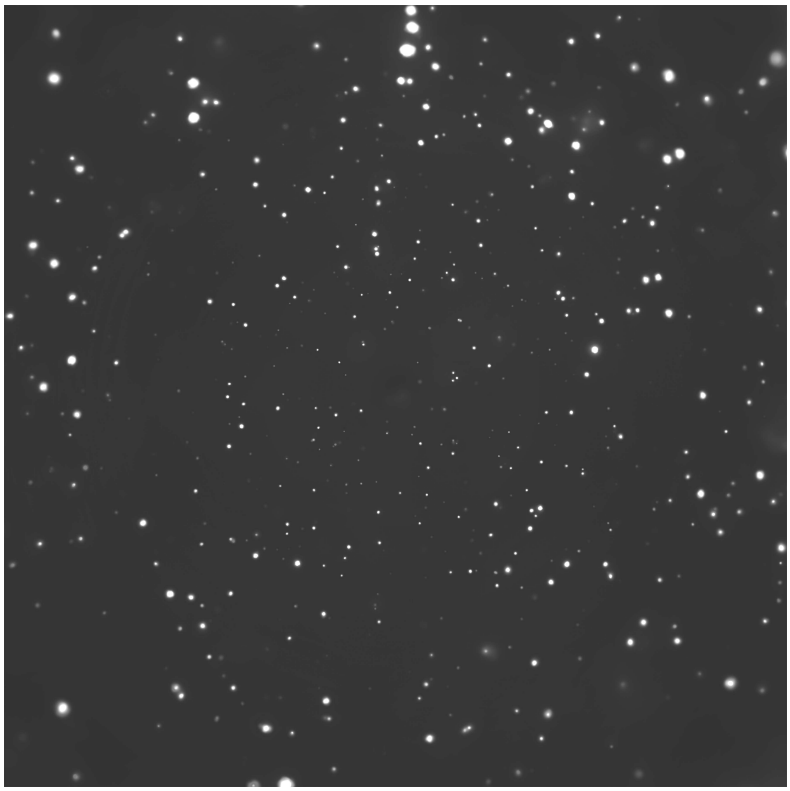
- This is a composite image of *Chandra X-ray Observatory* data having an angular extent somewhat less than the diameter of the full Moon;
- it is the *deepest X-ray image ever obtained*.



The black holes are of course invisible in this image but

- X-rays produced by gas being accelerated as it accretes mark the position of black holes.
- The central part of this image is thought to contain about 5000 black holes.

This implies $\sim 10^9$ black holes potentially visible in X-rays out to this redshift.



The study combined

- long-exposure Chandra X-ray data with
- very deep field optical images from the Hubble Space Telescope for the same region of the sky.

X-ray emission was studied from more than 2000 galaxies at redshifts $3.5 < z < 6.5$.

- About 70% of the X-ray objects are thought to be *super-massive black holes* with $M \sim 10^5 - 10^{10} M_\odot$.
- The remainder are thought to be *stellar-size black holes in X-ray binaries* hosted by unresolved galaxies.

This deep X-ray survey reached *two conclusions of potential importance* for the role of supermassive black holes in the evolution of the Universe.

1. The *preferred seeds* for the growth of supermassive black holes were black holes of mass $10^4\,M_\odot$ to $10^5\,M_\odot$.
2. The growth of supermassive black holes in the first several billion years occurred in *sporadic episodes*.

These conclusions are significant if they can be corroborated.

- The first favors the *"heavy seeds" model* of supermassive black hole formation where
  - intermediate-mass black holes form by *direct collapse of gas clouds*,
  - or possibly by *rapid merger of massive first-generation stars*,
- The second suggests that the rapid growth of supermassive black holes in the early Universe may have been greatly assisted by galaxy collisions and mergers, during which
  - central black holes grew at high rates, with
  - much slower growth during intervening periods.

However, this is but one set of data and the mechanism by which supermassive black holes form remains an open question.

14.8 Show Me an Event Horizon!

Compelling evidence exists for objects exhibiting many of the features expected for a black hole.

- Certainly these objects would be *difficult to explain* by any other hypothesis.
- But the essential observational characteristic of a black hole is its *event horizon,* which should have properties *unlike anything else encountered in the Universe*.
- Therefore, rigorous proof that black holes exist will require *"imaging" an event horizon*.
- This is challenging: black holes emit no light directly and are very small, and no nearby black holes are known.
- The best prospects are for the approximately 4 million solar mass black hole Sgr A* at the center of the galaxy.
- The task is to resolve an object
  - *obscured by dust* that is
  - less than *20 times the solar diameter*
  - at a *distance of* 8 kpc.

Very Long Baseline Interferometry (VLBI) with radio telescopes spread across the globe can *penetrate the dust* and *achieve such resolution*, so imaging the Sgr A* event horizon may be possible soon.

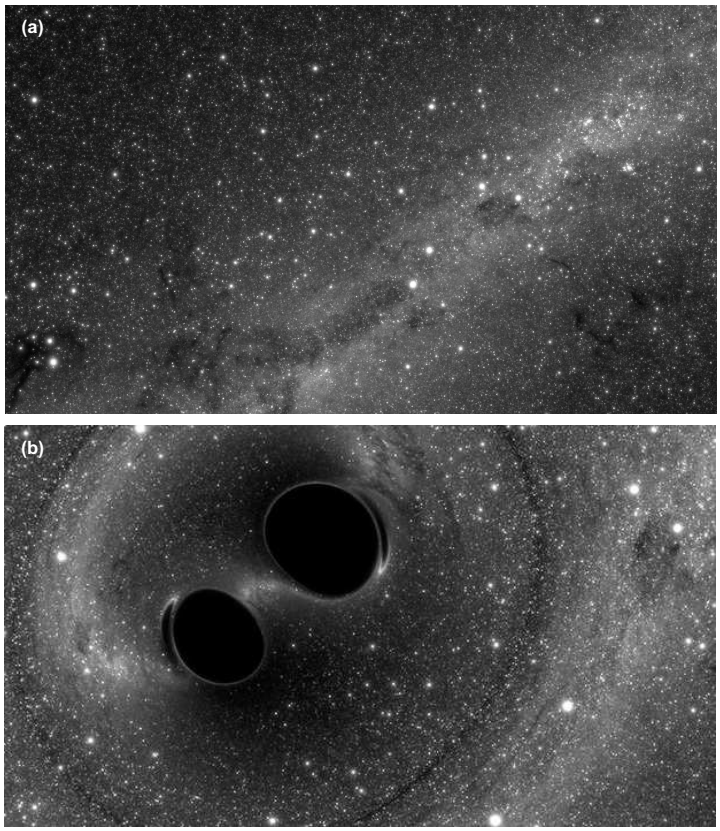
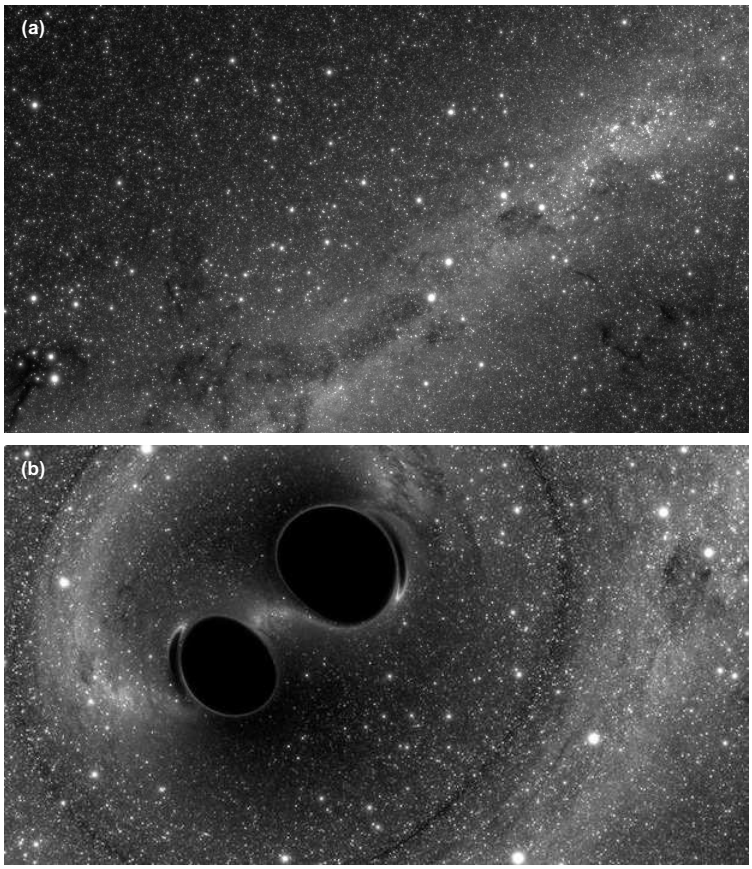


Figure 14.9: Computer simulation of how the two black holes might have appeared to a nearby observer just prior to merger in the gravitational wave event GW150914, which will be described in later chapters. (a) Background stars without the black holes. (b) Image including black holes.

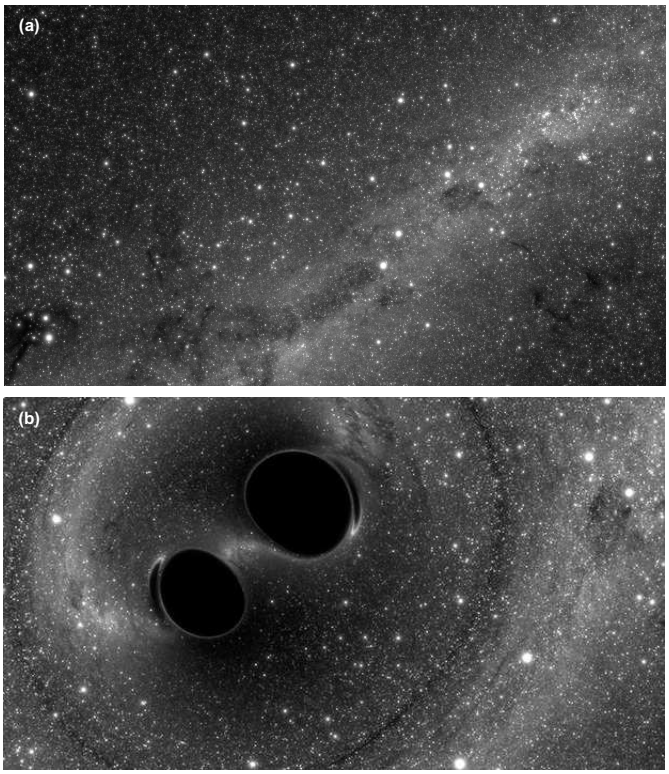
What would we expect to see that would identify an object unequivocally as the event horizon of a black hole?

- From *analysis of the gravitational wave* GW150914 (merger of $29\,M_\odot$ and $36\,M_\odot$ black holes)
- a *computer simulation* of the merger has been constructed.

A *frame from the simulation* is shown in Fig. 14.9.

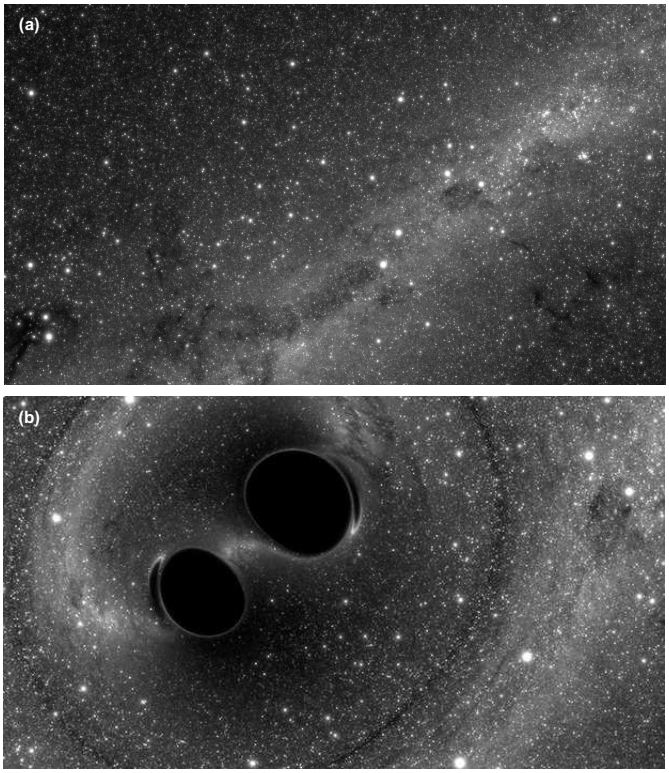


- All *stars are in the distant background* and are *unaffected by the black holes*,
- but *gravitational lensing* severely distorts the path of their light and hence apparent positions in (b).
- The dark, well-defined shapes are the *shadows of the event horizons* blocking all light from behind.
- Flattened dark features and the marked displacement of the apparent background star images are caused by *gravitational lensing* arising from the strong curvature.



- The ring around the black holes is an *Einstein ring* (gravitational lensing of light from stars behind the black holes).
- For the above simulation the black holes are
  - assumed to be isolated, with any surrounding material
  - already having been accreted by the black holes,

  so all other objects are distant stars in the background.
- Hence the image is dominated by
  - *shadowing by horizons* of light from background,
  - and by extremely strong *gravitational lensing effects*.



- The environment near Sgr A* is not likely to be as pristine:
  - The black hole is probably *accreting matter* and
  - *emitting radiation* from accretion.
- We might expect the dominant feature to be
  - complete, sharply-defined *shadowing by the horizon*,
  - strongly *distorted by gravitational lensing*.

However, it is unknown how the local environment of Sgr A* will alter this picture.

14.9 Summary: A Strong But Circumstantial Case

The evidence cited in this chapter is not yet iron-clad proof of the existence of black holes.

- The defining characteristic of a black hole is an event horizon.
- No known black hole candidates have had their event horizons imaged,
- Though angular resolution sufficient to resolve the event horizon of the Milky Way's central black hole may be imminent.

Nevertheless, extensive data on

- *X-ray binary systems* with massive companions.
- The *motions of stars and other objects* in the central regions of galaxies.
- Even more direct evidence from analysis of *gravitational waves generated by black hole mergers* that will be discussed in later chapters.
- The *enormous power sources* to be discussed in the next chapter, which are hard to explain without black holes

provide *extremely strong circumstantial evidence* for the existence of black holes.

Chapter 15

Black Holes as Central Engines

Black holes imply a fundamental modification of our understanding of space and time.

- But at a more mundane level they also are of *great practical importance* in astronomy because
- they can be extremely *efficient sources of energy*.

As we shall now discuss, rotating black holes are believed to be the engine powering a whole set of phenomena associated with

- quasars,
- active galaxies,
- gamma-ray bursts, . . .

15.1 Black Holes as Energy Sources

Black holes have *event horizons* that cut their interiors off from the outside world, so they might seem unlikely power sources.

1. However, the discussion of *Penrose processes* demonstrates conceptually that it is possible to extract energy from rotating classical black holes.
2. Penrose processes themselves are not likely to lead to large extraction of energy from black holes in astrophysical environments.

However, two classes of phenomena can lead to

- large energy release for black holes
- in regions *outside the event horizon*,

where released energy is accessible to external processes:

1. *Accretion onto Schwarzschild or Kerr black holes* can convert large amounts of gravitational energy into other usable forms of energy *before the matter falls through the event horizon*.
2. Rotating black holes may be accompanied by *strong external magnetic fields* that can
    - *tap the rotational power* of the black hole to
    - *accelerate charged particles in relativistic jets* along the axes of rotation, *outside the event horizon*.

A consideration of plausible *formation mechanisms for black holes* suggests that

- they are likely to form with *some angular momentum*, and that
- they are likely to have *no net charge*.
- Thus practically we may focus our attention on *Kerr black holes*.
- The *Schwarzschild solution* may then be viewed as
  - the *limit of the Kerr solution* for
  - for the special case of *vanishing angular momentum*.

Let us now turn to an overview of extracting energy from rotating black holes, either by *accretion*, or by *coupling of rotation to external fields*.

15.2 Accretion and Energy Release for Black Holes

Accretion of matter can be a large source of power because it is an *efficient mechanism for gravitational energy conversion*.

- This is especially true for accretion onto a *compact object:*
  - *white dwarfs*,
  - *neutron stars*, or
  - *black holes*.
- Accreting particles typically form an *accretion disk* orbiting the object (angular momentum conservation).
- Collisions of the particles in the accretion disk
  - *heat it* and
  - cause it to *radiate energy away*,

  allowing particles to *spiral inward and accrete*.
- For white dwarfs or neutron stars, the inspiraling material accumulates on the surface.
- For black holes there is *no surface* but the event horizon clearly sets a boundary for energy extraction.
- Calculations indicate that large amounts of gravitational energy can be extracted from in-spiraling matter before it falls through a black hole horizon, either through
  1. *Emission of radiation*, or through
  2. Production of *relativistic jets*.

Table 15.1: Newtonian gravitational energy for hydrogen accretion

Accretion onto	Max energy released (erg g^{-1})	Ratio to fusion
Black hole	1.5×10^{20}	25
Neutron star	1.3×10^{20}	20
White dwarf	1.3×10^{17}	0.02
Normal star	1.9×10^{15}	10^{-4}

15.3 Maximum Energy Release in Spherical Accretion

The most spectacular consequence of accretion is that it is an efficient mechanism for extracting gravitational energy.

- The energy released by accretion is approximately

$$\Delta E_{\text{acc}} = G\frac{Mm}{R},$$

  where M is the mass of the object, R is its radius, and m is the mass accreted.

- In Table 15.1 the energy released per gram of hydrogen accreted onto the surface of various objects is summarized.

- From Table 15.1, we see that accretion onto very compact objects is a much more efficient source of energy than is hydrogen fusion.

- But accretion onto normal stars or even white dwarfs is much less efficient than converting the equivalent amount of mass to energy by fusion.

Assume for the moment that

- all kinetic energy generated by conversion of gravitational energy in accretion is radiated from the system.
- Then the *accretion luminosity* is

$$L_{\rm acc} = \frac{GM\dot{M}}{R} \simeq 1.3 \times 10^{21} \left(\frac{M/M_\odot}{R/{\rm km}}\right) \left(\frac{\dot{M}}{{\rm g\,s^{-1}}}\right) {\rm erg\,s^{-1}},$$

  if a steady accretion rate $\dot{M}$ is assumed.

> We shall address the issue of *efficiency for realistic accretion* shortly)

Table 15.2: Some Eddington-limited accretion rates

Compact object	Radius (km)	Max accretion rate ($\mathrm{g\,s^{-1}}$)
White dwarf	$\sim 10^4$	10^{21}
Neutron star	~ 10	10^{18}

15.3.1 Limits on Accretion Rates

The *Eddington luminosity* is

$$L_{\mathrm{edd}} = \frac{4\pi G M m_{\mathrm{p}} c}{\sigma},$$

with σ ithe effective cross section for photon scattering.

- For fully ionized hydrogen, we may approximate σ by the Thomson cross section to give

$$L_{\mathrm{edd}} \simeq 1.3 \times 10^{38} \left(\frac{M}{M_\odot}\right) \mathrm{erg\,s^{-1}}.$$

- If L_{edd} is exceeded (luminosity is *super-Eddington*), accretion will be *blocked by the radiation pressure*.
- This implies that there is a *maximum accretion rate*.
- Equating L_{acc} and L_{edd} gives for this maximum rate

$$\dot{M}_{\mathrm{max}} \simeq 10^{17} \left(\frac{R}{\mathrm{km}}\right) \mathrm{g\,s^{-1}}$$

Eddington-limited accretion rates calculated from this formula are given in Table 15.2.

15.3.2 Accretion Efficiencies

- For gravitational energy released by accretion to be extracted,
  - it must be radiated or
  - matter must be ejected at high kinetic energy (for example, in AGN jets).
- Generally, only a fraction of the potential energy available from accretion can be extracted to do external work.
- This issue is particularly critical when black holes are the central accreting object, since
  - they have no accretion "surface" and
  - the horizon makes energy extraction problematic.
- Let's modify our previous equation for accretion power by introducing an *efficiency factor* η *through*

$$L_{\rm acc} = \frac{GM\dot{M}}{R} = \eta \dot{M} c^2 \qquad \eta \equiv \frac{GM}{Rc^2}.$$

- Specializing to black holes, it is logical to take *some multiple of the Schwarzschild radius*

$$r_{\rm s} = \frac{2GM}{c^2} = 2.95 \left(\frac{M}{M_\odot} \right) \text{ km},$$

to define the *"accretion radius"*.

- Then for a spherical black hole

$$L_{\rm acc}^{\rm bh} = \eta \dot{M} c^2 \qquad \eta = \frac{r_{\rm s}}{2R},$$

  and $r_{\rm s}/2R$ measures the efficiency for converting rest mass to energy by the accreting black hole power source.

- For spherical black holes, a typical choice for R is the *radius $R_{\rm ISCO} = 3r_{\rm s}$ of the innermost stable circular orbit* in the Schwarzschild spacetime.

- For hydrogen fusion, the mass to energy conversion efficiency is $\eta \sim 0.007$.

- For compact spherical objects like Schwarzschild black holes or neutron stars, reasonable estimates suggest $\eta \sim 0.1$.

- Shortly we shall see that for rotating *Kerr black holes* efficiencies of $\eta \sim 0.3$ might be possible.

15.3.3 Innermost Stable Circular Orbit and Binding Energy

A fundamental property of the Kerr metric is

- the *binding energy of the innermost stable circular orbit*,
- because this is related to the *energy that can be extracted from accretion* on a rotating black hole.

If the radius of the innermost stable orbit is denoted by R,

- for circular orbits we have that $dr/d\tau = 0$ and
- from earlier equations of motion in the Kerr metric

$$\frac{\varepsilon^2 - 1}{2} = \frac{1}{2}\left(\frac{dr}{d\tau}\right)^2 + V_{\text{eff}}(r, \varepsilon, \ell) \quad \rightarrow \quad \frac{\varepsilon^2 - 1}{2} = V_{\text{eff}}(R, \varepsilon, \ell).$$

- To remain circular the *radial acceleration must vanish*,

$$\left.\frac{\partial V_{\text{eff}}}{\partial r}\right|_{r=R} = 0,$$

- and to be a stable orbit the *potential must be a minimum*

$$\left.\frac{\partial^2 V_{\text{eff}}}{\partial r^2}\right|_{r=R} \geq 0,$$

with *equality holding for the last stable orbit*.

This set of equations may be solved to determine the *innermost stable orbit* and its *binding energy*.

For *extremal black holes* ($a = M$), we find that

- *Co-rotating orbits* (revolving in the same sense as the rotation of the hole) are more stable than the corresponding counter-rotating orbits.
- For *co-rotating orbits*,

$$\varepsilon = \frac{1}{\sqrt{3}} \qquad \ell = \frac{2M}{\sqrt{3}} \qquad R = M.$$

- The quantity ε is the *energy per unit rest mass measured at infinity*, so
- the *binding energy per unit rest mass*, B/M, is given by

$$\frac{B}{M} = 1 - \varepsilon.$$

- These equations imply that the *fraction f of the rest mass that could theoretically be extracted as energy* is

$$f = 1 - \varepsilon = 1 - \frac{1}{\sqrt{3}} \simeq 0.42.$$

  for a transition

  - from a distant unbound orbit
  - to the innermost circular bound orbit of a Kerr black hole.

Therefore, in principle *42% of the rest mass of accreted material could be extracted as usable energy* by accretion onto an extreme Kerr black hole.

- One expects *less efficiency for realistic scenarios*.
- But we might expect as much as 20–30% efficiencies in realistic accretion scenarios.
- This should be compared with the maximum theoretical efficiency for conversion of rest mass into energy of
  - 6% for *accretion on a spherical black hole* and
  - 0.7% efficiency for *hydrogen fusion to helium*.

Accretion onto rotating black holes is a *very efficient mechanism* for conversion of mass to energy.

As we shall see shortly, it is the *high mass-to-energy conversion efficiency* available from accretion onto compact objects that

- provides the most convincing argument that *active galactic nuclei (AGN) and quasars* must be powered by accretion onto rotating supermassive ($M \sim 10^9 M_\odot$) black holes.
- For example, *observed luminosities* and *temporal luminosity variation* indicate that a quasar could be powered by
  - accretion of as little as several solar masses per year
  - onto an object of mass $\sim 10^9 M_\odot$, and that
  - this mass must occupy a volume the size of the Solar System or smaller.

It is now broadly agreed that *rotating, supermassive black holes* are the only plausible energy source for quasars and active galactic nuclei.

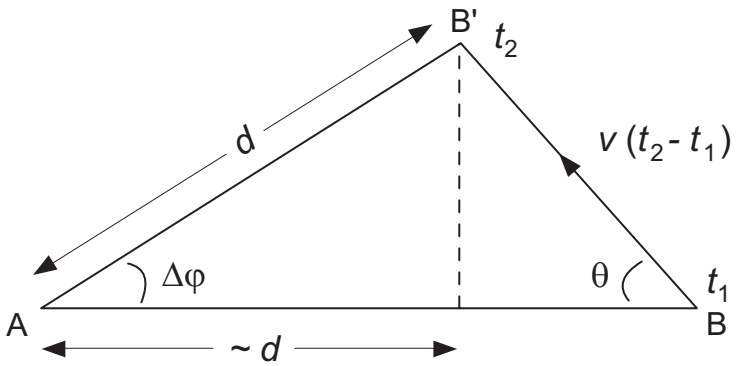
15.4 Jets and Magnetic Fields

A magnetic field cannot be anchored to the black hole itself because the event horizon pinches off magnetic field lines.

- However, the accretion disk lies outside the event horizon and it could have a strong magnetic field.
- Accretion of matter by the rotating black hole can lead to ejection at $v \sim c$ along the poles of the black hole rotation for the matter that does not cross the event horizon.
- The rotating magnetic field of the accreting disk can be carried away with the ejected matter, leading to *bipolar jets perpendicular to the accretion plane*.
- These jets contain
  - charged particles moving at near light velocity and
  - twisting and spiraling magnetic fields.
- The magnetic fields probably are essential to focusing and confining the relativistic jets into the narrow cones that are observed, but the details are not well understood.
- If the jets are not at right angles to our line-of-sight, we refer to the one pointed more toward us as the *jet* and the other as the *counterjet.*
- Because of relativistic beaming effects when relativistic jets are oriented near the line of sight, *the counterjet may be faint and difficult to see* relative to the jet.

15.5 Relativistic Jets and Apparent Superluminal Velocities

As an example of relativistic beaming, consider the following figure where a distant source at B moves with $v \sim c$ toward B′.



- At time t_1 the source at B emits a light signal that is detected at time t_1' by an observer at A.
- When the source reaches B′ at time t_2, it emits another light signal that is detected by observer A at time t_2'.
- From Lorentz invariance the *apparent transverse velocity* $\beta_T = v_T/c$ is related to the *actual velocity* $\beta = v/c$ by

$$\beta_T \equiv \frac{v_T}{c} = \frac{\beta \sin\theta}{1 - \beta\cos\theta}.$$

Thus for actual velocity $\beta < 1$, the *apparent transverse velocity can exceed c by any amount* because the maximum value

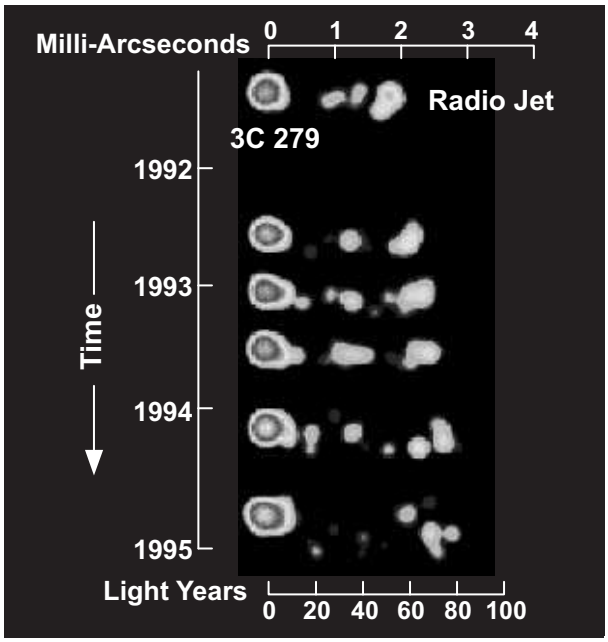
$$\beta_T^{max} = \beta/(1-\beta^2)$$

is unbounded as $\beta \to 1$.

The illusion of an apparent velocity exceeding that of light because of

$$\beta_{\rm T} \equiv \frac{v_{\rm T}}{c} = \frac{\beta \sin\theta}{1-\beta\cos\theta}.$$

is observed frequently for jets in radio astronomy where it is called *superluminal motion.* The figure below



- shows a radio jet in the blazar 3C 279 (redshift $z = 0.534$) exhibiting apparent superluminal motion,
- with an inferred transverse velocity $v \sim 7c$.

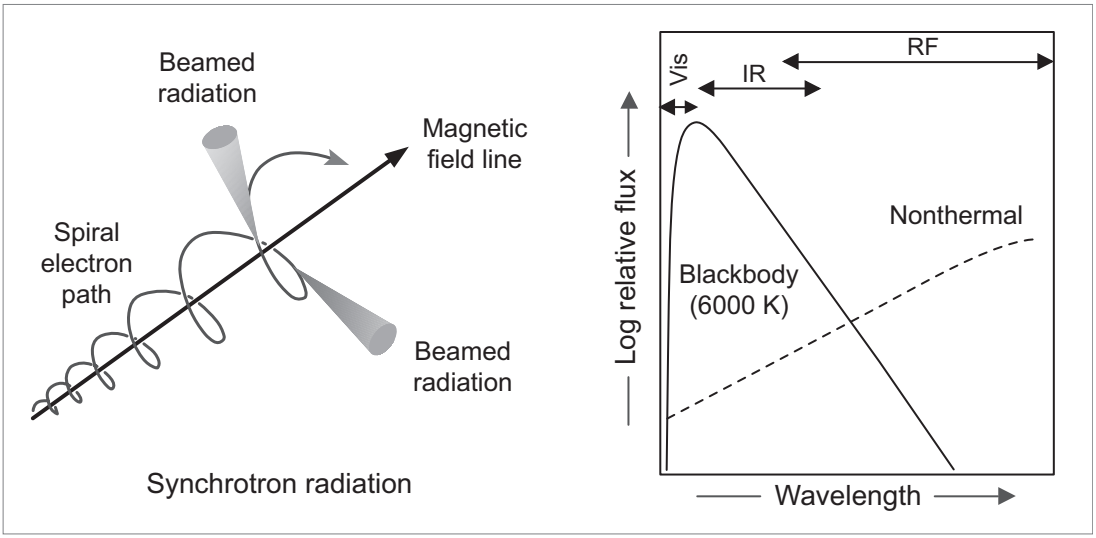


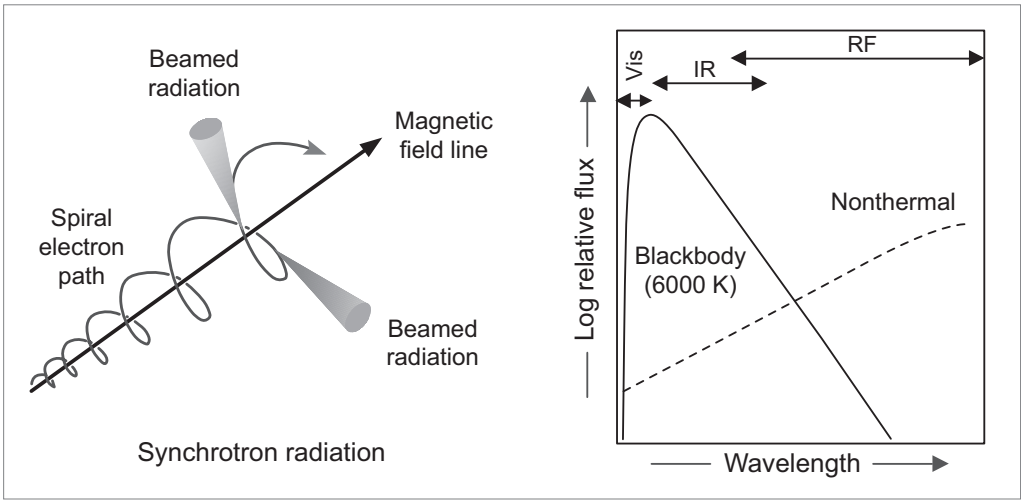
Figure 15.1: Thermal and nonthermal emission

15.5.1 Nonthermal Emission

The Planck law describes *thermal emission,* characterized by emission of radiation from a hot gas in approximate thermal equilibrium; the resulting spectrum is a *blackbody spectrum.*

- The characteristic Planck law curves for thermal emission peak at some wavelength, and fall off rapidly at longer and shorter wavelengths (curve "Blackbody" in Fig. 15.1).
- The position of the peak moves to shorter wavelength as the temperature of the gas is increased (the Wien law).
- Light from most stars, and light from normal galaxies, is dominantly thermal in character.

Sometimes we observe emission of *nonthermal* radiation, with a spectrum that increases in intensity at very long wavelengths.



The most common form of nonthermal emission in astronomy is *synchrotron radiation.*

- Created when high-velocity electrons (or other charged particles) in strong magnetic fields follow
- a spiral path around the field lines, radiating *highly-beamed and partially-polarized light* (figure above left).
- The figure above right contrasts
  - a *thermal spectrum* characteristic of 6000 K and
  - a *nonthermal spectrum*.
- The wavelength of the synchrotron radiation is related to how fast the charged particle spirals in the magnetic field.
- Thus, as the particle emits radiation, it *slows and emits longer wavelength radiation*.
- This explains the *broad distribution in wavelength of synchrotron radiation* as compared with thermal radiation.

Nonthermal emission is less common than *thermal emission*.

- However, the presence of a nonthermal component in a spectrum typically signals
  - *violent processes* and
  - *large accelerations* of charged particles.
- High-frequency synchrotron radiation also implies the presence of *very strong magnetic fields*, since
  - the frequency increases with tighter electron spirals,
  - which are characteristic of strong fields.
- The resulting synchrotron radiation has a *nonthermal spectrum* and is *partially polarized*.
- It is strongly focused in the forward direction by *relativistic beaming*.
- Fluctuations of the jet in time will also be compressed into shorter apparent periods by relativistic effects.
- For an observer in the general direction of a jet, these effects will exaggerate both the apparent intensity and the time variation of the nonthermal emission.
- The *nonthermal continuum emission* originates largely in the *synchrotron radiation produced in the jets*.
- The *thermal continuum* is typically *produced in the accretion disk* and the surrounding matter that it heats.

15.6 Quasars

In the 1950s astronomers began to catalog systematically objects in the sky that emitted radio waves.

- As these radio sources were cataloged, an effort was made to correlate the objects emitting radio waves with sources visible in optical telescopes.
- The resolution of the single-dish radio telescopes in use at the time was much poorer than that of large optical telescopes,
- so there often were many possible optical sources that might potentially be correlated with the fairly uncertain position of a radio source.
- Nevertheless, some progress was made in these identifications.

15.6.1 "Radio Stars" and a Spectrum in Disguise

Most radio sources were found to be correlated with certain galaxies.

- However, a few optical sources identified with the radio sources appeared to be points with no obvious spatial extension, as expected for stars.
- These were often called *"radio stars"*, but they had very strange characteristics for stars.
- In 1963, the first two of these radio stars were associated with the radio sources 3C48 and 3C273, respectively.
- Although these objects had the appearance of stars in optical telescopes, they had spectra unlike any stars that had ever been observed.
    1. There is a very strong continuum across a broad range of wavelengths that is non-thermal in character.
    2. Sitting on top of this continuum are emission lines, but they are very broad, and their wavelengths do not correspond to the lines for any known spectra.

Later in 1963, Dutch astronomer Maarten Schmidt found while studying the spectrum of 3C273 that the strange emission lines were really very familiar spectral lines in disguise.

- They were *lines of the Balmer series of hydrogen* (and a line in ionized magnesium), but *redshifted by a very large amount*.
- The redshift of 3C273 corresponds to a recessional velocity about 15 percent of the speed of light, if interpreted as a Doppler shift *(later we shall see that such cosmological redshifts should not be interpreted as Doppler shifts, though)*.
- Once this was realized for 3C273, it quickly became apparent that the spectrum of 3C48 could be interpreted in the same way, but with a redshift that was even larger.
- The reason that it took some time to arrive at this interpretation of the spectra for these objects is that
  1. They were thought initially to be relatively nearby stars because they seemed to be pointlike.
  2. Thus no one had any reason to believe they should be strongly redshifted.

These objects were named *quasistellar radio sources,* which was soon contracted to *quasars.*

15.6.2 Quasar Characteristics

Quasars are now known to have the following characteristics.

- They appeared to be *star-like in the initial images*, but
- more careful study shows *fuzziness and faint jets* associated with some quasars.
- Because quasars *emit strongly in the ultraviolet*, they are *distinctly blue at optical wavelengths*.
- Many of the first quasars discovered were also radio sources, but *most quasars are not strong radio emitters*.
- They exhibit a *non-thermal continuum spectrum*
  - that is *stronger than most other sources at all wavelengths*,
  - that *varies substantially in time*, and
  - exhibits the basic characteristics of *synchrotron radiation* (nonthermal and polarized).
- They usually exhibit *broad emission lines*.
- If this broadening is attributed to random motion of sources, velocities $\sim 10,000\,\text{km/s}$ are indicated by the widths of the emission lines.
- Many quasars exhibit *large redshifts*, implying by
  - the *Hubble law* that they are *at great distances*
  - and that the light that we see was emitted when the Universe was *much younger than it is now*.

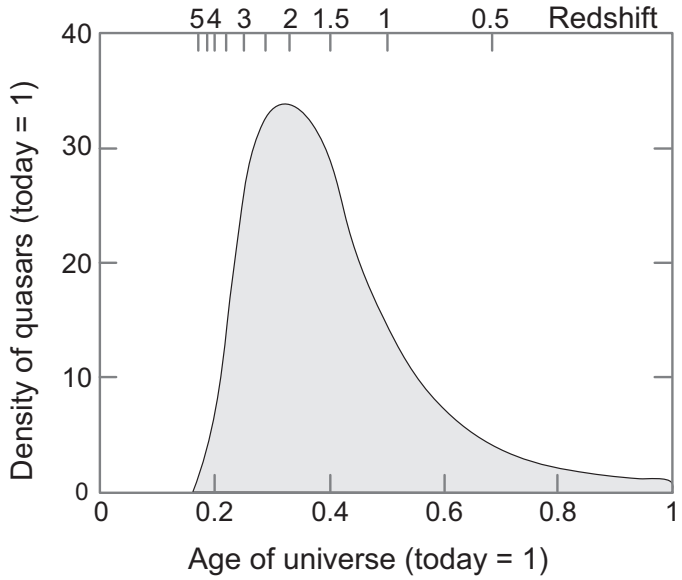


Figure 15.2: Observed quasar densities as a function of the fractional age of the Universe. We see that quasars were much more abundant earlier in the history of the Universe than they are today. The decrease of quasar abundances at the very earliest times (redshifts larger than about 3) may represent partially the finite amount of time for quasars to begin forming after the formation of the Universe and partially observational bias, since it is more difficult to detect objects at the largest distances and therefore the earliest times.

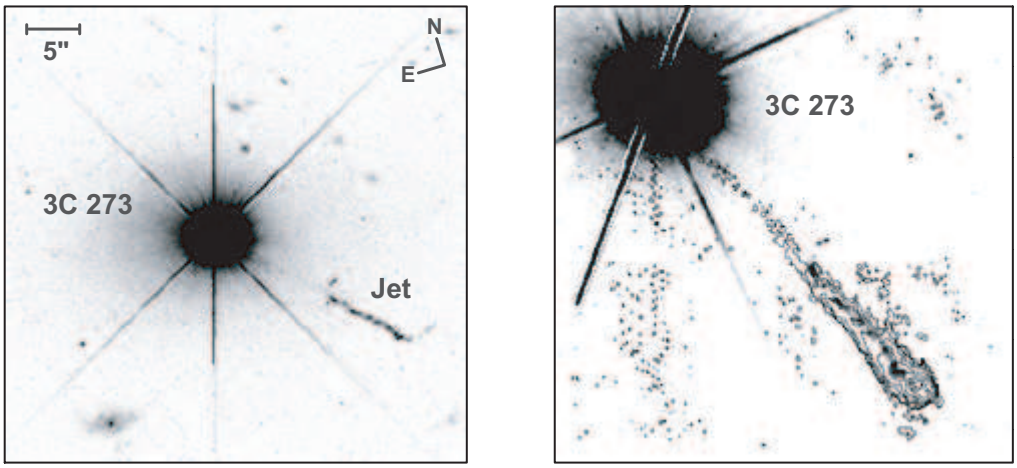
- Quasar counts as a function of redshift indicate that they were *more abundant earlier in the history of the Universe than they are today*, as illustrated in Fig. 15.2.

From their

- huge energy output,
- large redshifts, and
- jets and other structure

it is clear that quasars are not stars.

Now in favorable cases we have images showing the host galaxy of the quasar: *Quasars are highly energetic phenomena that occur in certain galaxies*.



(a) Quasar 3C 273 and jet (b) Optical jet + radio contours

Figure 15.3: (a) The quasar 3C273, which has a redshift of 0.158. The sharp radial lines are optical spike artifacts common in quasar images because of their star-like brightness. (b) Radio contours superposed on the optical jet.

Optical and radio images of the quasar 3C273 are illustrated in Fig. 15.3.

- The left image of 3C273 shows the quasar and a jet.
- The right image, which has been rotated and enlarged relative to the left image, superposes on this optical image *contours of radio frequency intensity*.
- Despite its distance, 3C273 has an apparent visual magnitude of +12.9, so it must be incredibly luminous.

When summed over all wavelengths, an average quasar like 3C273 is typically some *1000 times more luminous than a bright normal galaxy*.

15.6.3 An Implied Enormous and Compact Energy Source

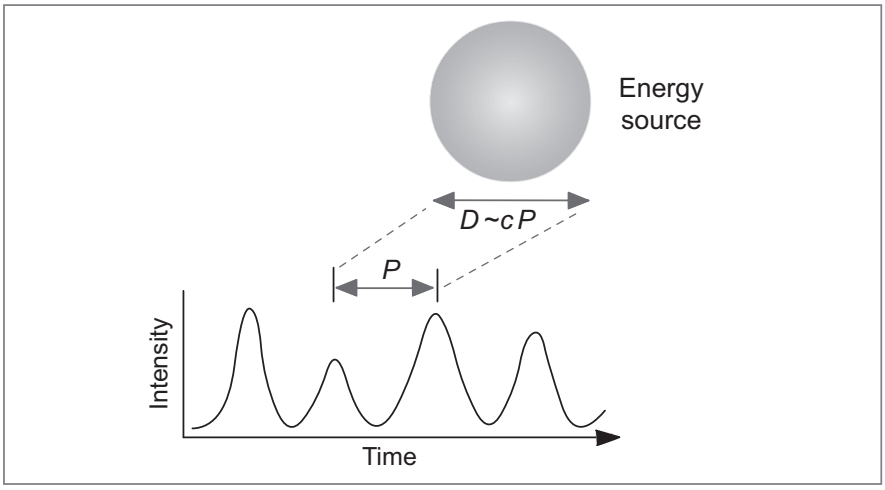
Quasars are *extremely luminous at all wavelengths*, which implies an enormous energy source.

- Many also exhibit *variability* in this luminosity on timescales as little as *months, weeks, or even hours*.
- As will be discussed next, this variability implies that they have a *very compact energy source*.

15.6.4 Causality and Maximum Size of an Energy Source

If an energy source of a certain size is to exhibit a well-defined period in its luminosity,

- Some signal must travel through the object to tell it to vary.
- The signal can travel *no faster than light velocity*.
- Thus the maximum size D of an object varying with some characteristic time P is the distance that light could have traveled during that time, $D \sim cP$.
- The figure below illustrates.



- For example, if a source is observed to vary its light output substantially over a period of a week, then
- the spatial extent of the energy-producing region can be no larger than a "light week", which is 1.82×10^{11} km.

The *distances covered by light* for various fixed times are summarized in the following table.

Time	km	AU	Parsecs
Year	9.46×10^{12}	63,240	3.07×10^{-1}
Month	7.88×10^{11}	5270	2.58×10^{-2}
Week	1.82×10^{11}	1216	5.90×10^{-3}
Day	2.59×10^{10}	173	8.41×10^{-4}
Hour	1.08×10^{9}	7.21	3.50×10^{-5}
Minute	1.80×10^{7}	0.120	5.84×10^{-7}
Second	3.00×10^{5}	0.002	9.73×10^{-9}
Millisecond	3.00×10^{2}	0.000002	9.73×10^{-12}

These arguments place only an *upper limit on source sizes*.

- Signals causing the variation may travel at *less than light speed* and
- the size may be *less than the upper limit* imposed by these arguments.
- But this is a *very powerful argument* because
  - it depends only on a general principle (finite speed of light) and
  - not on the internal details of the source.

Example: If the source varies over a period of an hour, from the table we see that the maximum size of the emitting region is ~ 7 AU.

When it was first realized that

- quasars possess *extremely powerful energy sources* and that
- this energy is produced in a *region no larger than the Solar System*,

serious issues were raised about whether known physical processes could account for these energy characteristics.

- Since the discovery of quasars we have achieved a much deeper understanding of black holes and a likely mechanism to produce energy on this scale in such a compact region has emerged.
- There is relatively uniform agreement that
  - *rotating black holes*
  - containing of order $10^9 M_\odot$,
  - which would have *radii smaller than the Solar System*

  are the most plausible candidates for quasar energy production.

However, before turning to a more detailed discussion of this idea, let's examine another class of observational phenomena that appears to have much in common with quasars.

15.7 Active Galactic Nuclei

A collection of billions of ordinary stars is expected to produce a spectrum that is *approximately blackbody*, because the spectrum of the individual stars is blackbody.

- Such a spectrum is dominated by a continuum that often peaks at visible wavelengths.
- For example, the Milky Way emits radio waves, but its radio luminosity is about a million times smaller than its visible luminosity.
- In addition, the spectral lines observed for normal galaxies (as for its stars) are *mostly absorption lines*, with *few emission lines*.

Thus, spectra for normal galaxies, as for the stars that they contain, are

- typically a continuum peaking at visible-light wavelengths representing thermal emission, with
- absorption lines superposed on the continuum and few emission lines.

However, some galaxies differ from this norm:

- They exhibit nonthermal emission from the RF to X-ray region of the spectrum and/or
- jets and unusual structure associated with the visual appearance of the galaxy.
- We refer to these as *active galaxies.*
- Since the source of the activity is normally concentrated in the nucleus of the galaxy, they are also called *active galactic nuclei* or *AGN.*

Generally, active galaxies exhibit some combination of the following characteristics:

- Unusual appearance, particularly of the nucleus.
- Jets emanating from the nucleus.
- High luminosities relative to normal galaxies, but generally smaller than for quasars.
- Excess radiation at RF, IR, UV, and X-ray wavelengths.
- Nonthermal continuum emission,
  - often polarized,
  - perhaps with broad and/or narrow emission lines.
- Often rapid variability from a compact energy source in the galactic nucleus.

Some galaxies with active nuclei are relatively nearby but

- They are more likely to be found at larger distances,
- implying that they were more common earlier in the Universe's history.
- Many of these characteristics are also exhibited by quasars, and
- we shall see that quasars may be closely related to the nuclei of active galaxies.
- In fact, we shall conclude below that quasars essentially are a particularly luminous form of active galactic nucleus.

There are several general classes of AGN that we now summarize.

15.7.1 Radio Galaxies

Radio galaxies are AGN associated with a *large nonthermal emission of radio waves*.

- Strong radio sources are often named by the constellation followed by a letter designating order of discovery.
- *Example:* The relatively nearby radio galaxy M87 is associated with the powerful radio source *Virgo A*.
- Powerful radio galaxies are elliptical and often exhibit jet structure from a compact nucleus.
- Weaker radio sources are found associated with smaller jets in some spiral galaxies.
- There are two broad classes of radio galaxies.
    1. *Core-halo radio galaxies* exhibit radio emission from a region near the nucleus of the galaxy that is comparable in size to the optically visible galaxy.
    2. *Lobed radio galaxies* display great lobes of radio emission that can extend millions of lightyears beyond the optical part of the galaxy.
- A radio galaxy can be a spectacular sight at RF wavelengths, but even powerful radio galaxies may appear to be rather normal elliptical galaxies at optical wavelengths.
- Often abnormalities at optical wavelengths become obvious only if the very core of the galaxy is resolved.

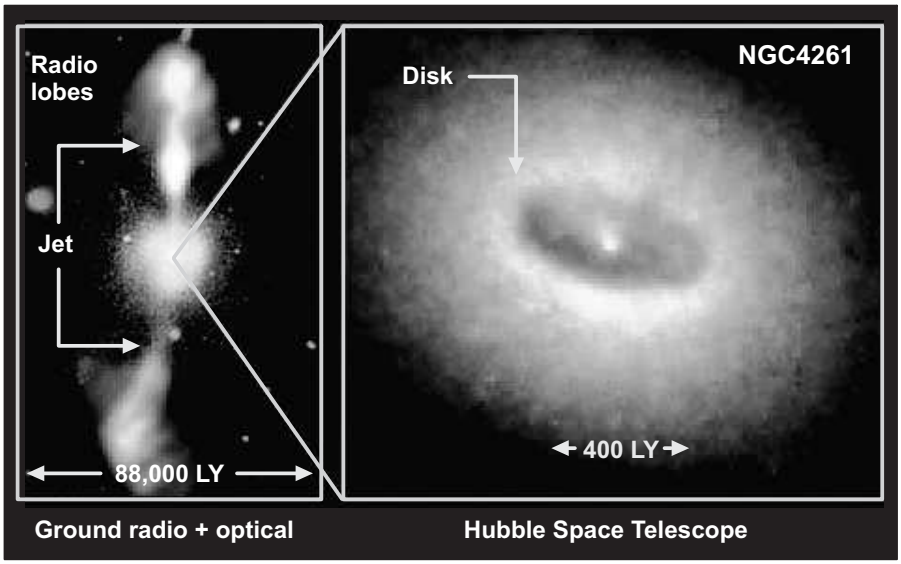


Figure 15.4: The radio galaxy NGC4261. On the left the false-color radio-wave map is superposed on the ground-based optical image of the galaxy. The image on the right is a high-resolution Hubble Space Telescope image of the core of the galaxy, showing a dusty accretion disk thought to surround a central black-hole engine that is driving the jets producing the radio lobes. The black hole engine lies inside the bright dot at the center of the disk.

The left side of Fig. 15.4 shows *huge radio lobes* superposed on the optical image of the elliptical galaxy NGC4261.

- The optical image of NGC4261 on the left side suggests a rather normal looking elliptical galaxy.
- However, the high-resolution image shown on the right indicates *something very unusual* at the center of the galaxy.
- The suggestion is that
  - there is a *supermassive black hole* at the center and
  - this is the *energy source powering the radio jets*.

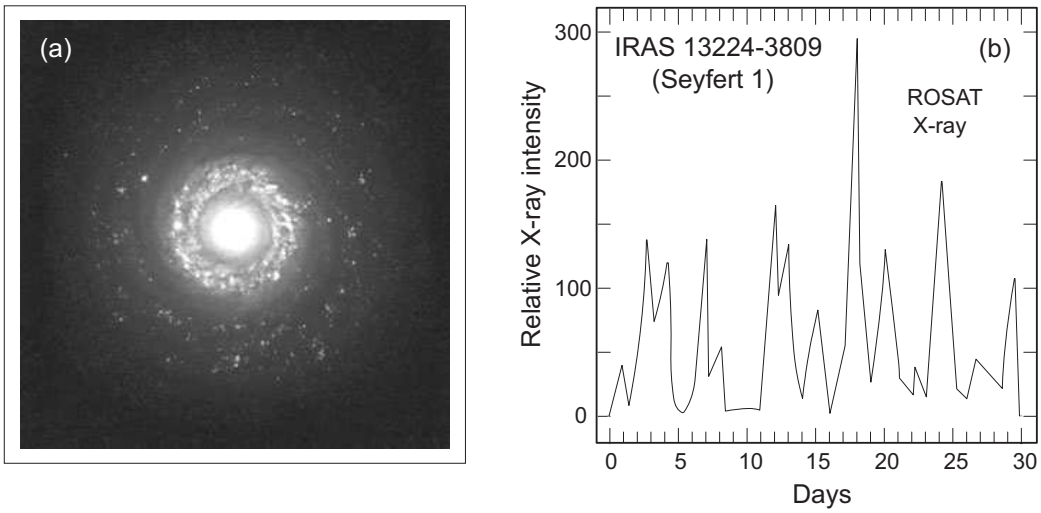


Figure 15.5: (a) The Seyfert 2 galaxy NGC 7742, which lies about 22 Mpc away in the constellation Pegasus. The nucleus is very compact and bright at optical wavelengths, which is characteristic of Seyfert spirals. (b) X-ray variability of the Seyfert 1 galaxy IRAS 13224-3809.

15.7.2 Seyfert Galaxies

Seyfert galaxies are spirals with very *bright and compact nuclei*. An example is displayed in Fig. 15.5.

- They are the *most commonly observed AGN* and
  - exhibit a *strong nonthermal continuum* from IR through X-ray regions of the spectrum,
  - with *emission lines of highly-ionized atoms* that are sometimes variable.
- Emission lines suggest a *low-density gas source*, while high ionization implies a *hot source*.
- Some Seyferts have *modest jets and RF emission*.

Seyfert galaxies can be classified roughly into *two subgroups:*

- *Seyfert 1 galaxies* have both broad and narrow emission lines and are very luminous at UV and X-ray wavelengths.
  1. *Broad emission lines* are associated with *allowed atomic transitions*.
  2. Their width comes from *Doppler broadening* in clouds with velocities as high as $\sim 10,000$ km/s.
  3. *Narrow lines* are associated with *forbidden transitions* and source velocities less than ~ 1000 km/s.
  4. These differences suggest that *broad and narrow lines originate in different regions* of a Seyfert 1.
- *Seyfert 2 galaxies* have relatively narrow emission lines suggesting source velocities less that ~ 1000 km/s.
  1. They are weak in X-ray and UV; very strong in IR.
  2. Generally, their continuum emission is weaker than for Seyfert 1 galaxies.
  3. In a Seyfert 2, forbidden and allowed lines are narrow.
  4. This suggests that
     - both *originate in the same region* with
     - relatively *low source velocities*.

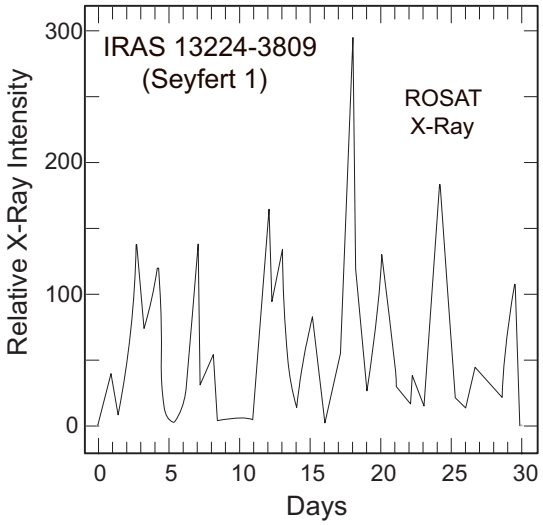


Figure 15.6: X-ray variability of the Seyfert 1 galaxy IRAS 13224-3809.

Seyfert galaxies can exhibit 50% changes in optical brightness over weeks and even larger changes over months.

- As illustrated in Fig. 15.6, the brightness variation at X-ray wavelengths can be even greater.
- By previous arguments for brightness variation in quasars,
  - the fluctuation on a timescale of approximately a day exhibited in Fig. 15.6
  - implies a *very compact energy source* for the X-rays.

We conclude that the X-ray energy source is *no larger than light days in diameter* (the Solar System is about half a light day in diameter).

15.7.3 BL Lac Objects

BL Lacertae objects (*BL Lac objects* or just *BL Lacs,* for short)

- Exhibit *no, or only weak,* emission lines,
- but have a *strong nonthermal continuum* stretching from RF through X-ray frequencies.
- They are generally *radio-loud*, and exhibit *strong polarization* of their emitted light
- (Polarization is typically *several percent*, as compared with 1 percent or less for most other AGN.)
- BL Lacs are members of a more general class of AGN called *blazars.*
- *Blazars are relatively uncommon* among AGN.
- For example, there are about
  - 100 times more quasars and
  - 10 times more Seyfert galaxies

  known than blazars.
- However, they are *extremely powerful*, typically being
  - 10,000 times more luminous than the Milky Way and
  - 1000 times more luminous than a Seyfert galaxy.

By *masking the bright core* that is responsible for producing the *intense and rather featureless continuum*,

- it is possible to acquire a spectrum of the "fuzz" seen faintly around many BL Lacs.
- This spectrum, and the variation of the light intensity of the fuzz with distance from the core, indicate that
- the fuzz is the *outer part of a giant elliptical galaxy* in most of the cases that have been analyzed.
- This suggests that most BL Lacs are *active cores of giant elliptical galaxies*.
- From the *redshifts of faint spectral lines* it is found that
  - blazars often correspond to more distant objects than Seyfert galaxies or radio galaxies, but
  - they are closer to us on average than quasars.
- Blazars can exhibit *dramatic variability* in their light output (and a corresponding variability in their polarization).
  - They are observed to vary significantly on *timescales as short as days*.
  - This implies a power source the *size of the Solar System or smaller*.

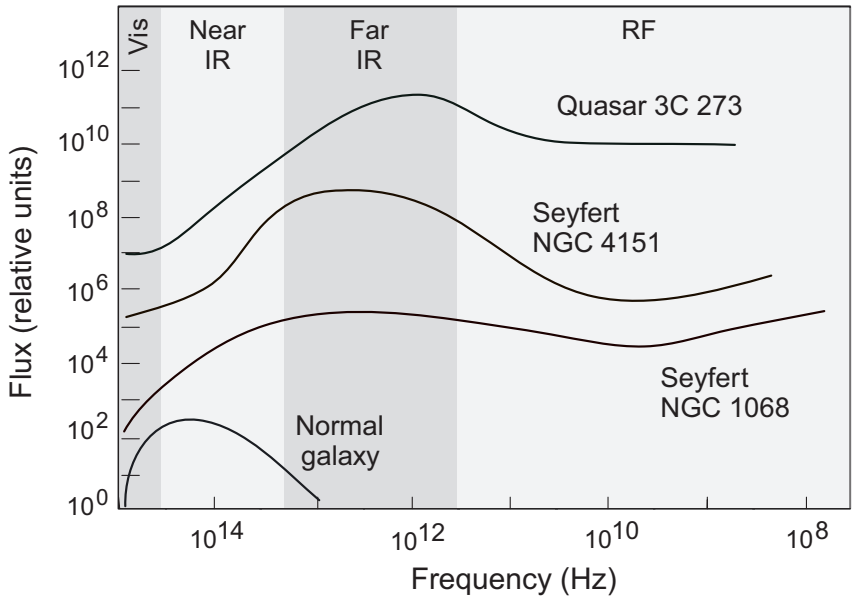
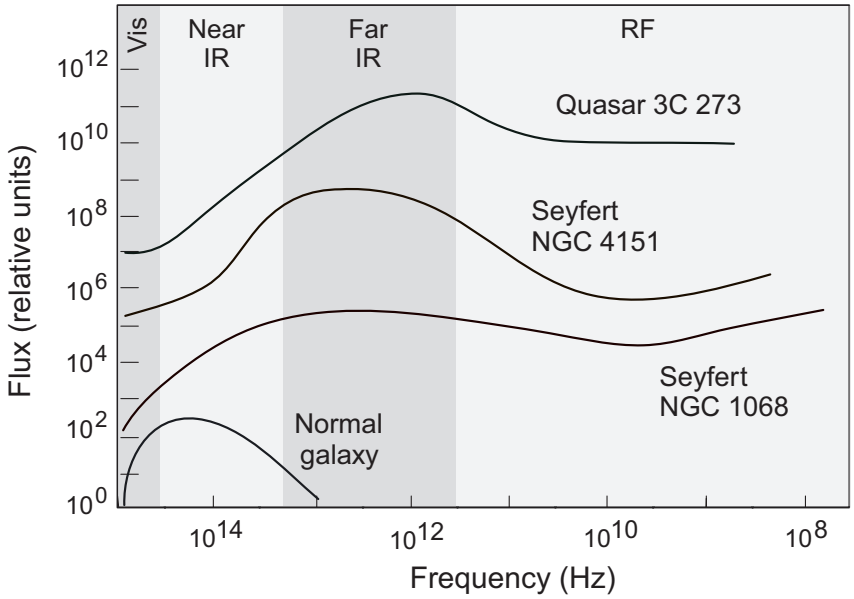


Figure 15.7: Some representative spectra for normal galaxies, active galaxies, and quasars. The spectrum of the galaxy is approximately thermal but the spectra of the active galaxies and quasars are highly nonthermal. Because electrons must radiate energy continuously as they spiral in magnetic fields, the energy driving the huge sustained synchrotron emission from these nonthermal sources must be replenished constantly. Strong, polarized, nonthermal emission is an indirect sign not only of strong magnetic fields, but also of a very large energy source for the quasars and active galaxies.

15.8 The Unified Model of AGN and Quasars

Some representative spectra of normal galaxies, quasars, and active galaxies are shown in Fig. 15.7.

- These spectra imply that active galaxies and quasars are *very different from normal galaxies*.
- On the other hand, the similarity of the spectra for quasars and Seyfert galaxies implies that there might be a *relationship between quasars and active galaxies*.



- That relationship is suggested by the
  - *huge nonthermal emission* from AGN and quasars,
  - which implies a *strong and very compact energy source*.
- The *only plausible candidate* for the engine driving these phenomena is a *rotating, supermassive black hole* of mass millions to billions of solar masses at their center.

In the past several decades a large amount of observational data and theoretical understanding supporting this point of view has emerged.

The present belief is that (despite observational differences)

- AGN such as *Seyfert galaxies, blazars, and radio galaxies are quite closely related*, and that
- *Quasars* are just a *particularly energetic form of AGN*.
- All are now thought to be
  - *active galaxies* with
  - bright nuclei powered by *supermassive Kerr black holes*.

In the unified model that we now discuss, the observational differences among quasars and various AGN mostly reduces to a matter of

1. How rapidly matter is accreting onto the black hole *(feeding the monster).*
2. Whether the central engine is masked from our view *(hiding the monster).*
3. How far away the AGN or quasar is from us *(proximity to the monster).*

We are thus led to an hypothesis:

Perhaps all large galaxies have supermassive black holes at their centers.

- The question of whether the galaxy exhibits an active nucleus is then primarily one of *"feeding the monster"* at the center.
- *In a quasar*, the black hole is accreting matter at a higher rate, leading to very high luminosity.
- *In a normal galaxy* like the Milky Way, the black hole is rather quiet because it is presently accreting little matter.
- *Seyfert and other active galaxies* are somewhere in between: the black hole is active because it is accreting matter, but at a rate lower than that for a quasar.

15.8.1 The AGN Central-Engine Model

We now discuss a *unified AGN model* where all active galaxies and quasars are powered by central supermassive black holes.

1. In this unified model, the differences among AGN arise primarily from
   - differences in *orientation angle* and
   - differences in *local environment*

   for the central engine.
2. The first determines whether the view of the central engine is blocked by dust;
3. the second determines the rate at which fuel flows to the central engine.

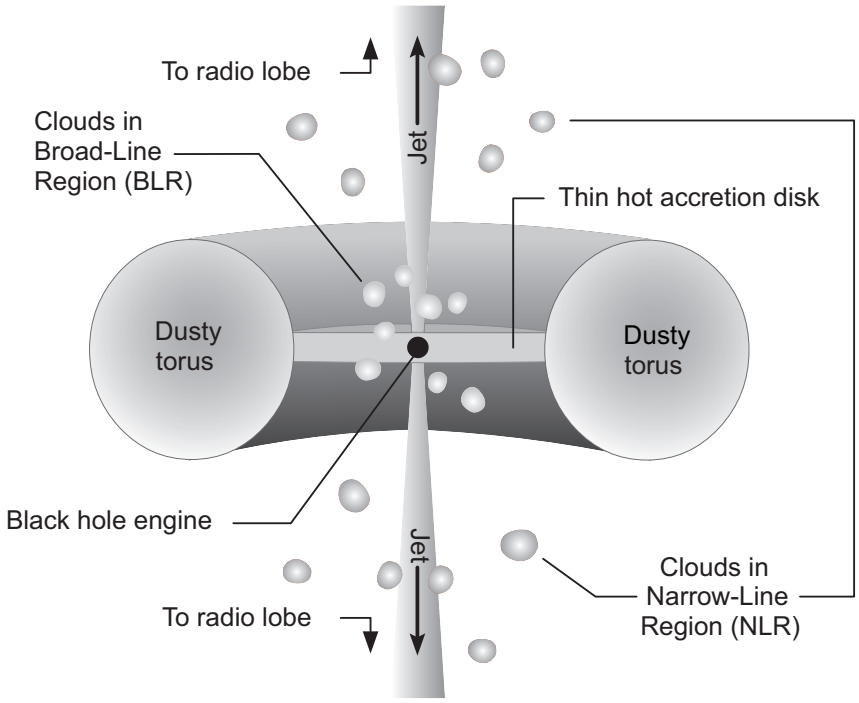


Figure 15.8: Illustration of the AGN black-hole central engine paradigm.

Fig. 15.8 illustrates the *standard central engine paradigm*.

- The black hole occupies a tiny region in the center (its size is *greatly exaggerated* in this figure).
- It is surrounded by
  - a *dense torus of matter* revolving around the hole and
  - a *flattened accretion disk* inside of that where the matter whirls even faster.
- Part of the matter is *sucked into the black hole* and
- part is *flung out in high-velocity jets* along the polar axes of the rotating black.

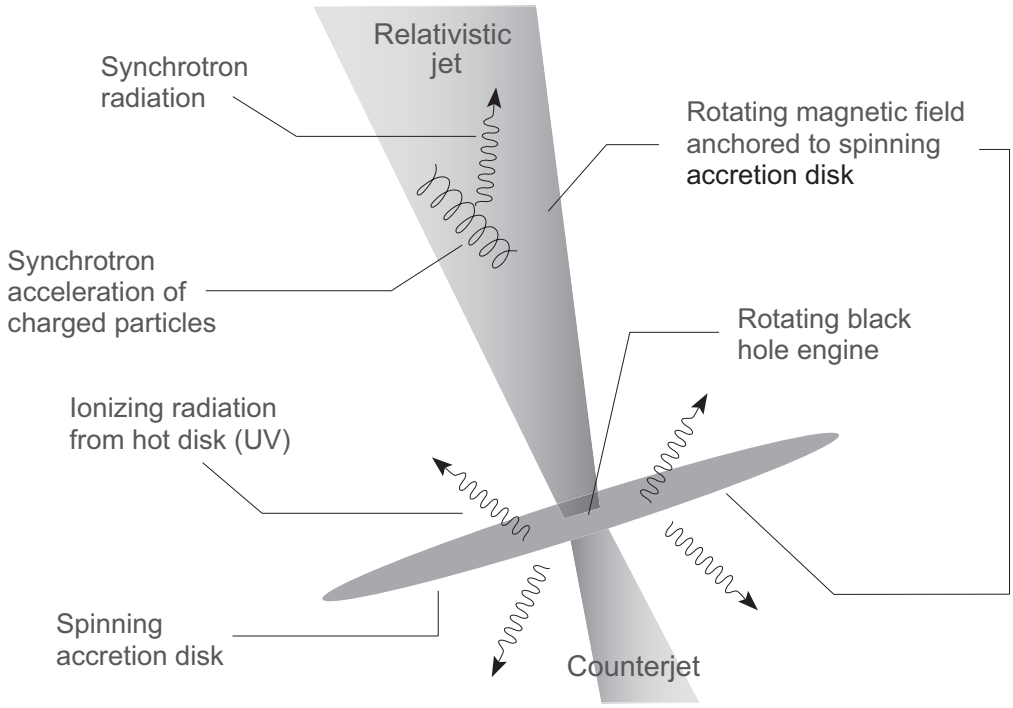
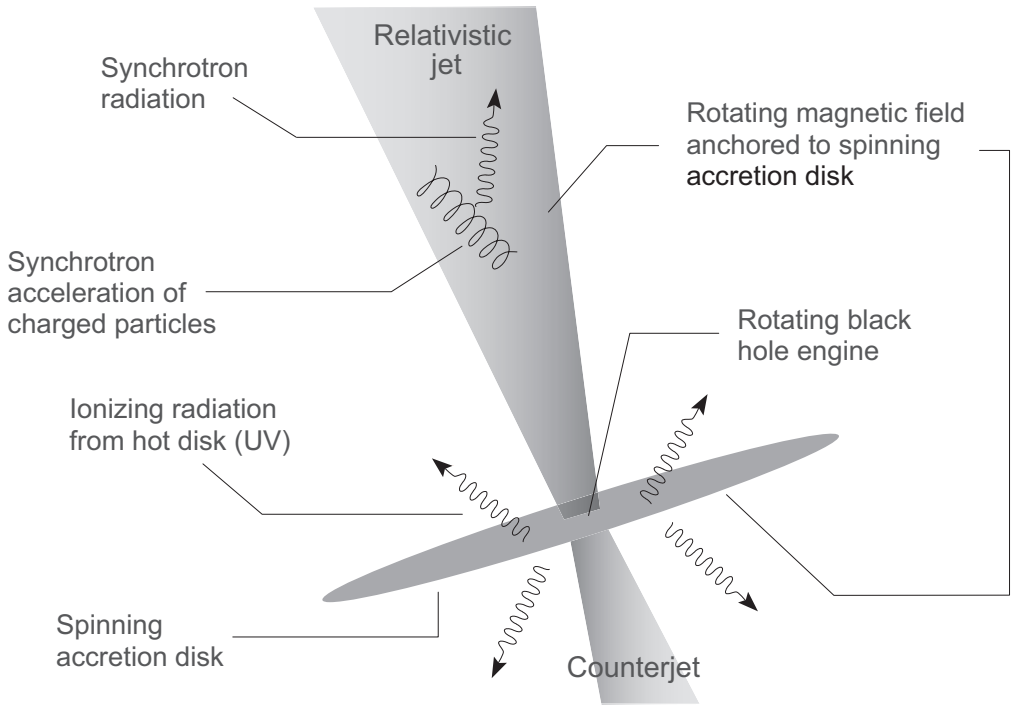


Figure 15.9: Schematic view of the central engine for an active galactic nucleus or quasar. Beamed synchrotron radiation is emitted by the electrons in the jet and UV photons are emitted by the hot accretion disk.

Figure 15.9 illustrates schematically what the central engine of an AGN or quasar would look like if we could *turn the radiation off* and *clear away the gas and dust*.

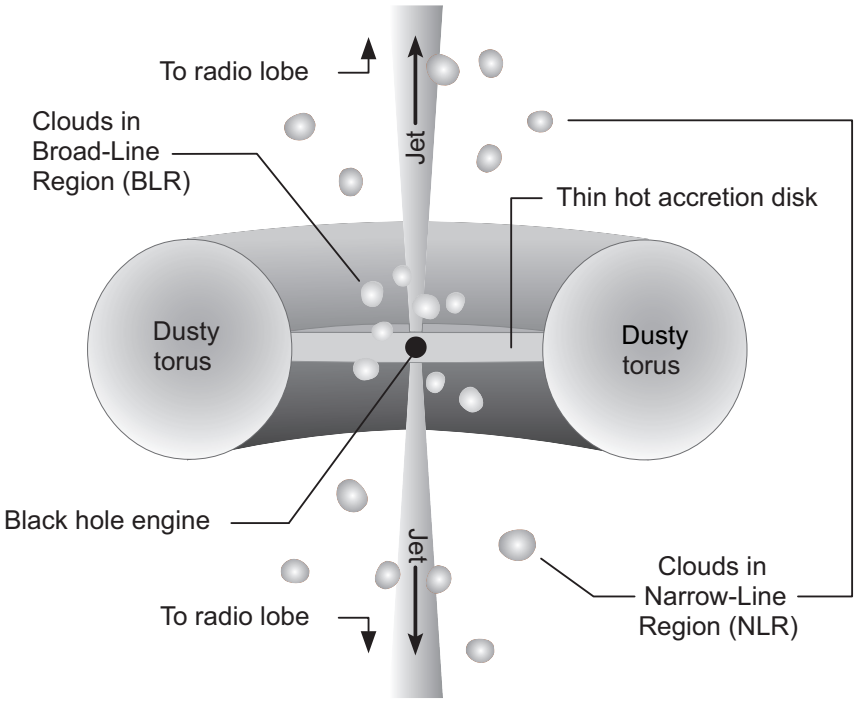
- The central engine consists of a *supermassive Kerr black hole* surrounded by a *thin accretion disk*.
- The accretion disk is very *hot because of collisions* in the rapidly swirling gas that it contains.
- Because it is hot, it *radiates strongly in the UV*.



Radiation from the accretion disk is *mostly thermal*.

- Typical estimates of the temperature for the accretion disk of a $10^9 M_\odot$ black hole are about 10,000 K;
- This corresponds to a peak wavelength of 3000 Angstroms (Wien law), which lies in the UV.
- Accretion disk temperature depends *inversely on the black hole mass*, so smaller ones have hotter accretion disks.

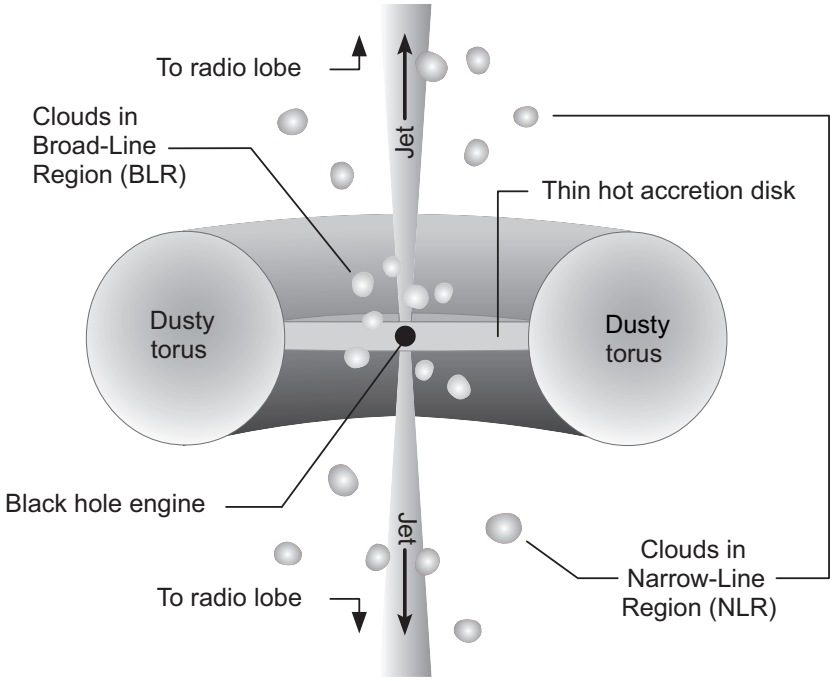
Stellar black holes have accretion disks that are even hotter than the ones considered here, *radiating more in the X-ray* region of the spectrum.



Photons from the accretion disk are responsible for much of the continuum observed from AGN,

- either directly, or
- by heating surrounding matter,

which then re-radiates at longer wavelengths.



Photons emitted by the accretion disk

- ionize atoms in the nearby clouds of gas where
- velocities are very high, producing the broad-line emission spectrum of the AGN *(BLR).*

They also

- ionize clouds further away from the central engine where velocities are lower,
- producing the narrow-line emission spectrum *(NLR).*

Whether both broad and narrow emission lines are visible depends on *location of the observer* relative to the accretion disk and torus.

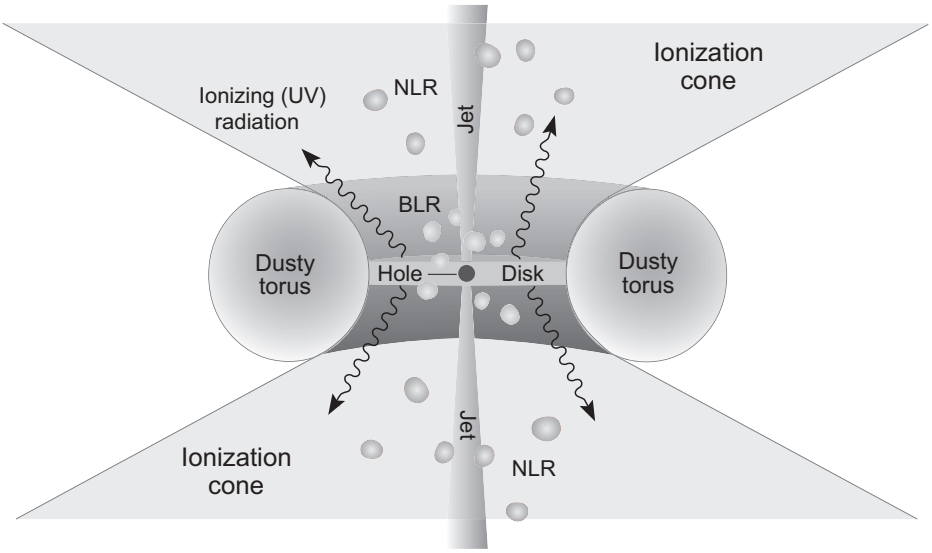
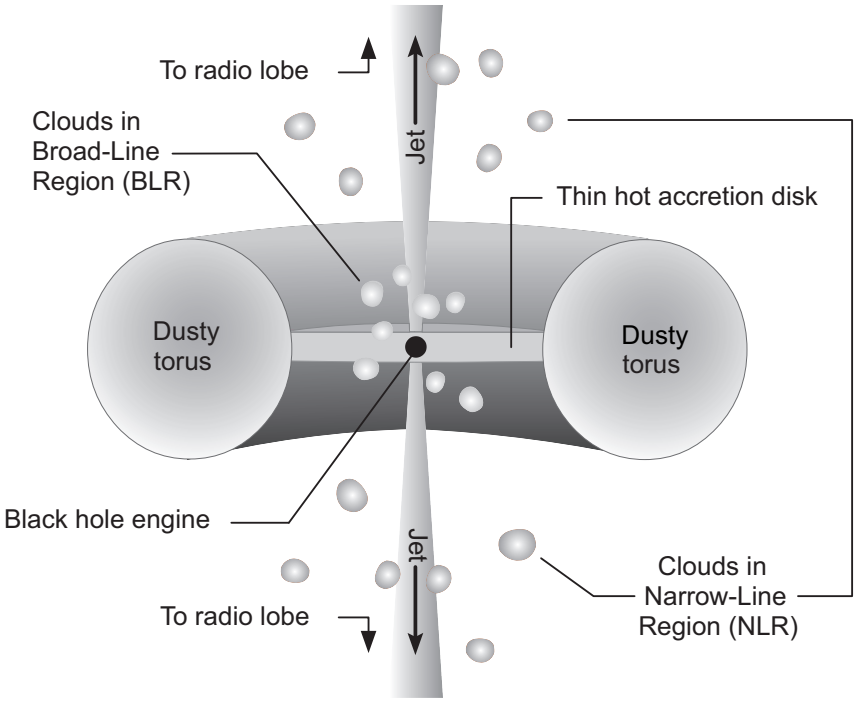


Figure 15.10: AGN ionization cones. Shaded regions see directly the UV radiated by the accretion disk. Other regions cannot see the accretion disk directly because it is blocked by the torus. Clouds in the broad-line region are labeled BLR and clouds in the narrow-line region are labeled NLR.

If the preceding picture is correct, there should be evidence in AGN for *anisotropic ionization:*

- Ideally, there should be *cone-shaped regions of ionization*
- corresponding to the directions from the hot accretion disk that are *not blocked off* by the dust torus.
- These *ionization cones* are illustrated in Fig. 15.10.
- All of the region shaded in gray can "see" the hot central accretion disk and the ionizing radiation that it is emitting.

Observations suggest such *anisotropic ionization zones* near AGN central engines.



Let's now summarize the basic *unified model*. We hypothesize that *AGN and quasars* are powered by *rotating, supermassive black holes* having all or most of the following features:

- *Accretion disks*.
- Possibly a *dusty torus* surrounding the accretion disk.
- *High-velocity clouds* near the black hole that produce *broad emission lines*.
- *Slow, lower-density clouds* further from the black hole that produce *narrower emission lines*.
- Possible *relativistic jet outflow* perpendicular to the plane of the accretion disk.

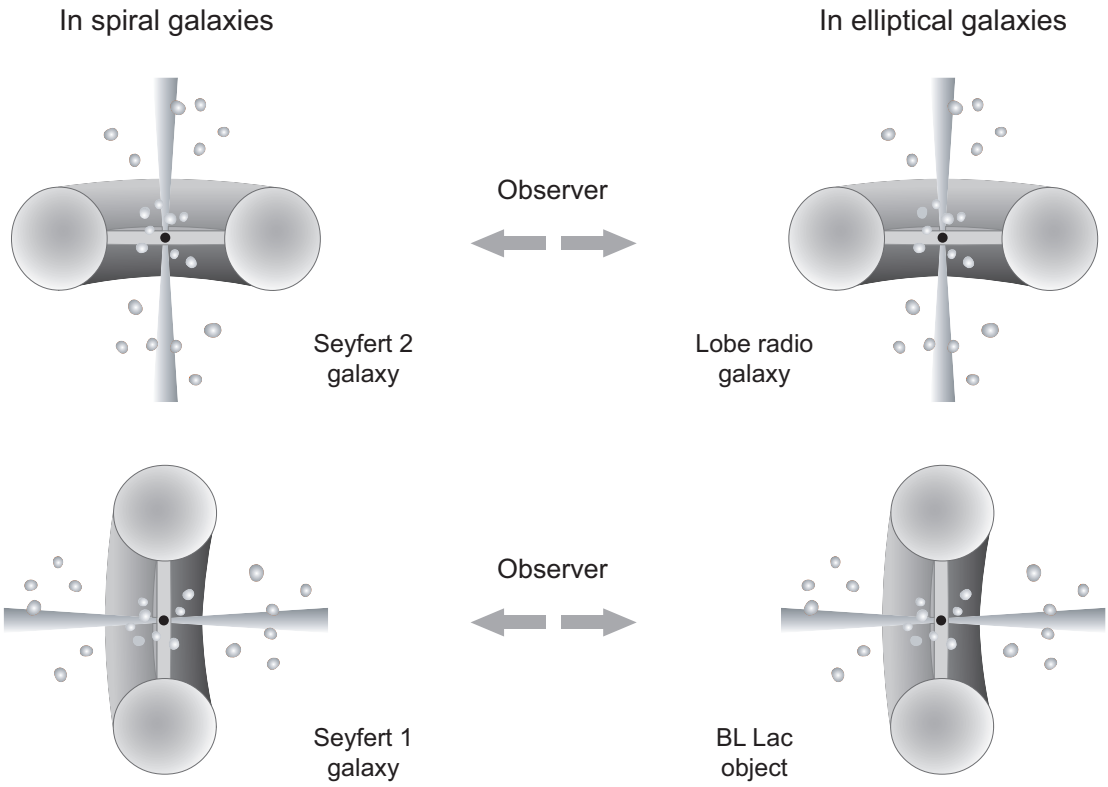
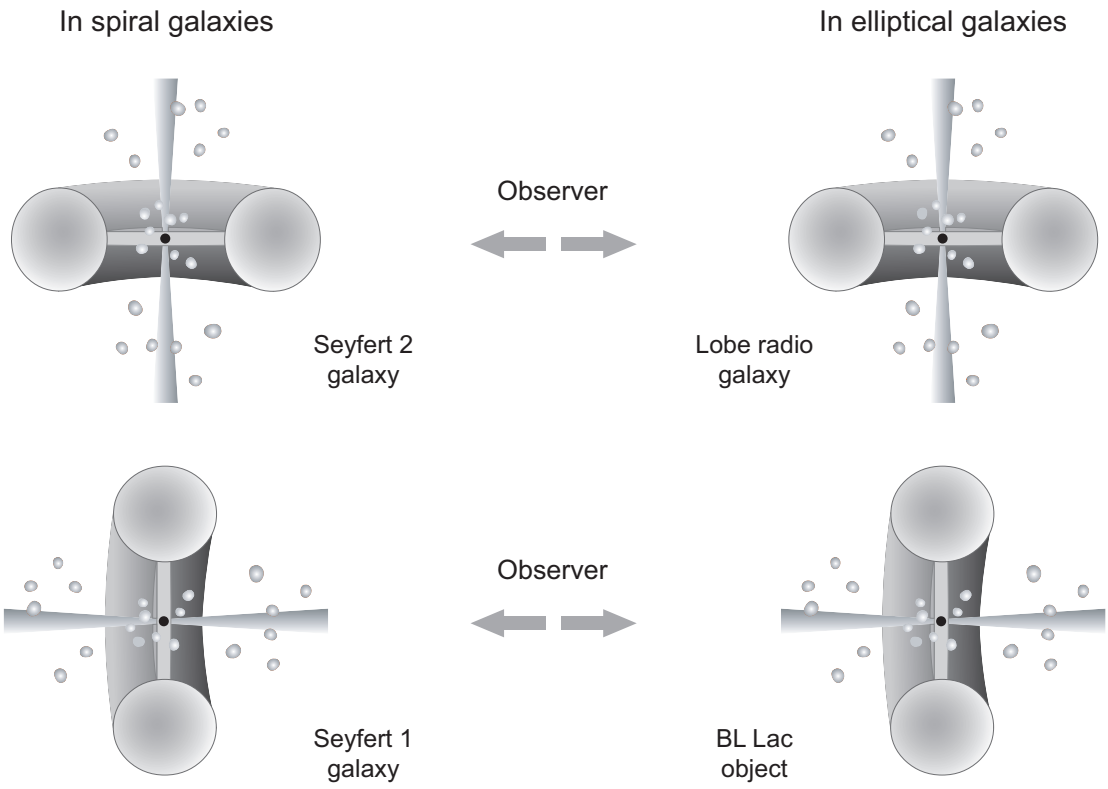


Figure 15.11: A unified model of active galactic nuclei based on geometry. All cases are powered by supermassive Kerr black holes and observational differences are determined primarily by viewing angles relative to the plane of the black hole accretion disk and the jets. Quasars are then hypothesized to be similar to AGN, but more powerful because the black hole is accreting at a higher rate.

- The cartoon in Fig. 15.11 illustrates this *unified AGN model*. If the orientation of the system relative to the observer is as displayed in the top row of Fig. 15.11,
  1. We see a *Seyfert 2* if the AGN is in a *spiral galaxy*.
  2. We see a *lobed radio source* if the AGN is in an *elliptical galaxy with strong jets*.



- On the other hand, if the orientation of the AGN central engine is as in the bottom row of the above figure,
  1. We see a *Seyfert 1* if the host is a *spiral galaxy*.
  2. We see a *blazar (BL Lac)* if the host is *elliptical with a strong jet*.
  3. We see perhaps a *core-halo radio galaxy* if the host is an *elliptical system with a weak jet*.

We then conjecture that

- *quasars* are just particularly *energetic forms of AGN* and
- are described by the *same unified model*.

The primary difference is that

- quasars are very luminous because
- their black holes are being fed matter at a higher rate than for moderately luminous AGN.

We surmise then that the reason quasars are *more abundant at larger redshift* is because

- large redshift corresponds to earlier in the Universe's history, when
- *matter was more dense* and
- *collisions more frequent* between galaxies.

Therefore, quasars may be just *"better fed" AGN* from a time when *more fuel was available* to power the black hole engines.

As a corollary, we may conjecture that

- many nearby normal galaxies once sported brilliant quasars in their cores,
- but have since used up the available fuel.
- Their massive black holes lie dormant, ready to blaze back to life should tidal interactions with another galaxy divert matter into the black hole.

This may be true of our own galaxy, which has a $\sim 4 \times 10^6 M_\odot$ black hole at its center, but presently exhibits only *weak nonthermal emission*.

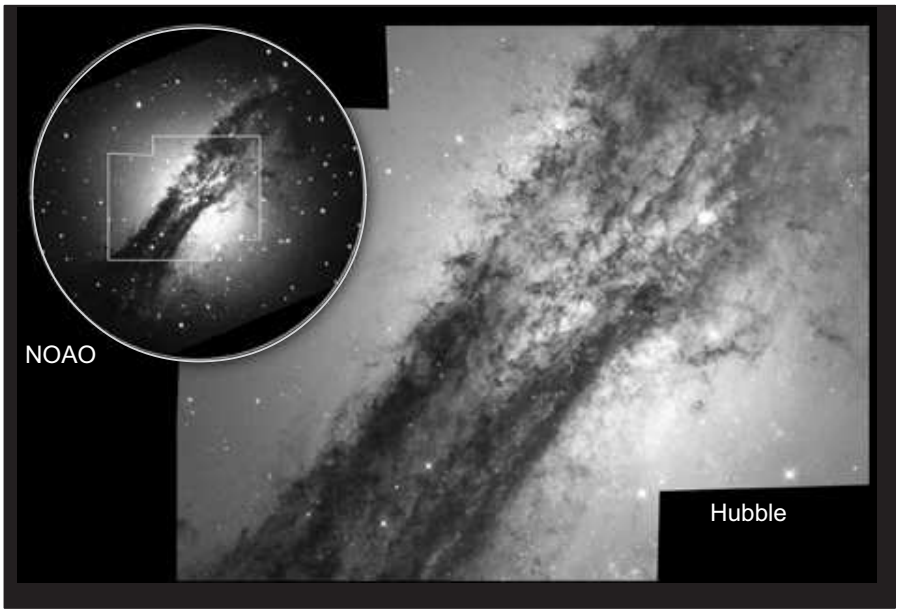
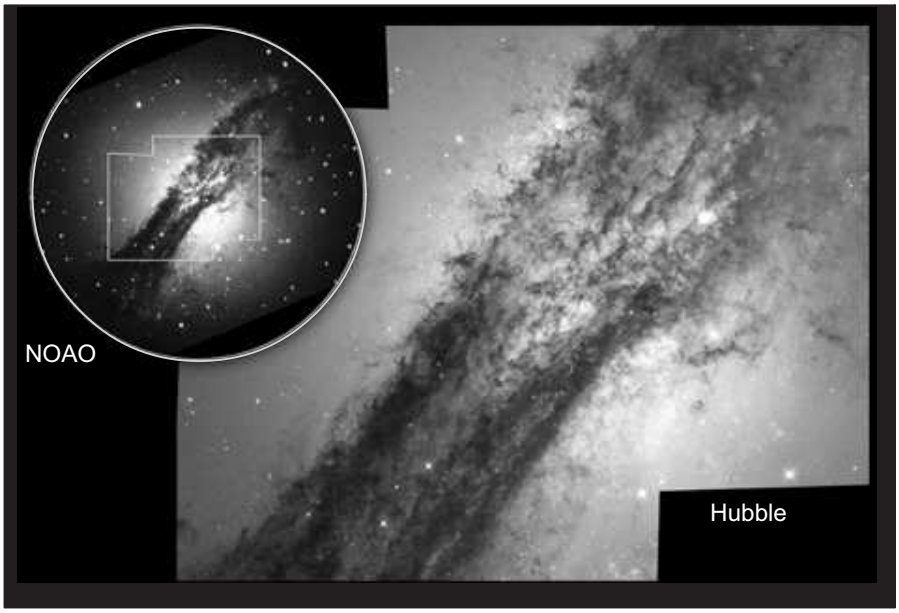


Figure 15.12: One of the nearest active galactic nuclei, Centaurus A (in NGC 5128).

15.8.2 Example: Feeding a Nearby Monster

Fig. 15.12, displays a nearby AGN: the giant elliptical galaxy NGC 5128, which harbors the radio source Centaurus A.

- Centaurus A is only about 10^7 ly away. Its radio lobes span 10° in our sky (20 times the Moon's diameter)
- It has a faint optical jet and strong radio jet, and violent central activity causing the jets and huge radio lobes.
- IR observations suggest that behind the dark dust lanes lies a gas disk 130 ly in diameter, surrounding the
- $\sim 10^9 M_\odot$ black hole powering Centaurus A.



The firestorm of starbirth in the above figure was caused by

- a *collision* between a *smaller spiral galaxy* and a *giant elliptical galaxy* within the last billion years.
- Dark diagonal dust lanes may be the *remains of the spiral galaxy*, encircling the core of the giant elliptical.
- The logical conclusion is that the black hole at the center of Centaurus A has been *triggered into pronounced activity* by the collision of its parent galaxy with another.
- This collision has led to *strong radio emission* caused by *relativistic jets* powered by *enhanced accretion*.
- It has also produced *enhanced star formation* because of the *compression of gas and dust* in the collision.

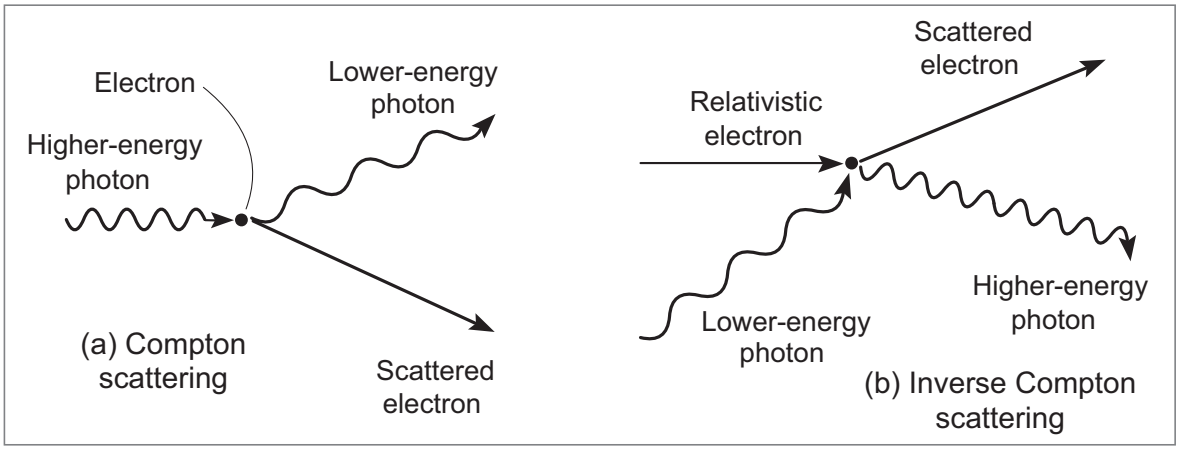
Figure 15.13: X-ray jet from center of Centaurus A superposed on an optical image of the galaxy. In this Chandra X-ray Observatory map the highest X-ray intensity is indicated by white.

This black hole central-engine interpretation is supported further by Fig. 15.13.

1. This image shows an *X-ray map* obtained by the *Chandra X-ray Observatory* superposed on an optical image of Centaurus A.
2. The Chandra data reveal a bright central region (white ball near the center) that likely surrounds the black hole.
3. It also indicates a strong jet oriented to the upper left $25,000$ ly in length that is probably being ejected on the polar axis of the central black hole.
4. There also is a fainter jet oriented in the opposite direction, making it harder to see.

Blazars sometimes emit *extremely high-energy photons*.

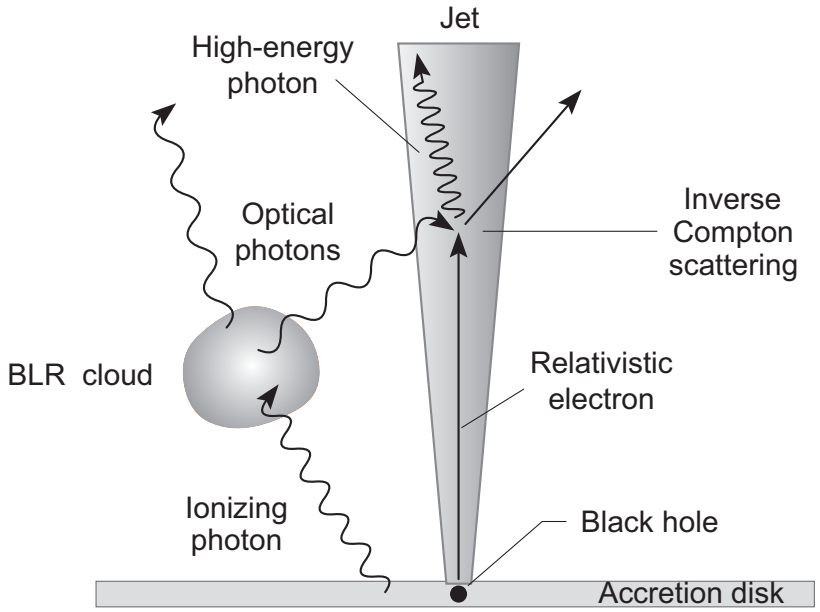
- For example, 10^{12} eV gamma-rays have been detected coming from Markarian 421, a BL Lac at a redshift of $z = 0.031$.
- Such gamma-rays *cannot be produced by AGN thermal processes*, but
- the unified model of AGN and quasars provides a possible explanation.
- The next slide illustrates a mechanism called *inverse Compton scattering* through which lower-energy photons can be boosted to extremely high energy in an AGN.



15.8.3 Producing Very High Energy Photons

Some AGN emit extremely high-energy photons that are thought to involve a process called *inverse Compton scattering,* illustrated in the figure above.

- In *Compton scattering* a high-energy photon strikes an electron, giving up energy.
- *Inverse Compton scattering* is the reverse: a high-energy electron strikes a photon, imparting energy.

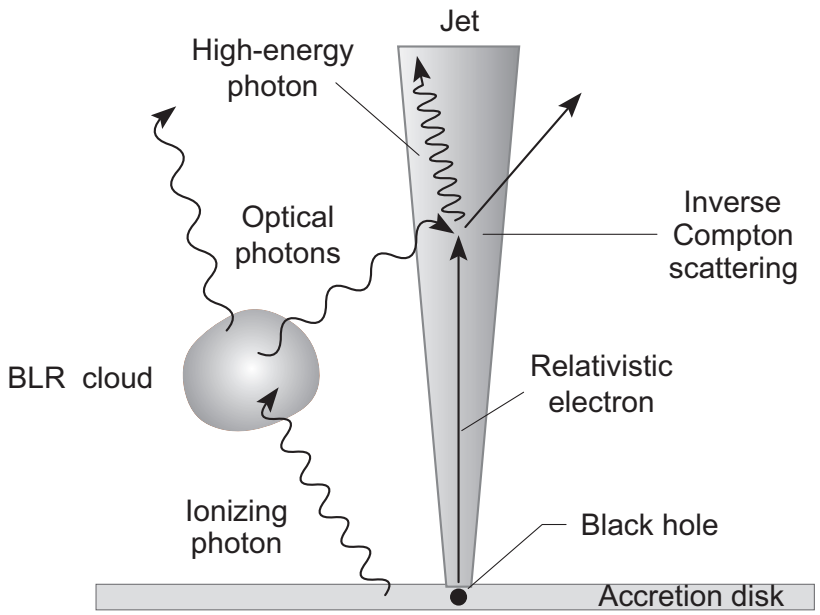


An AGN could produce a high-energy photon by the *inverse Compton process* illustrated in the figure above.

- UV from the disk can photoionize clouds in the broad line region (BLR), which then emit at optical wavelengths.
- If these photons enter the jet they can be inverse Compton scattered to much higher energies.
- The *frequency boost factor* is given by

$$\frac{\nu}{\nu_0} = \frac{1}{1 - v^2/c^2}$$

  where v is the jet velocity.
- Thus in a highly relativistic jet *boost factors of* 10^6 *can be obtained*.



The inverse Compton process illustrated above can convert

- optical photons into intermediate-energy gamma-rays and
- X-rays into 10^{12} eV gamma-rays.

Thus, the 10^{12} eV gamma-rays observed coming from Markarian 421 could be produced if

- X-rays produced by irradiation from the accretion disk
- enter a highly-relativistic jet and
- are inverse Compton scattered.

15.9 Gamma-Ray Bursts

Discovered serendipitously by military satellites,

- gamma-ray bursts (GRB) were initially one of the *great mysteries in modern astronomy*.
- On average, about *one burst a day* is observable somewhere in the sky.
- However, until the 1990s we had little idea about
  - *where they originated* or even
  - *how far away* they were.

New observations have begun to clear away the mystery and we now have a much better understanding of these remarkably energetic events.

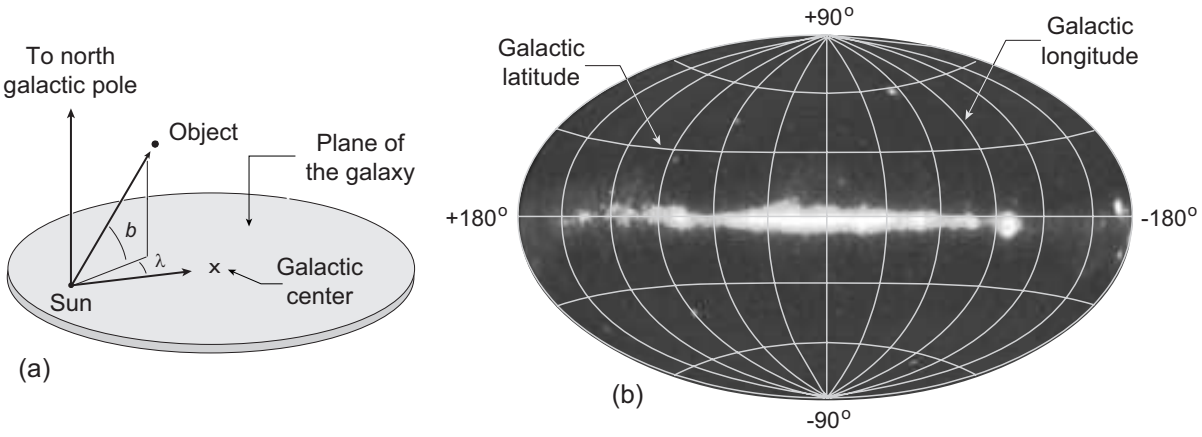


Figure 15.14: (a) Galactic coordinate system. The angle b is the galactic latitude and the angle λ is the galactic longitude. These may be related to right ascension and declination by standard spherical trigonometry. (b) The sky at gamma-ray wavelengths in galactic coordinates. White denotes the most intense and black the least intense sources. The diffuse horizontal feature at the galactic equator is from gamma-ray sources in the plane of the galaxy. Bright spots right of center in the galactic plane are galactic pulsars. Brighter spots above and below the plane of the galaxy are distant quasars.

15.9.1 The Gamma-Ray Sky

Our sky glows in gamma-rays.

- Since gamma-rays are high-energy photons, they are not easy to produce and tend to indicate unusual phenomena.
- Thus they are of considerable astrophysical interest.
- They are absorbed by the atmosphere so a systematic study requires an orbiting observatory.

Fig. 15.14(b) shows the continuous glow of the gamma-ray sky, measured by the orbiting Compton Gamma-Ray Observatory.

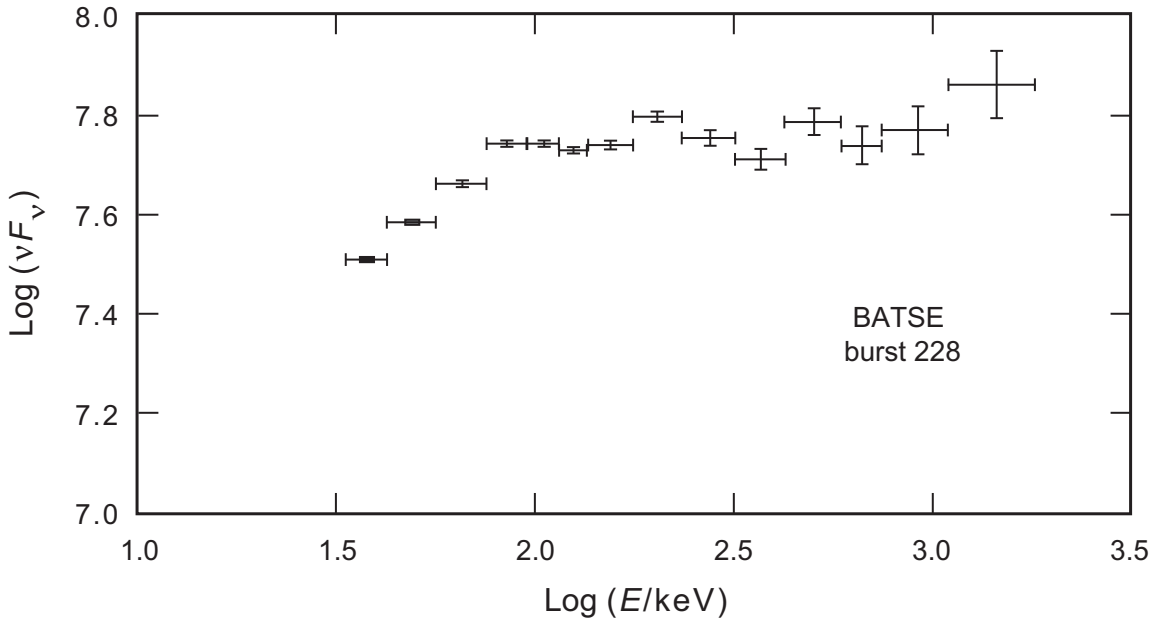


Figure 15.15: Spectrum of a gamma-ray burst. As is characteristic, the spectrum is nonthermal.

Superposed on the steady flux are *sudden bursts*.

- These *gamma-ray bursts* can be
  - as *short as tens of milliseconds* and
  - as *long as several minutes*.
- The spectrum is
  - *nonthermal* and
  - dominated by photons in the *hundreds of keV to several MeV* range,

as illustrated in Fig. 15.15.

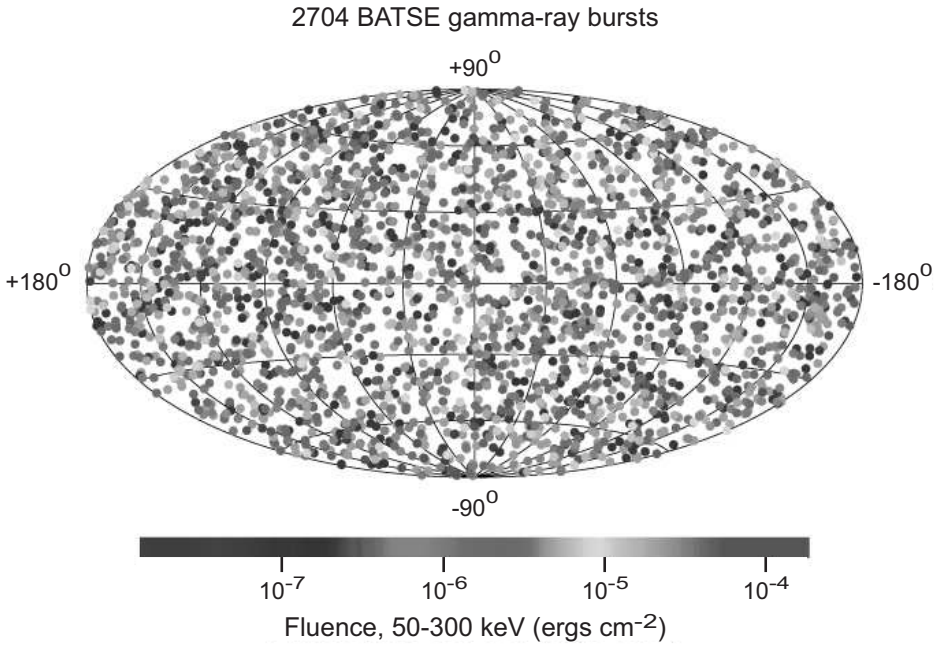
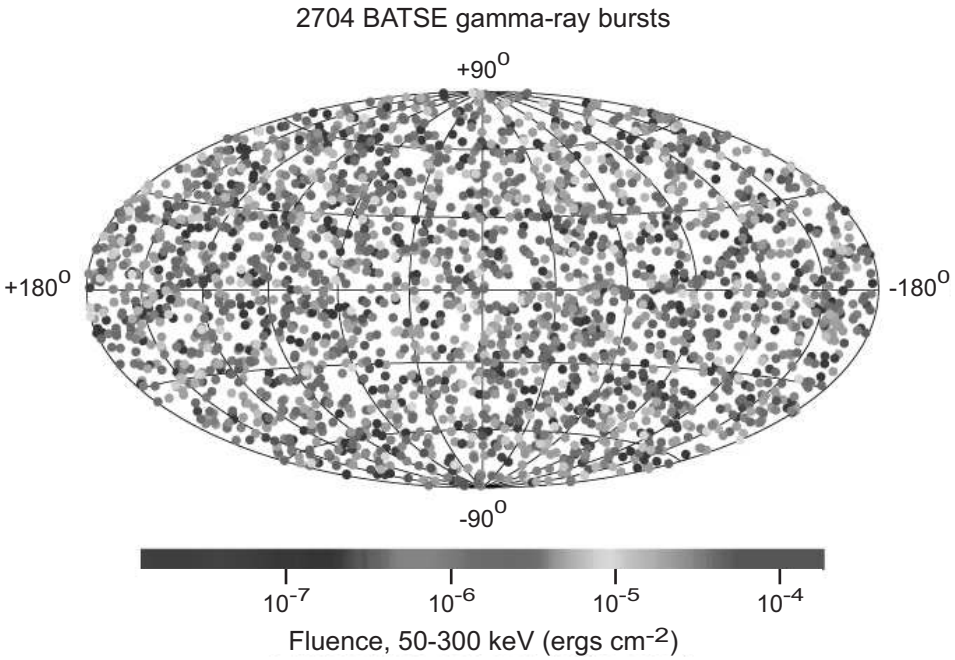


Figure 15.16: Location on the sky of 2704 gamma-ray bursts plotted in galactic coordinates with the grayscale indicating the fluence (energy received per unit area) of each burst. The data were recorded by the Burst and Transient Source Experiment (BATSE) of the Compton Gamma-Ray Observatory (CGRO). Highly isotropic distribution of events over a broad range of fluences argues strongly that they occur at cosmological distances.

Figure 15.16 shows the sky position of

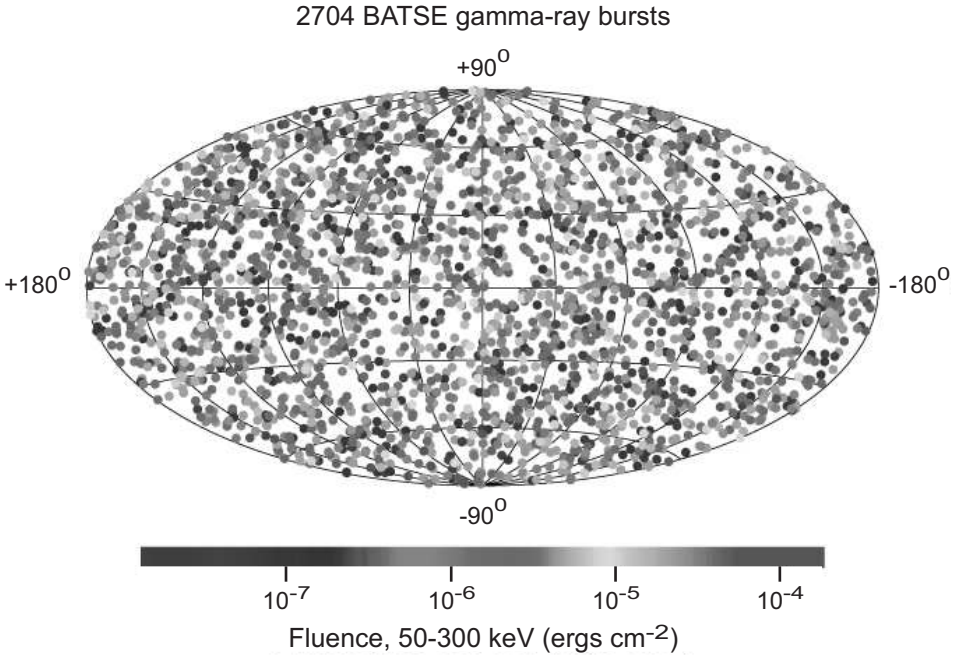
- *2704 gamma-ray bursts* observed by the
- *Burst and Transient Source Experiment (BATSE)* of the
- orbiting *Compton Gamma-Ray Observatory (CGRO).*

This plot indicates that the distribution of bursts is *isotropic on the sky.*



- If bursts were of local origin, such as in the disk of the galaxy,
- they should be concentrated along the galactic equator in this figure, *not randomly scattered over the sky*.
- This tells us that either
  - the gamma-ray bursts come from events at great distances (cosmological distances), or
  - perhaps they come from events in the more spherical halo of our galaxy.

Spectral Doppler shifts confirm that *gamma-ray bursts are cosmological in origin.*



- Their cosmological distances make GRB quite remarkable; to even be seen at such large distances they must correspond to events in which
  - energy *comparable to a supernova* is liberated
  - in a short burst in the *form of gamma-rays*.
- Furthermore, the mechanism producing the gamma-rays
  - must allow them to *escape with little interaction* with surrounding matter because
  - *even a small interaction with baryons* would down-scatter gamma-rays to light of longer wavelength,
  - *thermalizing the spectrum* (contrary to observation).

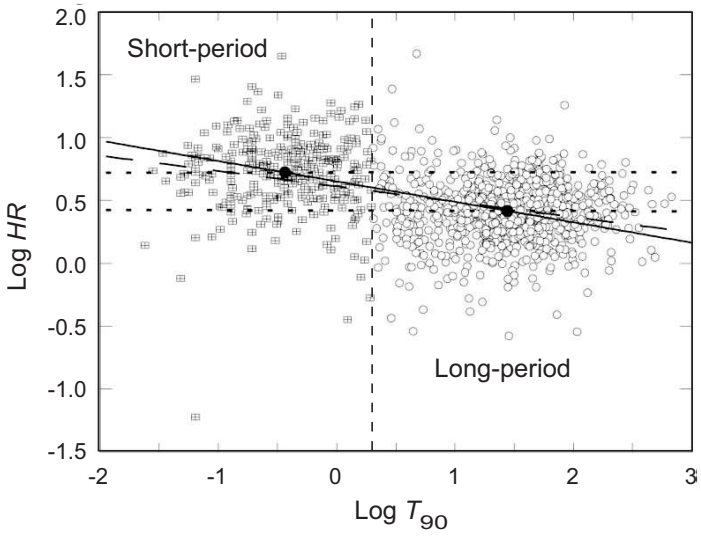


Figure 15.17: Hardness (propensity to contain higher-energy photons) of the spectrum versus duration of the burst, illustrating the separation of the GRB population into long, soft bursts and short, hard bursts. The parameter HR measures the hardness of the spectrum and T_{90} is defined to be the time from burst trigger for 90% of the energy to be collected. Bursts with T_{90} shorter than 2 seconds are classified as short-period and those with T_{90} greater than 2 seconds are classified as long-period bursts. This plot indicates that shorter bursts generally have a harder spectrum.

15.9.2 Two Classes of Gamma-Ray Bursts

There are *two classes* of gamma-ray bursts (Fig. 15.17):

1. *Short-period bursts*
   - last less than two seconds and
   - have harder (higher-energy) spectra.
2. *Long-period bursts*
   - have softer (lower-energy) spectra and
   - last from several seconds to several hundred seconds.

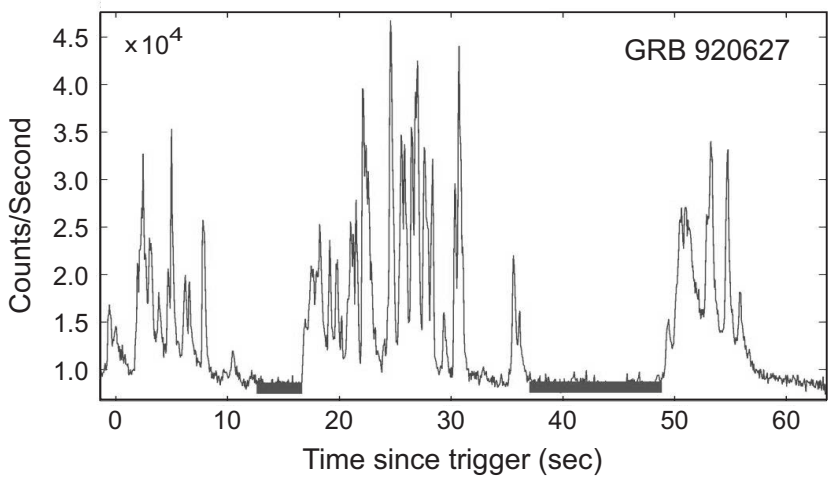


Figure 15.18: Time profile of a typical gamma-ray burst. This is an example of a long-period GRB.

The time profile of a *long-period burst* is shown in Fig. 15.18.

The existence of two classes of gamma-ray bursts indicated that there were *at least two mechanisms*.

- To determine these mechanisms, it was necessary first to understand *where they originated*.
- For example, were they *associated with known galaxies*?
- *Initial progress was slow* because BATSE observations could localize the position of a burst *only within several degrees* on the sky.

Therefore, it was very difficult to know exactly where to point telescopes to find evidence associated with the gamma-ray burst at other wavelengths.

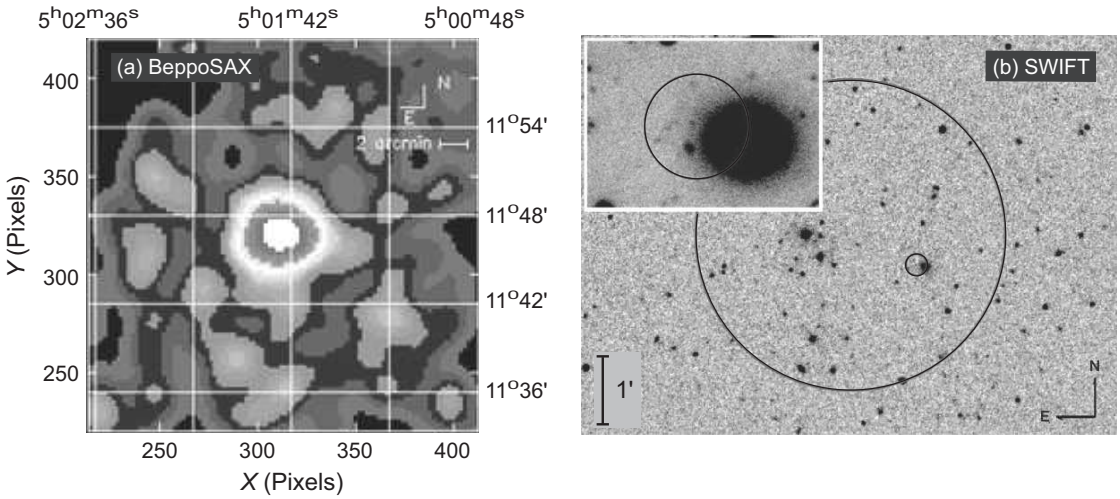


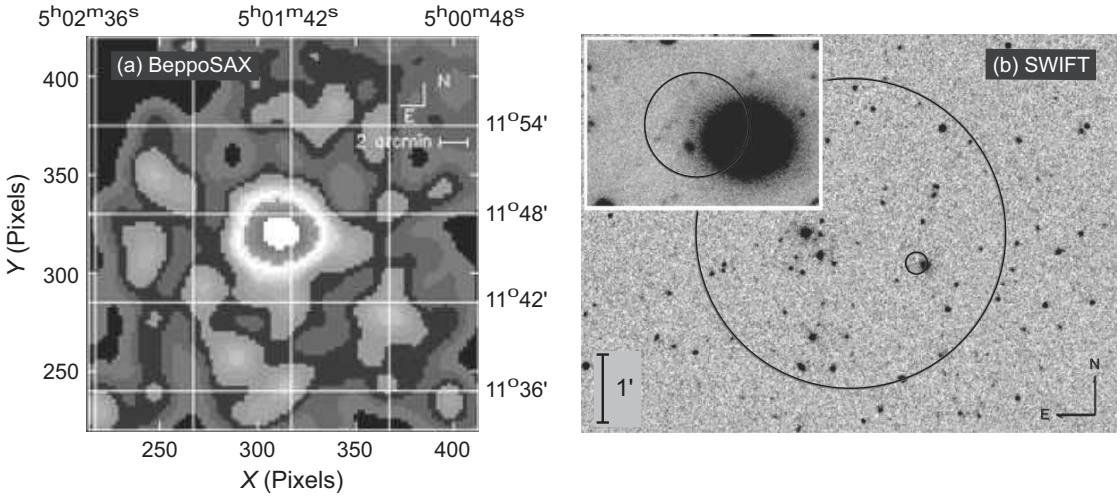
Figure 15.19: (a) First localization of an X-ray afterglow for a GRB, obtained by the satellite BeppoSAX. (b) Optical association of short-period GRB 050509B with a large elliptical galaxy at a redshift of $z = 0.225$ by SWIFT. The larger circle is the error circle for the Burst Alert Telescope (BAT). The smaller circle is the error circle for the X-Ray Telescope (XRT), which slewed to point at the event when alerted by the BAT. The XRT error circle is shown enlarged in the inset at the upper left, suggesting that the GRB occurred in the outer regions of a large elliptical galaxy (dark blob partially overlapped by the XRT error circle).

15.9.3 Localization of Gamma-Ray Bursts

A major breakthrough came when it became possible to correlate some GRB with *visible, RF, IR, UV, and X-ray* sources.

- BeppoSAX could pinpoint the position of X-rays following GRB with *2 arc-minute resolution* in a matter of hours.
- This allowed other instruments to look quickly at the site.

Figure 15.19(a) shows an *X-ray transient* observed by BeppoSAX at the location of a gamma-ray burst.



- A localization for a *short-period burst* by the SWIFT satellite is illustrated in the figure above right.
- These *transients* or *afterglows* are thought to be associated with
- a *rapidly fading fireball* produced by the primary gamma-ray burst.

Correlation of GRB with sources at other wavelengths

- have allowed *distances to be estimated* to GRB because
- spectral lines and their associated *redshifts* have been observed in the transients after the burst.

These observations show conclusively that *GRB occur at cosmological distances*.

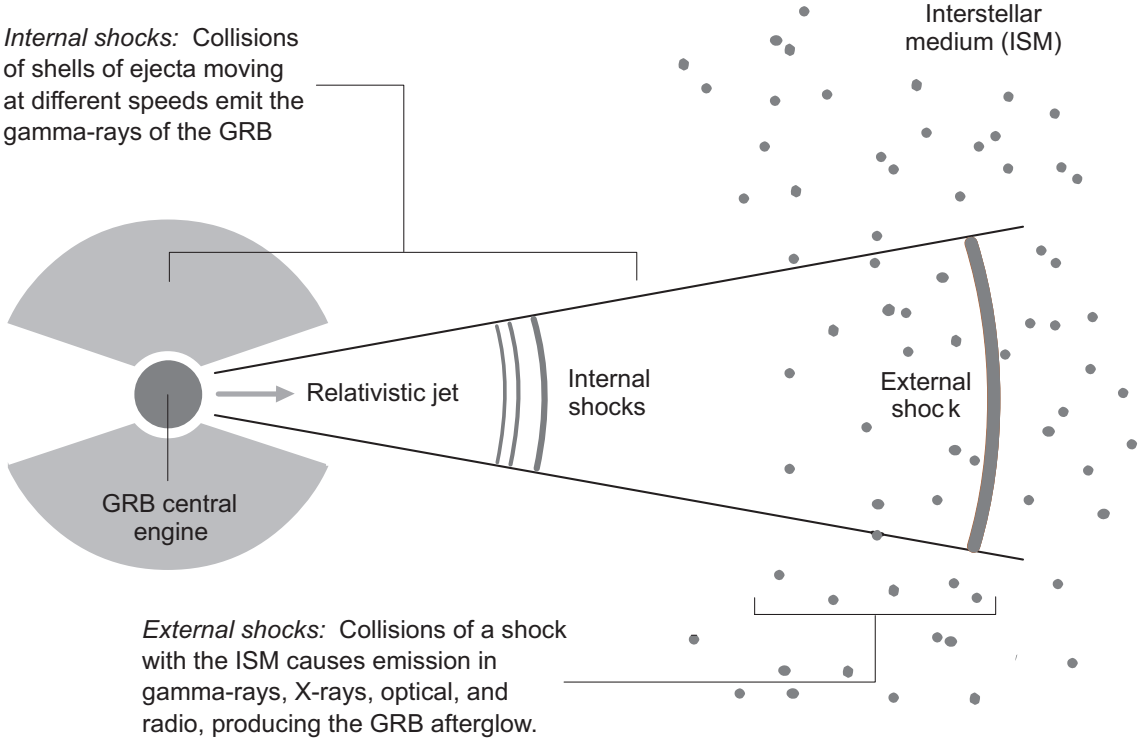


Figure 15.20: Relativistic fireball model for transient afterglows following gamma-ray bursts. In this model the internal shocks in the relativistic jet produced by clumps of ejected material with different velocities overtaking each other produce the gamma-rays and the external shocks resulting from the jet impacting the interstellar medium produce the afterglows.

There is broad agreement that the transient afterglows following gamma-ray bursts

- are described by the *fireball model* of Fig. 15.20.
- In this model some generic central engine deposits a large amount of energy in a small volume of space,
- which produces an *expanding relativistic fireball*.

The fireball is then responsible for the observed afterglow.

15.9.4 Necessity of Ultrarelativistic Jets

Gamma-ray bursts must involve *ultrarelativistic jets* because observed prompt emission is *nonthermal*.

- If the jet were *nonrelativistic* the ejecta would be *optically thick to pair production*.
- This would *thermalize the energy* (in contradiction to the *observed nonthermal spectrum* for GRB).
- The requirement that gamma-ray bursts be produced by ultrarelativistic jets can be understood in terms of
- the *opacity of the medium* with respect to formation of electron–positron pairs by $\gamma\gamma \to e^+e^-$.

Let us now elaborate on this crucial point.

15.9.5 Optical Depth for a Nonrelativistic Burst

We first assume that the burst involves nonrelativistic velocities.

- The initial spectrum is *nonthermal*.
- The number of counts $N(E)$ as a function of gamma-ray energy can be approximated for particular ranges as a power law,

$$N(E)dE \propto E^{-\alpha}dE,$$

  where the *spectral index* $\alpha \sim 2$ for typical cases.
- Because the *observed spectrum is nonthermal*
  - the medium must be *optically thin* (low opacity), since
  - scattering in an optically-thick (that is, highly-opaque) medium would *quickly thermalize the photons*.

Let's consider the optical depth for pair production associated with a typical gamma-ray burst to see whether this condition can be fulfilled.

For the reaction $\gamma\gamma \to e^+e^-$ to occur,

- *Energy conservation* requires the two photons with energies E_1 and E_2, respectively, to satisfy

$$(E_i E_2)^{1/2} \geq m_e c^2,$$

  where m_e is the *mass of an electron*.

- If f is the fraction of photon pairs fulfilling this condition, the *optical depth with respect to* $\gamma\gamma \to e^+e^-$ is then

$$\tau_0 = \frac{f\sigma_T F D^2}{R^2 m_e c^2} \simeq \frac{f\sigma_T F D^2}{\delta t^2 m_e c^4},$$

  - $\sigma_T = 6.652 \times 10^{-25}\,\mathrm{cm}^2$ is the *Thomson scattering cross section* for electrons,
  - F is the observed *fluence* for the burst,
  - D is the *distance to the source*, and
  - R is its *size*, which can be related to the observed *period* δt *for time structure* in the burst by $R = c\delta t$.

- Optical depth estimated using this formula are *enormous* ($\tau \sim 10^{14}$).

- This is *completely inconsistent* with the $\tau \lesssim 1$ required by the nonthermal GRB spectrum.

Nonrelativistic jets clearly won't do; what about relativistic jets?

15.9.6 Optical Depth for an Ultrarelativistic Burst

The above considerations will be altered in two essential ways if the burst is instead *ultrarelativistic* with a Lorentz factor $\gamma \gg 1$, so that *special-relativistic kinematics* apply:

1. The *blueshift* of the emitted radiation will modify the fraction f of photon pairs that have sufficient energy to make electron–positron pairs.
2. The *size* R of the emitting region will be altered by relativistic effects.

Specifically,

- The observed photons of frequency ν and energy $E = h\nu$ have been blueshifted from their energy in the rest frame of the GRB by a factor γ.
- Thus the source energy E_0 was lower than the observed energy E by a factor of γ^{-1} and $E_0 \sim h\nu/\gamma$.
- This means that *fewer photon pairs have sufficient energy to initiate* $\gamma\gamma \to e^+e^-$ than was inferred assuming nonrelativistic kinematics.
- From the spectrum

$$N(E)dE \propto E^{-\alpha}dE,$$

this means that f should be multiplied by a factor of $\gamma^{-2\alpha}$.

- Furthermore, relativistic effects *increase the size of the emitting region* by a factor of γ^2 over that inferred from the time period δt,
- so R should be multiplied by a factor of γ^2.
- Incorporating these corrections, the *ultrarelativistic modification is*

$$\tau \simeq \frac{\tau_0}{\gamma^{4+2\alpha}},$$

  where τ_0 is the result with no ultrarelativistic correction.
- Thus, even if τ_0 is very large an *optically thin medium* results if γ is large enough.

Typical estimates are that an optically thin medium requires $\gamma \sim 100$ or larger.

Observational confirmation that *gamma-ray bursts are associated with the large values of* γ comes from the observed location of *breaks* in the afterglow lightcurves.

- These breaks are thought to indicate the time when the initially-relativistic afterglow begins to slow rapidly through interactions with the interstellar medium.
- This in turn can be related to the *opening angle of the jet* that produced the afterglow.
- Such analyses typically find
  - *jet opening angles* in the range $\Delta\theta = 10 - 20^\circ$, which suggests
  - *Lorentz factors* $\gamma \sim 100$ or larger for many gamma-ray bursts.

Thus afterglow lightcurve breaks indicate directly that *gamma-ray bursts are produced by ultrarelativistic jets*.

15.9.7 Implications of Ultrarelativistic Beaming

The GRB beaming mechanism implies that *a fixed observer sees only a fraction of all gamma-ray bursts*.

- The ultrarelativistic nature of the jets means that the gamma-rays are *highly beamed in direction*.
- Afterglows are not strongly beamed after slowing, so
- they could be detected even for a GRB not on-axis (not aimed toward Earth).

Ultrarelativistic beaming for gamma-ray bursts solves a *potential energy-conservation problem*.

- If the energy from detected bursts were assumed to be *emitted isotropically*,
- total energies exceeding 10^{54} erg would be inferred for some gamma-ray bursts. This is comparable to
- the *rest mass energy of the Sun*, which would be difficult to explain by any mechanism that conserves energy.
- However, if GRBs are *emitted as collimated jets*, then
- total energy release would be much smaller than inferred by viewing it on-axis and assuming it to be isotropic,

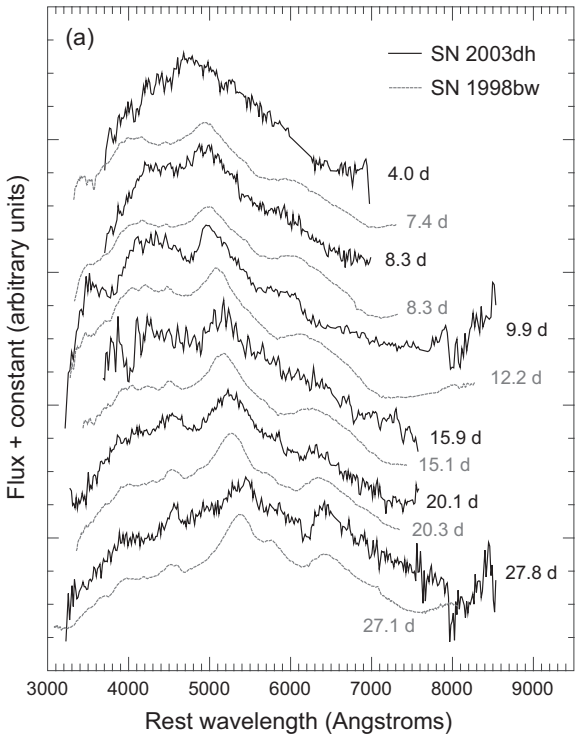
This places GRBs more in the total-energy range of well-studied events like supernova explosions.

15.9.8 Association of GRB with Galaxies

The *localization provided by afterglows* has permitted a number of long-period and short-period GRB to be associated with distant galaxies.

1. *Long-period (soft)* gamma-ray bursts
   - appear to be strongly correlated with *star-forming regions*.
   - (They have a strong correlation with blue light in host galaxies.)
2. *Short-period (hard)* gamma-ray bursts
   - are generally fainter and sampled at smaller redshift than long-period bursts.
   - They are *not correlated with star-forming regions*.
3. There is some evidence that *long-period bursts* are preferentially found in *star-forming regions having low metallicity*.

These observations provide further evidence that *long-period and short-period bursts are initiated by different mechanisms*.

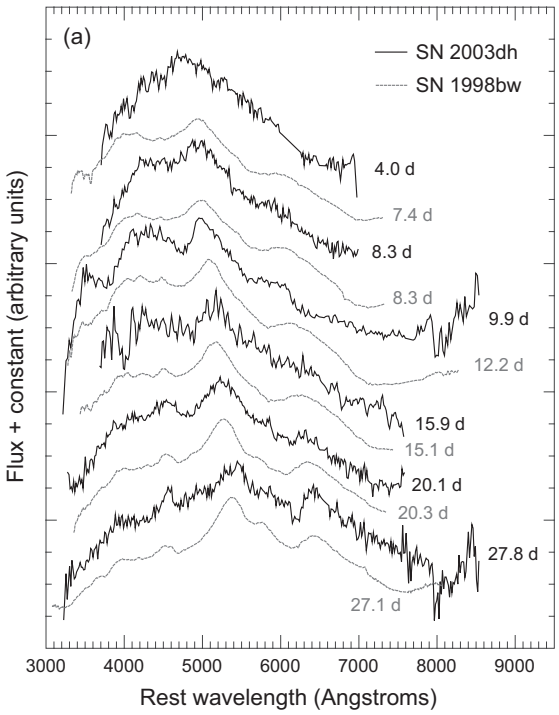


(b) Wolf-Rayet star HD56925

Figure 15.21: (a) Comparison of time evolution for spectral bumps in the rest-frame optical spectrum of SN2003dh (GRB 030329) in black compared with a reference supernova SN1998bw in gray. The initially rather featureless spectrum of the GRB 030329 afterglow develops bumps similar to those of the supernova SN1998bw over time, suggesting that as the afterglow produced by deposition of the GRB energy fades, an underlying spectrum characteristic of a supernova explosion is revealed. Hence GRB 030329 is also denoted by a supernova label, SN2003dh. (b) A Wolf–Rayet star (black arrow) surrounded by shells of gas that it has emitted. These massive, rapidly-spinning stars may be progenitors of Type Ib and Type Ic core collapse supernovae, and hence of long-period gamma-ray bursts.

15.9.9 Association of Long-Period GRB with Supernovae

There is a relationship between long-period gamma-ray bursts and core collapse supernovae, as suggested by Fig. 15.21.



(b) Wolf-Rayet star HD56925

There is a connection between *long-period GRB* and particular types of core-collapse supernovae called *Types Ib and Ic*.

- The *Ib and Ic mechanisms* are thought to involve *core collapse* in a rapidly-rotating, massive (15–30 $M_\odot$) *Wolf–Rayet star* [Fig. (b) above].
- These stars exhibit *large mass loss* and can *shed their H and even He envelopes* before their cores collapse.
  - It is thought that in a Type Ib supernova the *H shell has been removed* before collapse of the core, and
  - in a Type Ic supernova *both the H and He shells have been removed* before the stellar core collapses.
- These stars are so massive that they can *collapse directly to a rotating (Kerr) black hole*, instead of a neutron star.

The Kerr black hole formed from core collapse of a Wolf–Rayet star is thought to *power a long-period gamma-ray burst* in conjunction with a Type Ib or Ic supernova.

- On the other hand, there is *little observational evidence* that short-period bursts are associated with
  - star-forming regions, or
  - with supernovae.
- As we will see, the favored mechanism for short-period bursts involves the
  - formation of an accreting Kerr black hole by
  - merger of two neutron stars (or a neutron star and a black hole).

We now turn to a more detailed discussion of the mechanism for both short-period and long-period gamma-ray bursts.

15.9.10 Characteristics of Gamma-Ray Bursts

As preparation for discussing the mechanisms powering gamma-ray bursts, let's summarize their basic characteristics.

1. Isotropic sky distribution suggests a *cosmological origin*.
   - This has been confirmed by direct *redshift measurements* on emission lines in GRB afterglows.
   - Known redshifts for gamma-ray bursts range up to $z = 8.2$, with an average $z \sim 1$ (as of 2016).
2. The *spectrum is non-thermal*, typically peaking around 200 keV and extending perhaps as high as GeV.
3. Duration of individual bursts *spans 5 orders of magnitude*:
   - from about 0.01 seconds up to
   - several hundred seconds.
4. There is a *variety of time structure*, from rather smooth to millisecond fluctuations (implying a compact source).
5. Bursts are *ultrarelativistically beamed*.
6. *Transient afterglows* are observed that are described phenomenologically by the *fireball model* introducted earlier.
7. There appear to be *two classes of bursts*:
   - *long-period* gamma-ray bursts, and
   - *short-period* gamma-ray bursts,

   triggered by different events.

15.9.11 Mechanisms for the Central Engine

GRB central engines are not well understood, but an acceptable model must embody at least the following features:

1. All models require *highly-relativistic jets* to account for observed properties of gamma-ray bursts.
   - Lorentz γ factors of at least 200, perhaps as large as 1000.
   - Jets focused with opening angles ~ 0.1 rad and
   - power as large as $\sim 10^{52}$ erg
   - Long-period bursts must deliver $\sim 10^{52}$ erg to a much larger angular range (~ 1 rad) if the burst is accompanied by a supernova, and
   - the central engine must be capable of operating for 10 seconds or more in long-period bursts.
2. Large energies and potentially long timescales imply *compact-object accretion* to meter gravitational energy.
3. Thus, acceptable models must produce accretion disks.

Almost the only possible explanation is

- *gravitational collapse* to a *Kerr black hole*
- with formation of a large *accretion disk*.

At least *two mechanisms* can lead to this.

We now discuss in more detail the *two general classes of models* thought to account for GRBs.

1. *Short-Period Bursts:* The *merger of two neutron stars* (or a neutron star and a black hole), with
   - jet outflow perpendicular to the merger plane
   - producing a burst of gamma-rays as
   - the two objects collapse to a Kerr black hole.
2. *Long-Period Bursts:* A *hypernova,* where
   - a spinning massive star collapses to a Kerr black hole and
   - jet outflow from the region surrounding this collapsed object produces a burst of gamma-rays.

The unifying theme is the collapse of stellar-size amounts of spinning mass to a black hole central engine that powers the burst.

In both the hypernova and the merger of neutron stars,

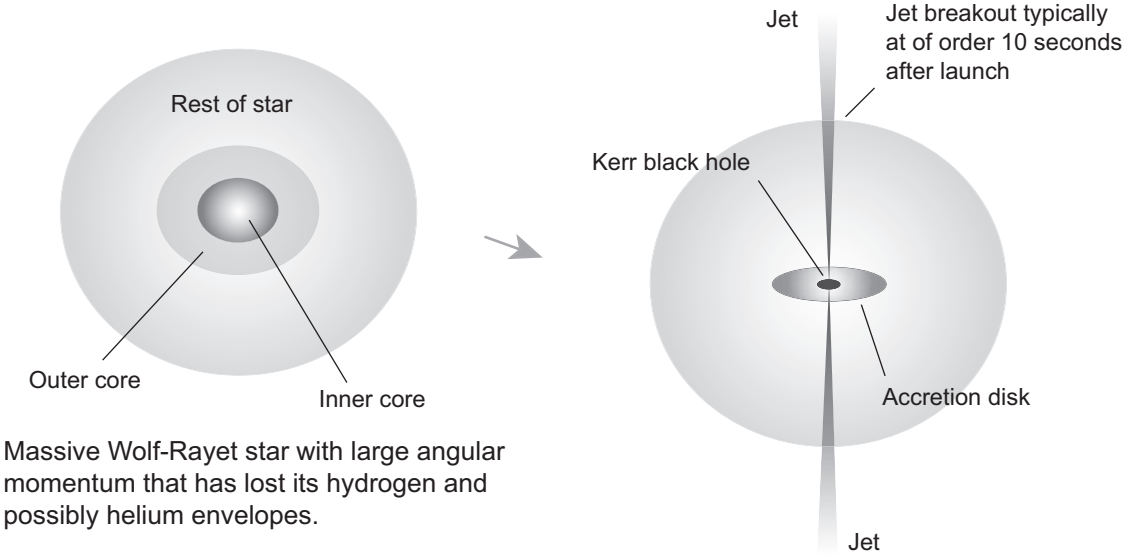
- the outcome is a
  - Kerr black hole having
  - large angular momentum and
  - strong magnetic fields,

  surrounded by an accretion disk of matter that hasn't yet fallen into the black hole.
- This scenario likely leads to *highly-focused relativistic jet outflow* on the polar axes of the Kerr black hole.
- These jets are powered by some combination of
  - *rapid accretion* from the disk,
  - *neutrino–antineutrino annihilation*,
  - strong *coupling to magnetic fields*.

Thus, the GRB black hole engine has many similarities with AGN–quasar engines, but on a stellar rather than galactic-core scale.

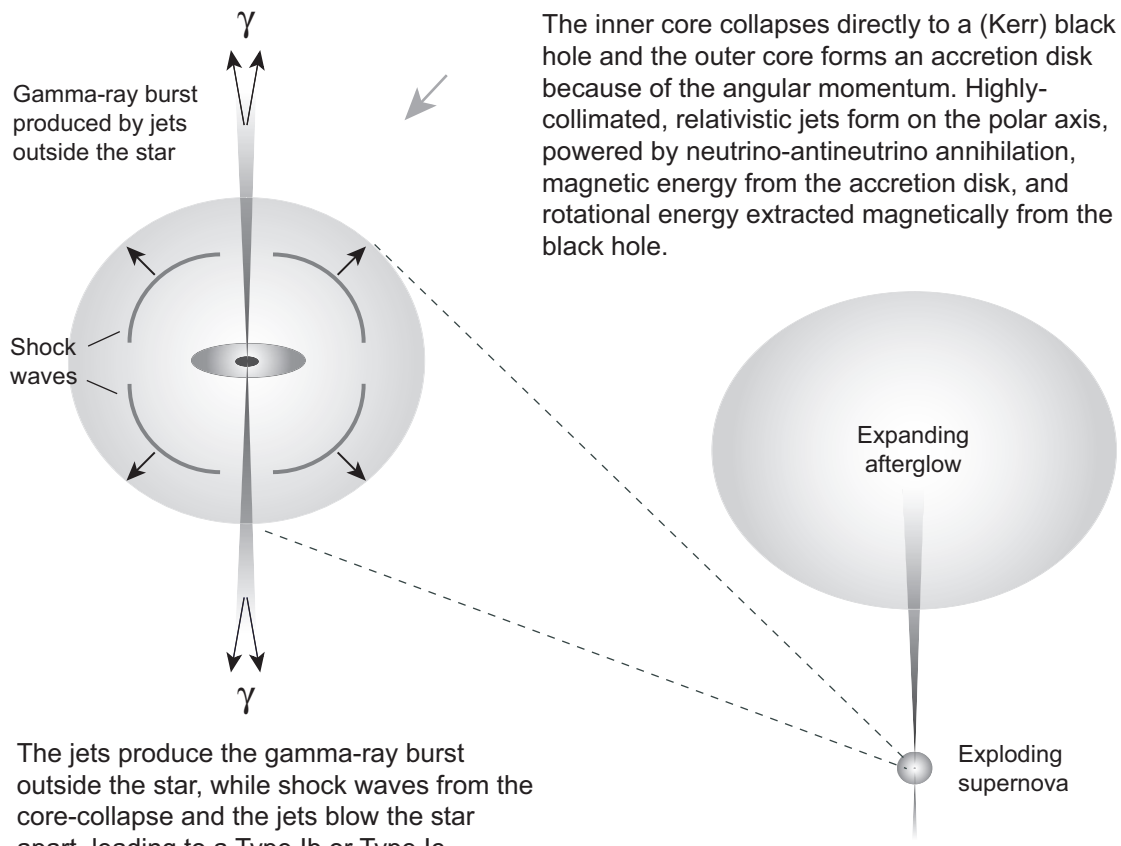
15.9.12 The Collapsar Model and Long-Period Bursts

An overview of the collapsar model is shown in Fig. 15.22 (next page).



Massive Wolf-Rayet star with large angular momentum that has lost its hydrogen and possibly helium envelopes.

The inner core collapses directly to a (Kerr) black hole and the outer core forms an accretion disk because of the angular momentum. Highly-collimated, relativistic jets form on the polar axis, powered by neutrino-antineutrino annihilation, magnetic energy from the accretion disk, and rotational energy extracted magnetically from the black hole.



The jets produce the gamma-ray burst outside the star, while shock waves from the core-collapse and the jets blow the star apart, leading to a Type Ib or Type Ic supernova.

At much larger distances the interaction of the jets with the surrounding medium begins to produce the afterglow that will be detected at longer wavelengths.

Figure 15.22: Collapsar model for long-period GRB and Type Ib or Ic supernova.

Simulations of relativistic jets breaking out of a Wolf–Rayet star in a collapsar model and a Wolf–Rayet star 20 seconds after core collapse are shown in Fig. 15.23 and Fig. 15.24 on the following page.

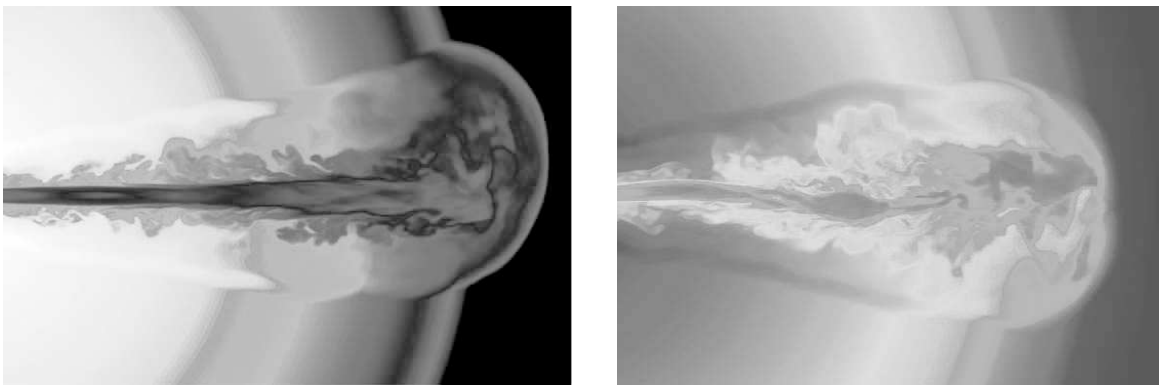


Figure 15.23: Simulations of relativistic jets breaking out of Wolf–Rayet stars. Breakout of the $\gamma \sim 200$ jet is 8 seconds after launch from the center of a 15 $M_\odot$ Wolf–Rayet star.

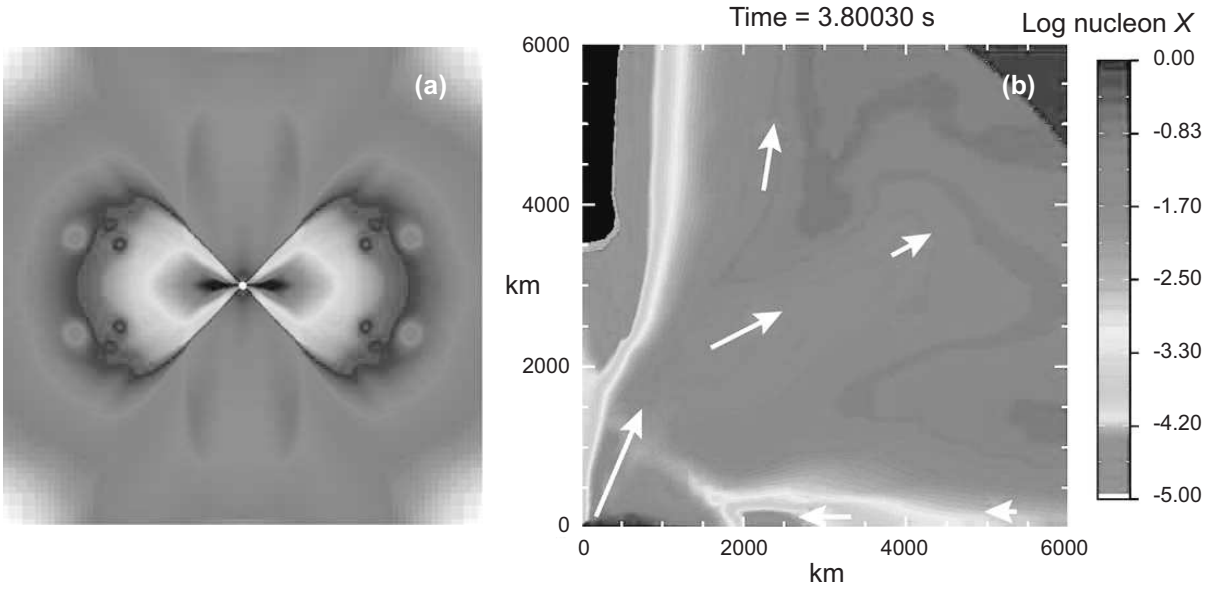
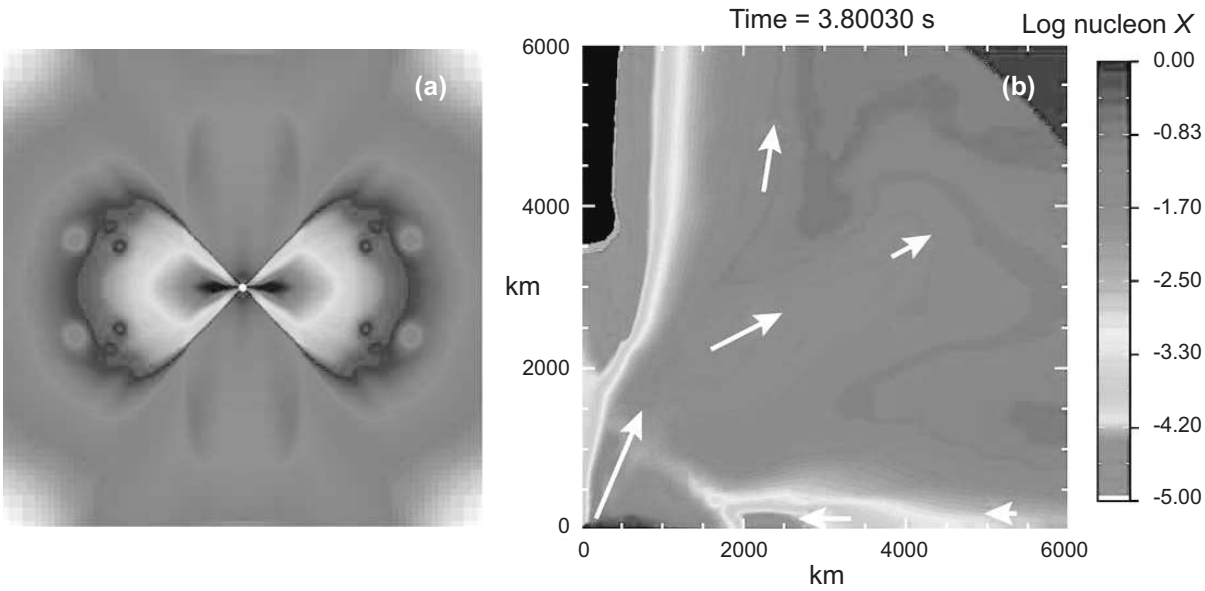


Figure 15.24: (a) A rapidly-rotating 14 $M_\odot$ Wolf–Rayet star, 20 seconds after core collapse. The polar axis is vertical, the density scale is logarithmic, and the 4.4 $M_\odot$ Kerr black hole has been accreting at $\sim 0.1 M_\odot\,\mathrm{s}^{-1}$ for 15 seconds at this point. (b) Simulation of the nucleon wind blowing off the accretion disk in a collapsar model. The gray-scale contours represent the log of the nucleon mass fraction X and arrows indicate the general flow.



In the figure above right a strong nucleon wind blowing off the collapsar accretion disk is shown. This wind

- produces the supernova and
- synthesizes the ^{56}Ni that powers the lightcurve of the supernova by radioactive decay.

The GRB and the supernova are powered in different ways in the collapsar model:

1. The GRB is powered by a relativistic jet deriving its energy from neutrino–antineutrino annihilation or rotating magnetic fields.
2. The accompanying supernova is powered by the disk wind illustrated in this figure.

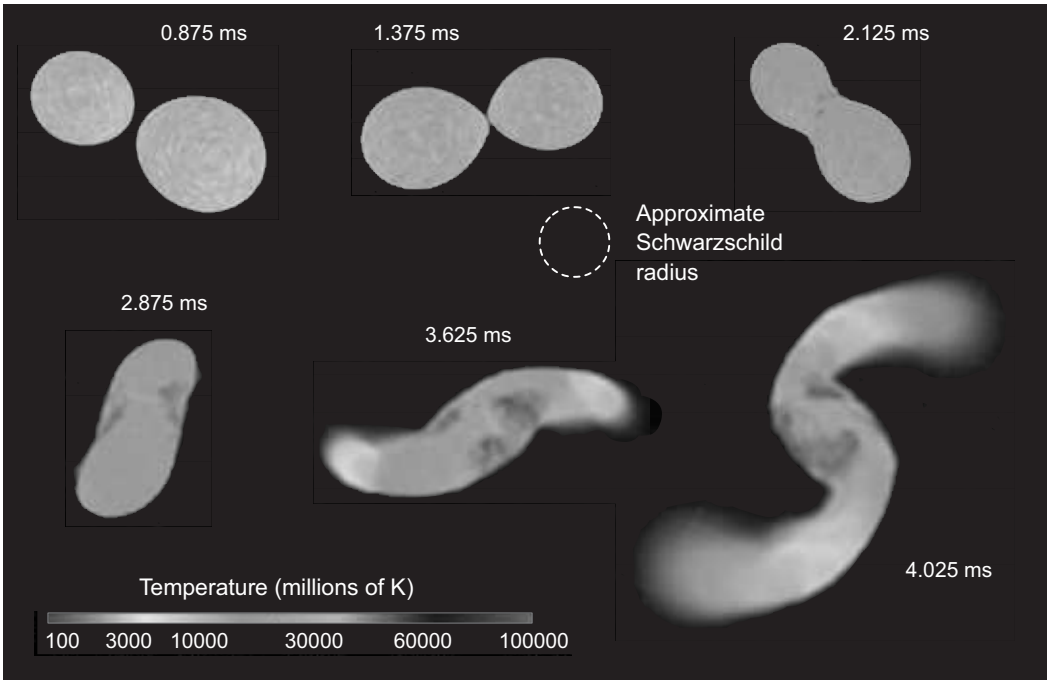


Figure 15.25: Simulation of the merger of two neutron stars. The elapsed time is about 3 ms and the approximate Schwarzschild radius for the combined system is indicated. The rapid motion of several solar masses of material with large quadrupole distortion and sufficient density to be compressed near the Schwarzschild radius indicates that this merger should be a strong source of gravitational waves (Source: S. Rosswog simulation).

15.9.13 Merging Neutron Stars and Short-Period Bursts

The collapser model is not valid for short-period bursts because

- they are not observed in star-forming regions.
- Instead, short-period gamma-ray bursts involve binary neutron stars merging to form a Kerr black hole.

A simulation of a neutron star merger is shown in Fig. 15.25.

An illustration of how the magnetic fields of the neutron stars can be magnified in the merger is illustrated in Fig. 15.26 (following page).

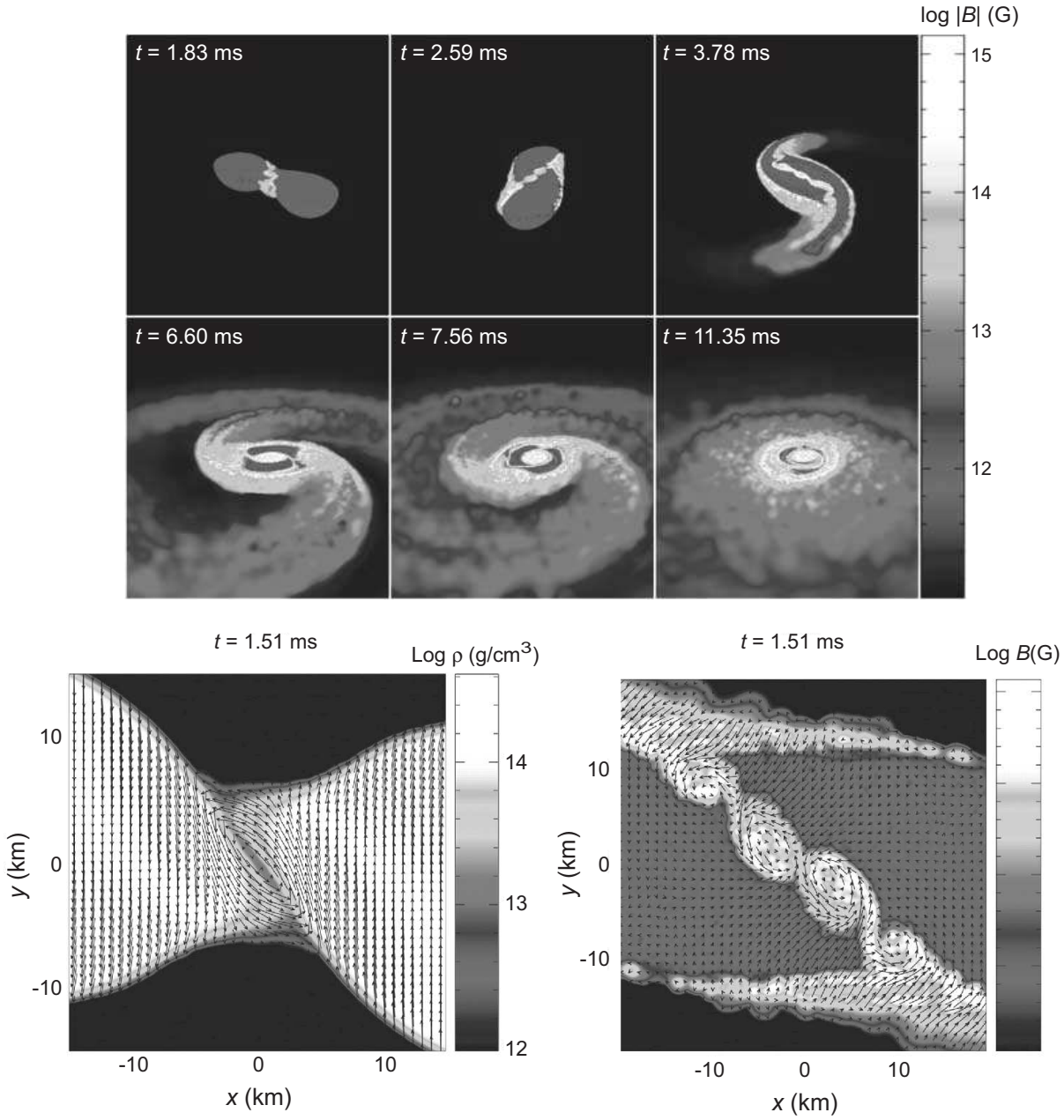


Figure 15.26: (Top) A simulation of merging neutron stars with the magnetic field strength indicated by the grayscale. (Bottom) Amplification of magnetic fields in merging neutron stars for the simulation shown in the top figure. Arrows indicate the direction of the magnetic field. In the simulation the shear produced at the merger boundary is capable of substantially amplifying the already significant magnetic fields that are present (Stefan Rosswog).

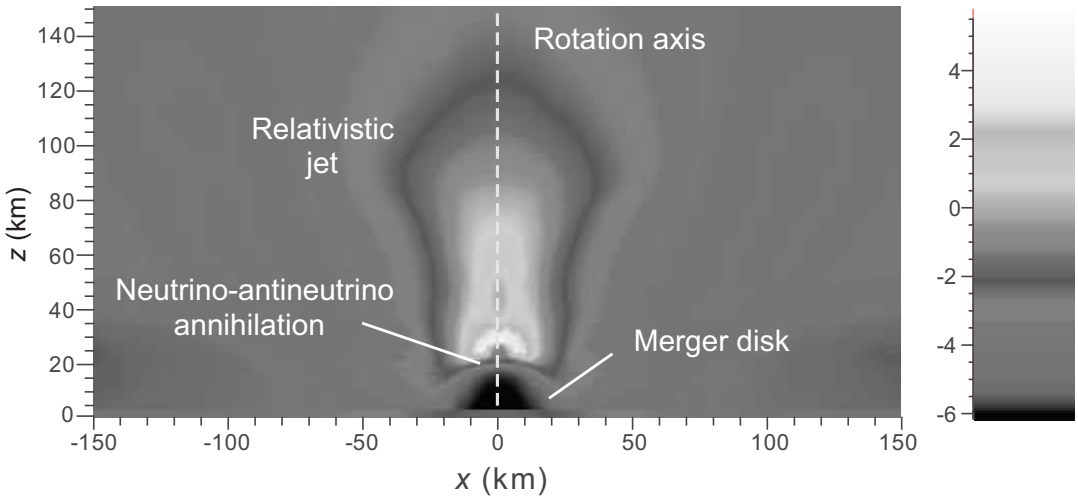


Figure 15.27: Illustration of relativistic jet powered by neutrino–antineutrino annihilation in a neutron-star merger.

- One possible mechanism for powering the jets in GRB produced by neutron star mergers is *neutrino–antineutrino annihilation* above and below the plane of the merger disk, as illustrated in Fig. 15.27.
- Another is *accretion onto the rotating black hole* from the surrounding disk in the merger.
- Another is to *tap the power of the very strong magnetic fields* that are expected, for example as illustrated in Fig. 15.28 on the following page.

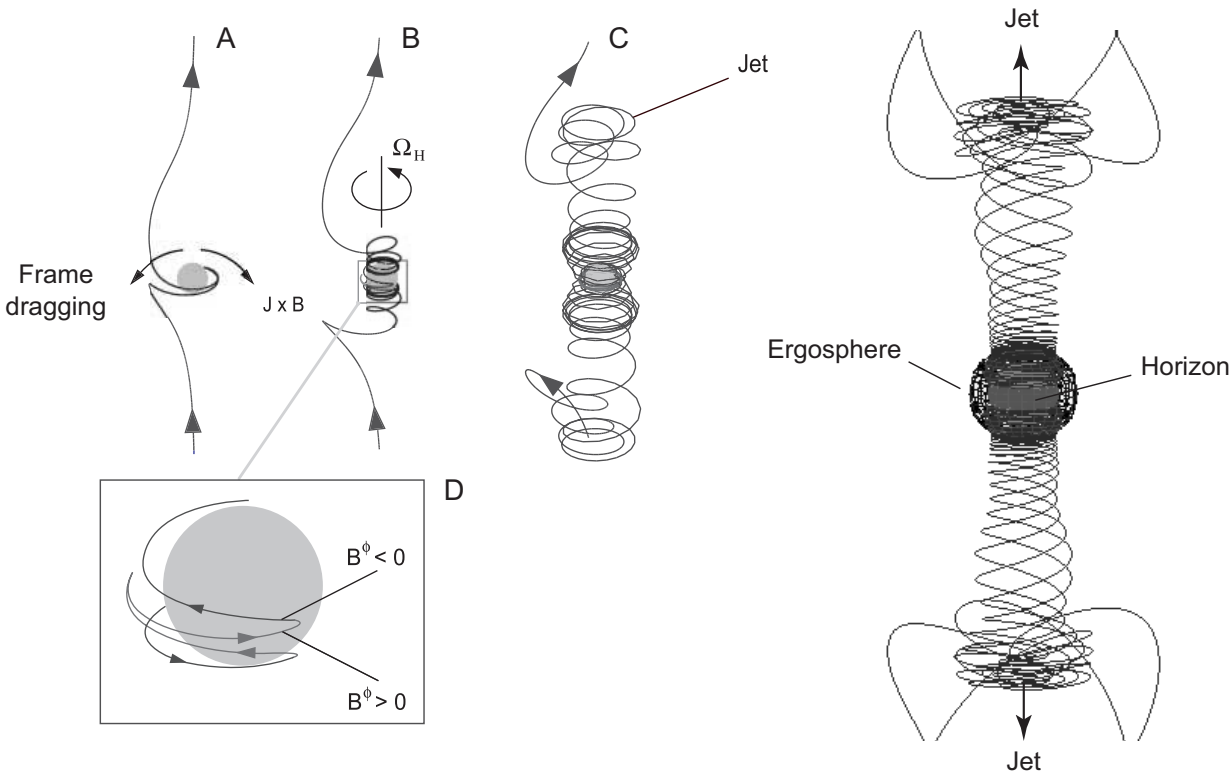


Figure 15.28: Illustration of relativistic jet powered by frame dragging of magnetic fields by a Kerr black hole.

In the model illustrated in Fig. 15.28,

- the *frame-dragging effects* associated with the Kerr black hole
- *wind the flux lines* associated with the magnetic field around the black hole and
- *spiral them off the poles* of the rotation axis,

thus *powering relativistic jets*.

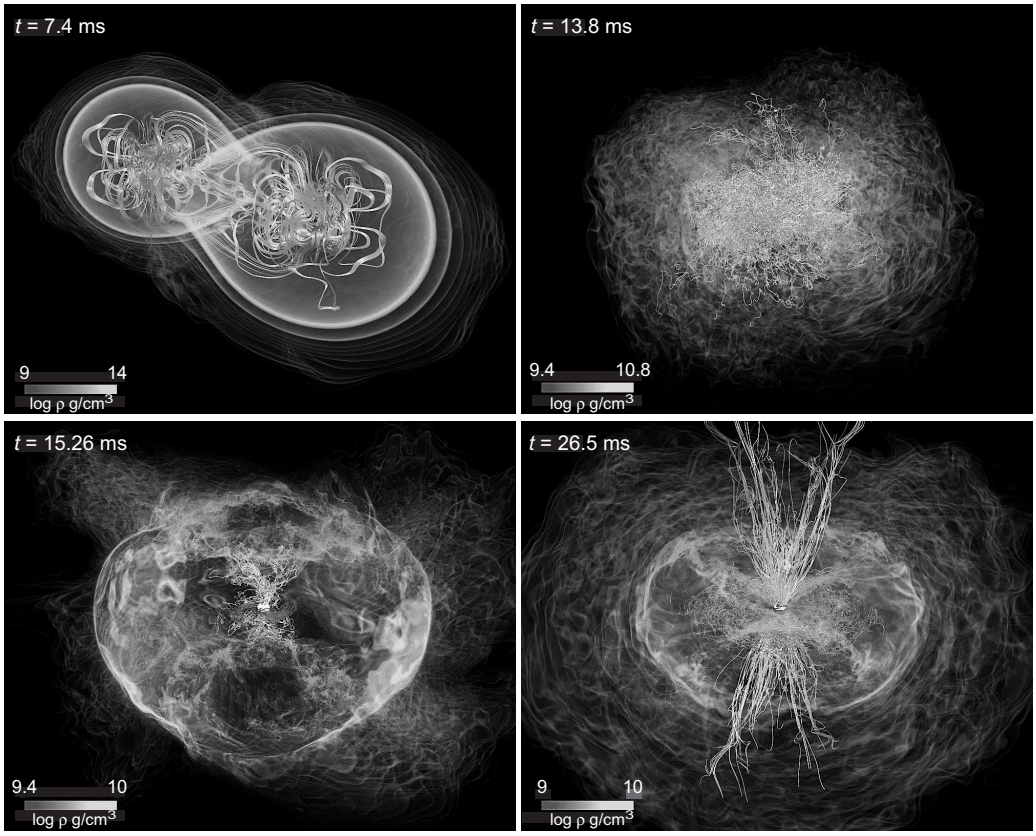


Figure 15.29: Neutron star merger simulation with strong magnetic fields. Grayscale indicates the log of the density; curves are magnetic field lines.

Simulation of a neutron-star merger to form a Kerr black hole with strong magnetic fields is illustrated in Fig. 15.29.

- The first panel shows the state shortly after initial contact.
- The second displays a merged neutron star configuration.
- In the bottom panels a Kerr black hole has formed with a disk around it, and
- the magnetic field is wound up to $\sim 10^{15}$ gauss, with the field lines directed in the polar direction.

15.9.14 Gamma-Ray Bursts and Gravitational Waves

Finally we note that

- neutron star mergers should
  - produce gravitational waves of sufficient strength to be
  - observable in earth-based gravitational wave detectors.
- This possibility will be discussed further in later chapters.
- This also raises the intriguing possibility of *multimessenger astronomy* where, for example,
- a gamma-ray burst might be observed in coincidence with gravitational waves from a binary neutron star merger.

Such an observation would presumably have a large impact on

- the understanding of neutron-star structure,
- the mechanism for gamma-ray bursts, and
- the nature of gravitational wave sources.

As we shall discuss more extensively later,

- in 2017 a
  - *short-period gamma-ray burst* and
  - *afterglows* at various wavelengths
  - were observed *in coincidence with gravitational waves*

  from a binary *neutron star merger*.
- This represents the *first multimessenger gravitational-wave event* observed in astronomy, and
- provides the first strong evidence for the conjectured *neutron star merger mechanism* for short-period gamma-ray bursts.

In addition it

- provides the first direct evidence that many of the heavier elements are synthesized in neutron star mergers,
- places an extremely strong constraint on any deviation of the speed of gravity from the speed of light,
- suggests a new way to determine the Hubble constant for expansion of the Universe, and
- begins to place new constraints on the deformability of neutron-star matter and the neutron star equation of state.

Part III

Cosmology

Chapter 16

The Hubble Expansion

The observational characteristics of the Universe coupled with theoretical interpretation to be discussed further in subsequent chapters, allow us to formulate a *standard picture* of the nature of our Universe.

16.1 The Standard Picture

The standard picture rests on but a few ideas, but they have profound significance for the nature of the Universe.

16.1.1 Mass Distribution on Large Scales

Observations indicate that the Universe is

- *homogeneous* (no preferred place) and
- *isotropic* (no preferred direction),

when considered *on sufficiently large scales*.

- When averaged over distances of order 50 Mpc, the fluctuation in mass distribution is of order unity, $\delta M/M \simeq 1$
- When averaged over a distance of 4000 Mpc (comparable to the present horizon), $\delta M/M \leq 10^{-4}$

> Thus, averaged over a large enough volume, *no part of the Universe looks any different from any other part*.

- The idea that the Universe is *homogeneous and isotropic* on large scales is called the *cosmological principle*.

> The cosmological principle, as implemented in general relativity, is the *fundamental theoretical underpinning* of modern cosmology.

The cosmological principle should not be confused with the *perfect cosmological principle,* which was the underlying idea of the *steady state theory* of the Universe.

- In the perfect cosmological principle, the Universe is not only homogeneous in space but also *homogeneous in time*.
- Thus it looks the same not only from any place, but from any time.

This idea once had a large influence on cosmology but is no longer considered viable because it is inconsistent with modern observations that show *a Universe evolving in time*.

16.1.2 The Universe Is Expanding

Observations indicate that the *Universe is expanding*;

- the interpretation of general relativity is that this is because *space itself is expanding*.
- The distance ℓ between conserved particles is changing according to the *Hubble law*

$$v \equiv \frac{d\ell}{dt} = H_0 \ell,$$

deduced from redshift of light from distant galaxies.

- H_0 is the *Hubble parameter* (or Hubble "constant", but it changes with time; the subscript zero indicates that this is the value *at the present time*).
- The Hubble parameter can be determined by fitting the above equation to the redshifts of galaxies at known distances.
- The uncertainty in H_0 is sometimes absorbed into a dimensionless parameter h by quoting

$$H_0 = 100h\,\mathrm{km\,s^{-1}\,Mpc^{-1}} = 3.24 \times 10^{-18} h\,\mathrm{s^{-1}},$$

where h is of order 1.

- When we need a definite value, we will often assume

$$H_0 = 72\,\mathrm{km\ s^{-1}\ Mpc^{-1}},$$

so $h = 0.72$. *Note:* Units of H_0 are actually $(\text{time})^{-1}$.

- We may define a *Hubble length* L_{H} through

$$L_{\mathrm{H}} = \frac{c}{H_0} \simeq 4000\,\mathrm{Mpc}.$$

- Thus, for a galaxy lying a Hubble length away from us,

$$v = \frac{d\ell}{dt} = H_0 \frac{c}{H_0} = c,$$

- This implies that the recessional velocity of a galaxy further away than L_{H}
  - *exceeds the speed of light*,
  - if the observed redshifts are interpreted as Doppler shifts,
  - as it often is in popular-level discussions.

- *Not to worry:* We shall find that in general relativity
  - the redshift of the receding galaxies is *not a Doppler shift* caused by velocities in spacetime,
  - but is a consequence of the *expansion of space itself*,

  which *stretches the wavelength* of all light.

The light speed limit of special relativity applies to velocities *in spacetime, not to spacetime itself.*

It is important to understand that *local objects are not partaking* of the general Hubble expansion.

- The Hubble expansion *is not caused by a force.*
- It occurs in pure form only when forces between objects are negligible.
- Deviations from the Hubble expansion caused by forces between objects are called *peculiar motion* by astronomers.
- Smaller objects, such as our bodies, are held together by chemical (electrical) forces. They do not Hubble expand.
- Larger objects like planets, solar systems, and galaxies are also held together by forces, in this case gravitational in origin. They generally do not Hubble expand with the Universe either.

It is only on

- much larger scales (beyond superclusters of galaxies) that
- gravitational forces among local objects are sufficiently weak to cause

negligible perturbation on the overall Hubble expansion.

16.1.3 The Expansion is Governed by General Relativity

It is possible to understand much of the expanding Universe using *Newtonian physics* and *insights borrowed from relativity*.

- However, in the final analysis there are *serious technical and philosophical difficulties* that arise and that
- require replacement of Newtonian gravitation with the Einstein's general theory of relativity for their resolution.
- Central to these issues is the nature of space and time compared with that in classical Newtonian gravitation.
  - In relativity, space and time are not separate but enter as a unified spacetime continuum.
  - Even more fundamentally, space and time in relativity are not a passive background where events happen.
  - Relativistic space and time are not "things" but are abstractions expressing a relationship between events.

  Thus, in this view, space and time do not have a separate existence apart from events involving matter and energy.

- On fundamental grounds, the gravitational curvature radius of the Universe could be comparable to the radius of the visible Universe, implying the *necessity of general relativity* for the description of cosmology.

16.1.4 There Is a Big Bang in Our Past

Evidence suggests that the Universe expanded from an *initial condition of very high density and temperature*.

- This *hot, dense initial state* is called the *big bang.*
- The *popular (mis)conception* that the big bang was a gigantic explosion *is in error* because it conveys the idea
  - that it happened in space and time, and that expansion
  - is due to forces generated by this explosion.

> The general relativistic view of the big bang is that it did not happen *in spacetime* but rather that *space and time were created in the big bang*.

- "What happened before the big bang?" or "what is the Universe expanding into?" are meaningless because these questions *presuppose the existence of a spacetime background* upon which events happen.
- The big bang should be viewed not as an explosion but as an *initial condition for the Universe.*

> *Loosely* we may view the big bang as an "explosion" because of the *hot, dense initial state*. But then this "explosion" happens at all points in space: *there is no "center" for the big bang.*

General relativity implies that the initial state was a *spacetime singularity*.

- Whether general relativity is correct on this issue
  - will have to await a full theory of *quantum gravitation*, since
  - general relativity cannot be applied too close to the initial singularity (on scales below the *Planck scale*)
  - without incorporating the principles of quantum mechanics.
- However, for most (but not all!) issues in cosmology the question of whether there was an initial spacetime singularity is not relevant.
- For those issues, all that is important is that
  - once the Universe expanded beyond the Planck scale
  - it was *very hot and very dense*.

This *hot and dense initial state* is what we shall mean in simplest form when we refer to *the big bang*.

16.1.5 Particle Content Influences Evolution of the Universe

The Universe contains a variety of particles and their associated fields that influence strongly its evolution.

- The "ordinary" matter composed of things that we find around us is generally termed *baryonic matter.*

> *Baryons* are the *strongly interacting particles of half-integer spin* such as protons and neutrons.

- Baryonic matter is the most obvious matter to us.
- However data indicate that *only a small fraction of the total mass in the Universe is baryonic*.
- The bulk of the mass in the Universe appears to be in the form of *dark matter,* which is *easily detected only through its gravitational influence*.
- We *don't know what dark matter is*.
- A popular idea ascribes it to *undiscovered elementary particles*, but no evidence supports this so far.
- There is also growing evidence that the Universe is strongly influenced by *dark energy,* which
  - *permeates even empty space* and
  - effectively causes *gravity to become repulsive*.
- We do not know the origin of dark energy.

A fundamental distinction for particles and associated fields is whether they are *massless* or *massive*.

- Lorentz-invariant quantum field theories require that
  - *massless particles* must move with $v = c$ and that
  - *particles with finite mass* must move with $v < c$.
- Therefore, massless particles like photons, gravitons, and gluons, and nearly massless particles like neutrinos, are highly relativistic.
- Roughly, particles with rest mass m are non-relativistic at those temperatures T where $kT << mc^2$
  - Electrons have $mc^2 = 511\,\text{keV}$ and they are non-relativistic at temperatures below about 6×10^9 K.
  - Protons have a rest mass of 931 MeV and they remain non-relativistic up to temperatures of about 10^{13} K.
  - Conversely massless photons, gluons, gravitons, and nearly-massless neutrinos are *always relativistic.*
- In cosmology, it is common to refer to massless or nearly massless particles as *radiation.*
- Conversely, massive particles have $v << c$ (unless temperatures are extremely high) and are non-relativistic.
- Non-relativistic particles are termed *matter* (or *dust*).
- Non-relativistic matter has *low velocity* and *exerts little pressure* compared with relativistic matter.

16.1.6 There Is a Cosmic Microwave Background

Radiation was important in the evolution of the Universe.

The most important feature beyond Hubble expansion is that *the Universe is filled with a smooth and isotropic microwave photon background*.

- Any theoretical attempt to understand the standard picture must as a minimal starting point accomodate the
  - *Hubble expansion* of the Universe and the
  - *cosmic microwave background (CMB)* radiation.
- Conversely, precise measurement of tiny fluctuations in the CMB is turning cosmology into a quantitative science.
- The CMB currently peaks in the microwave region of the spectrum but its wavelength
  - has been steadily redshifting since the big bang and
  - it was originally much higher energy radiation.
- For example, when electrons combined with protons to make neutral hydrogen 400,000 years after the big bang, the spectrum of the current CMB peaked in the *near-IR*.

The CMB contains $> 90\%$ of the current photon energy density (less than 10% is in starlight).

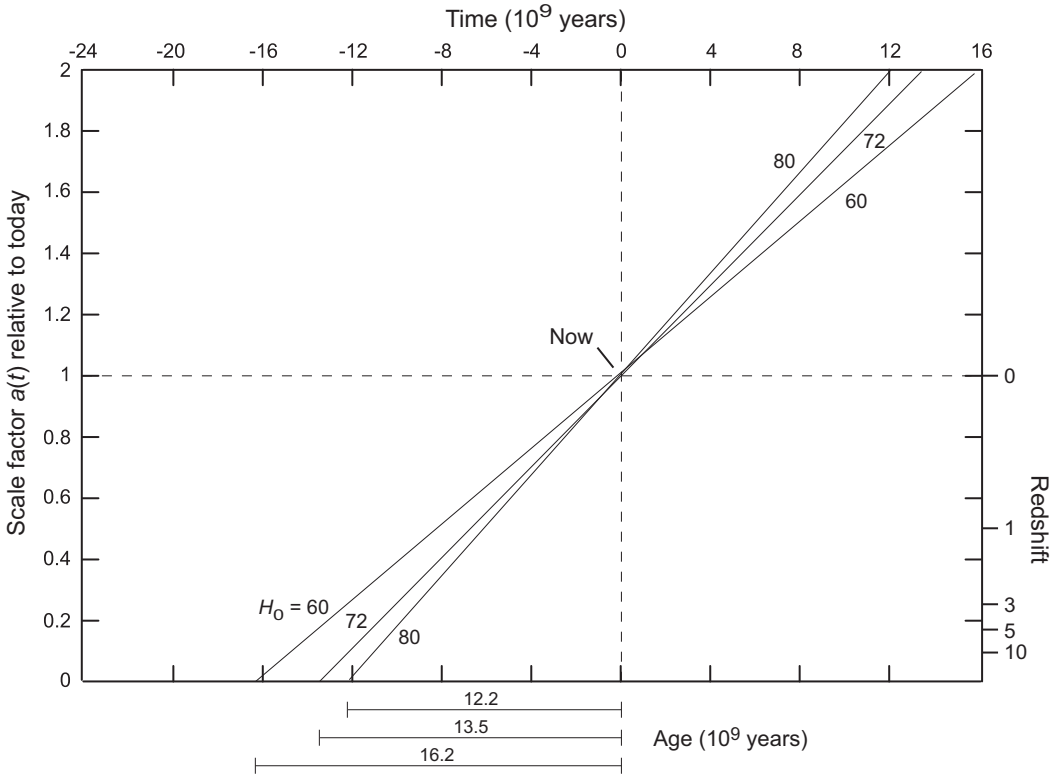


Figure 16.1: Expansion of the Universe for three values of the Hubble constant ($\mathrm{km\,s^{-1}\,Mpc^{-1}}$). Hubble times estimating the age of the Universe are indicated below the lower axis. Redshift is indicated on the right axis.

16.2 The Hubble Law

Hubble Law: $$v \equiv \frac{d\ell}{dt} = H_0 \ell \qquad H_0 \simeq 72\,\mathrm{km\,s^{-1}\,Mpc^{-1}},$$

The Hubble expansion is most consistently interpreted in terms of an *expansion of space itself*.

- Introduce a *scale factor* $a(t)$ that describes how distances scale because of the expansion of the Universe (Fig. 16.1).
- The *slopes of the lines* define the *Hubble constant* H_0.

We shall interpret the Hubble parameter as being

- characteristic of a space (thus constant for the Universe at a given time)
- but having possible time dependence

as the Universe evolves.

- The subscript zero on H_0 denotes that this is the value of the Hubble constant *today*,
- in anticipation that the coefficient governing the rate of expansion changes with time.
- One often refers to the
  - *Hubble parameter* $H = H(t)$, meaning an H that varies with time, and to the
  - *Hubble constant* H_0 to mean the *value of* $H(t)$ *today*.

Hubble's *original value* was

$$H_0 = 550\,\mathrm{km\ s^{-1}\ Mpc^{-1}}.$$

This is approximately *an order of magnitude larger* than the presently accepted value of

$$H_0 \sim 70\,\mathrm{km\ s^{-1}\ Mpc^{-1}}.$$

The large revision (which implies a corresponding shift in the perceived distance scale of the Universe) was because

- Hubble's original sample was a poorly-determined one based on relatively nearby galaxies.
- There was confusion over the extra-galactic distance scale at the time because of issues like
  - misinterpreting types of variable stars and
  - failing to account for the effect of dust on light propagation.

Table 16.1: Some peculiar velocities in the Virgo Cluster

Galaxy	Redshift (z)	v_r (km s^{-1})
IC 3258	−0.001454	−436
M86 (NGC 4406)	−0.000901	−270
NGC 4419	−0.000854	−256
M90 (NGC 4569)	−0.000720	−216
M98 (NGC 4192)	−0.000467	−140
NGC 4318	+0.004086	+1226
NGC 4388	+0.008426	+2528
IC 3453	+0.008526	+2558
NGC 4607	+0.007412	+2224
NGC 4168	+0.007689	+2307
M99 (NGC 4254)	+0.008036	+2411
NGC 4354	+0.007700	+2310

Source: SIMBAD

16.2.1 Redshifts

If a spectral line normally at wavelength λ_{emit} is shifted to a wavelength λ_{obs} when we observe it, the *redshift* z is

$$z \equiv \frac{\lambda_{\text{obs}} - \lambda_{\text{emit}}}{\lambda_{\text{emit}}}.$$

- A *negative value of* z corresponds to a *blueshift*.
- A *positive value of* z corresponds to a *redshift*.
- The *Hubble law gives rise only to redshifts*.
- Thus, any blueshifts correspond to *peculiar motion* of objects with respect to the general Hubble flow (Table 16.1).
- "Peculiar" = "property specific to an object".

The few galaxies observed to have blueshifts are *nearby*,

- in the Local Group or the Virgo Cluster,
- where peculiar motion is large enough to partially counteract the overall Hubble expansion.

The Andromeda Galaxy (M31), which is part of our Local Group of galaxies,

- is moving toward us with a velocity of about $300\ \mathrm{km\,s^{-1}}$ and
- will probably *collide with the Milky Way* in several billion years.

The *most extreme blueshifts* (negative radial velocities) found in the Virgo Cluster are the largest blueshifts known with respect to our galaxy.

16.2.2 Expansion Interpretation of Redshifts

The redshifts associated with the Hubble law may be *approximately* viewed as Doppler shifts for *small redshifts*.

- This interpretation is *problematic for large redshifts*.
- The Hubble redshifts (large and small)
  - are *most consistently interpreted* in terms of the *expansion of space*,
  - which may be parameterized by the *cosmic scale factor* $a(t)$.
- If all peculiar motion is ignored
  - the time dependence of the expansion is lodged entirely in the time dependence of $a(t)$, and
  - *all distances simply scale with this factor*.
- A simple analogy on a 2-dimensional surface will be exploited in later discussion:

  > Distances between dots placed on the surface of a balloon all *scale with the radius of the balloon* as it expands.

- In the general case $a(t)$ may be interpreted as *setting a scale for all cosmological distances*.

We shall show later that light traveling between galaxies at cosmological separation follows the null curve defined by

$$c^2dt^2 = a^2(t)dr^2$$

- The *scale factor* $a(t)$ sets the overall scale for distances in the Universe at time t and
- r is the *coordinate distance*. Thus,

$$\frac{cdt}{a(t)} = dr.$$

Consider a wavecrest of light with wavelength λ' that is

- *emitted* at time t' from one galaxy and
- *detected* with wavelength λ_0 at time t_0 in another galaxy.

Integrating both sides of the above equation gives

$$c\int_{t'}^{t_0} \frac{dt}{a(t)} = \int_0^r dr = r.$$

- The *next wavecrest* is *emitted* at $t = t' + \lambda'/c$ and is
- *detected* in the second galaxy at time $t = t_0 + \lambda_0/c$.

For the 2nd wave crest, integrating as above (neglect the interval between wavecrests compared with the expansion timescale)

$$c\int_{t'+\lambda'/c}^{t_0+\lambda_0/c} \frac{dt}{a(t)} = \int_0^r dr = r.$$

From the preceding equations we have

$$c\int_{t'}^{t_0}\frac{dt}{a(t)} = r \qquad c\int_{t'+\lambda'/c}^{t_0+\lambda_0/c}\frac{dt}{a(t)} = r.$$

Thus we may *equate the left sides* to obtain

$$\int_{t'}^{t_0}\frac{dt}{a(t)} = \int_{t'+\lambda'/c}^{t_0+\lambda_0/c}\frac{dt}{a(t)},$$

which may be rewritten as

$$\int_{t'}^{t'+\lambda'/c}\frac{dt}{a(t)} + \int_{t'+\lambda'/c}^{t_0}\frac{dt}{a(t)} = \int_{t'+\lambda'/c}^{t_0}\frac{dt}{a(t)} + \int_{t_0}^{t_0+\lambda_0/c}\frac{dt}{a(t)}$$

and upon *cancelling the terms in red*,

$$\int_{t'}^{t'+\lambda'/c}\frac{dt}{a(t)} = \int_{t_0}^{t_0+\lambda_0/c}\frac{dt}{a(t)}.$$

- The interval between wave crests is negligible compared with the expansion timescale, so
- we may bring $1/a(t)$ outside the integral to obtain

$$\frac{1}{a(t')}\int_{t'}^{t'+\lambda'/c} dt = \frac{1}{a(t_0)}\int_{t_0}^{t_0+\lambda_0/c} dt \quad\longrightarrow\quad \frac{\lambda'/c}{a(t')} = \frac{\lambda_0/c}{a(t_0)},$$

and thus that

$$\frac{\lambda'}{\lambda_0} = \frac{a(t')}{a(t_0)}.$$

The result that we have just obtained,

$$\frac{\lambda'}{\lambda_0} = \frac{a(t')}{a(t_0)},$$

demonstrates explicitly that

- The stretching of wavelengths (redshift) is caused by the *expansion of the Universe*
- (specifically by the change in $a(t)$ while the photon is propagating).
- The cosmological redshift is *not a Doppler shift* (*no velocities appear in the above formula*).

From the definition for the redshift z and the preceding result,

$$1 + z = 1 + \frac{\lambda_0 - \lambda'}{\lambda'} = \frac{\lambda_0}{\lambda'} = \frac{a(t_0)}{a(t')}$$

from which

$$z = \frac{a(t_0)}{a(t')} - 1.$$

It is conventional to

- *normalize the scale parameter* so that its value in the present Universe is unity, $a(t_0) \equiv 1$,
- in which case

$$z = \frac{1}{a(t')} - 1,$$

where $a(t')$ is the *scale factor of the Universe when the light was emitted.*

Thus the *redshift z that enters the Hubble law*

- Depends *only* on the ratio of the scale parameters at the *time of emission* and *time of detection* for the light,

$$1+z=\frac{a(t_0)}{a(t')}.$$

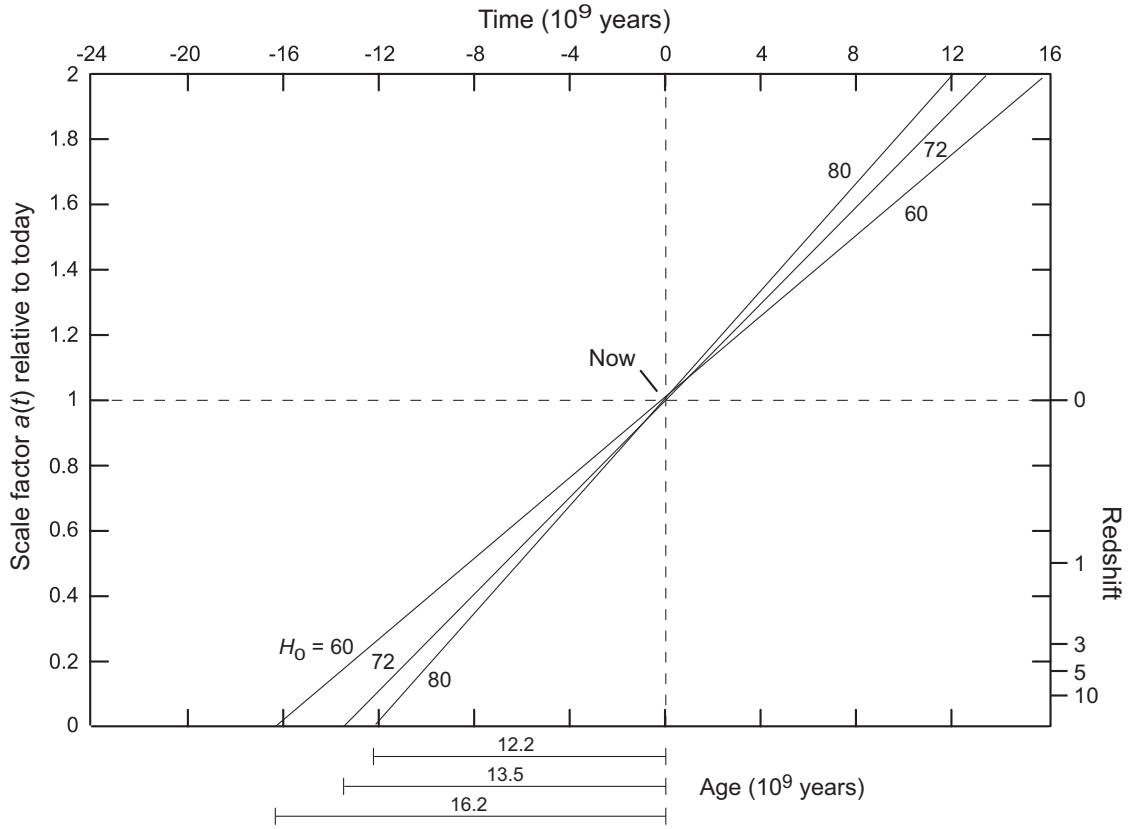
- It is *independent of the details* of how the scale parameter changed between the two times.
- Redshift is then determined completely by
  - the scale parameter of the expanding Universe *at the time the light was emitted*,
  - relative to the *scale parameter today*.
- The ratio of the scale parameter at two different times depends on the cosmological model assumed.
- Thus, measuring redshifts tests cosmological models.

EXAMPLE: If from the spectrum of a quasar

$$z\equiv\frac{\Delta\lambda}{\lambda}=5,$$

the scale factor of the Universe at the time that light was emitted from the quasar is given by

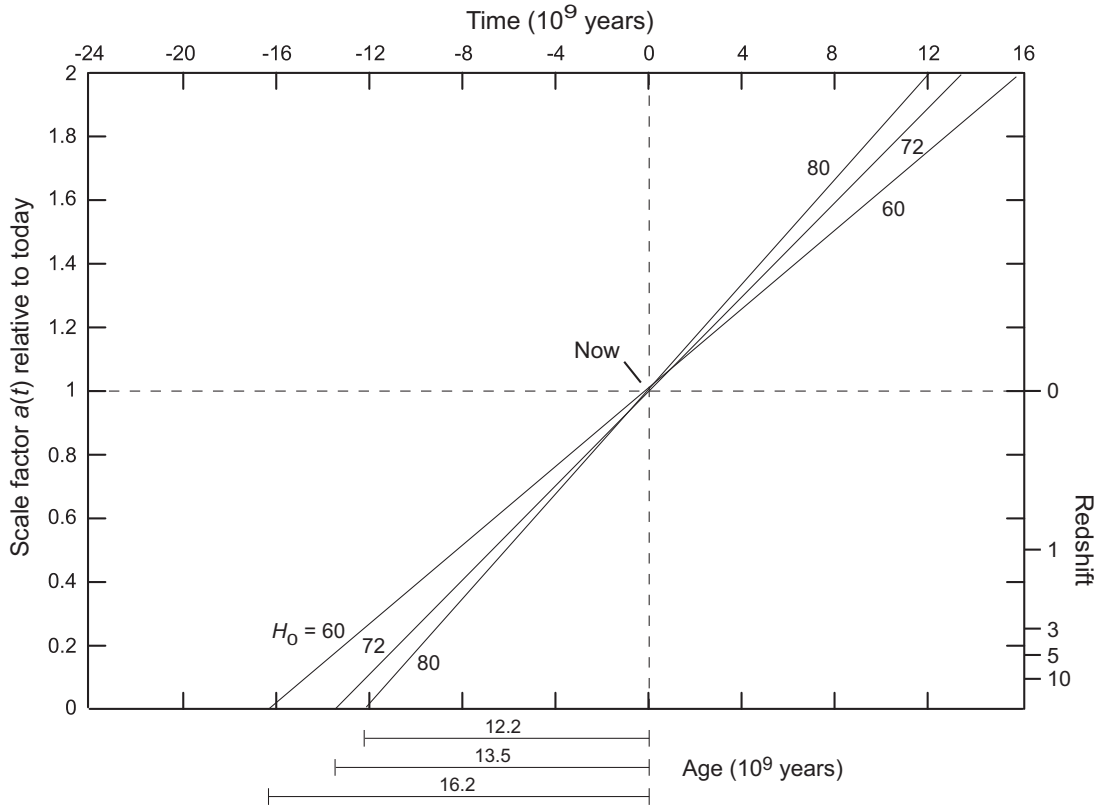
$$\frac{a(t')}{a(t_0)}=\frac{1}{z+1}=\frac{1}{6}.$$



The preceding discussion indicates that we may use

- the scale factor $a(t)$ or
- the redshift z

interchangeably as time variables for a universe in which the scale parameter changes monotonically (compare the right and left axes of the above figure).



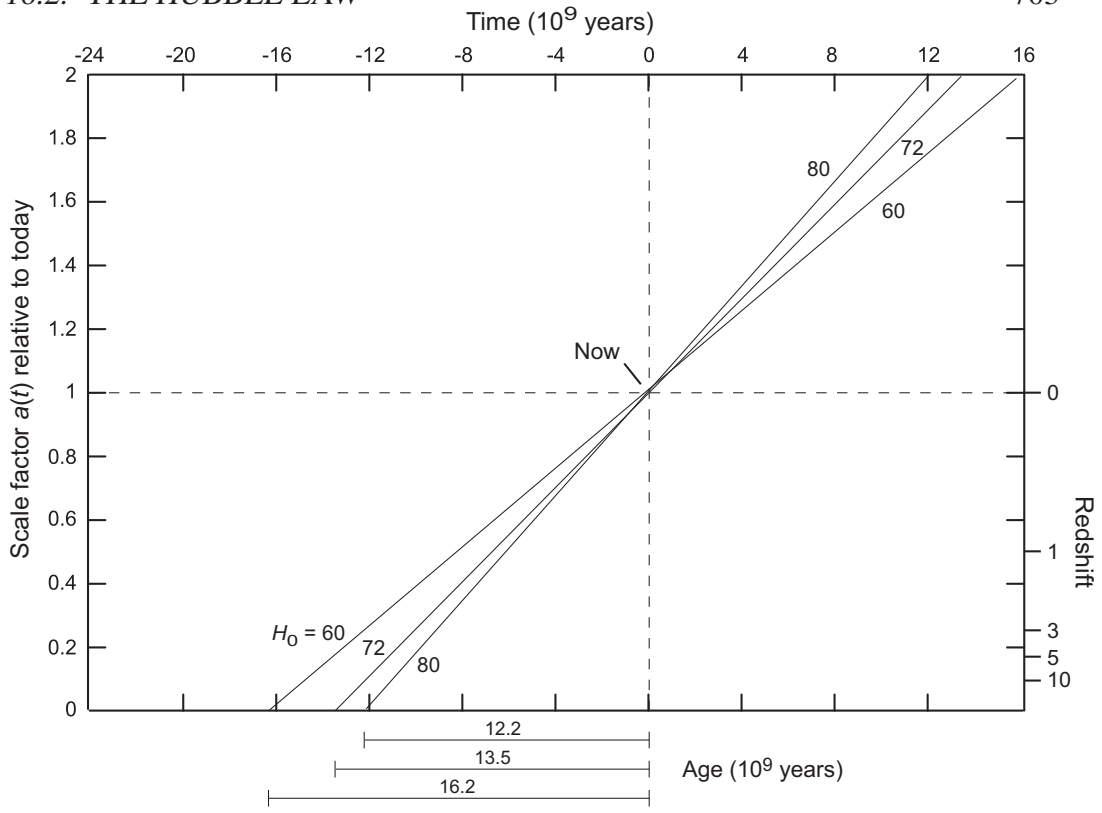
The *Hubble parameter* has units of *inverse time:*

$$[H_0] = [\mathrm{km\,s^{-1}\,Mpc^{-1}}] = \mathrm{time}^{-1}.$$

Thus, $1/H_0$ defines a time called the *Hubble time* τ_{H},

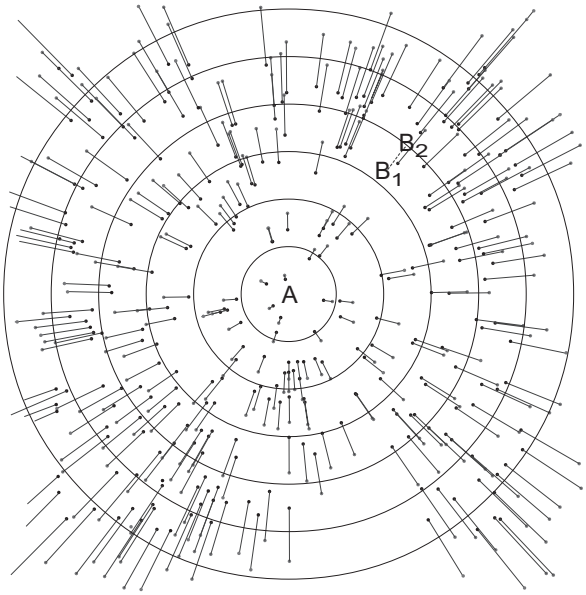
$$\tau_{\mathrm{H}} \equiv \frac{1}{H_0} = 9.8h^{-1} \times 10^9\,\mathrm{y}.$$

- If the Hubble law is obeyed with constant H_0, the intercept with the time axis gives the time when the scale factor was zero.
- Hence, the value of $\tau_{\mathrm{H}} = 1/H_0$ is sometimes quoted as an estimate of the *age of the Universe*.

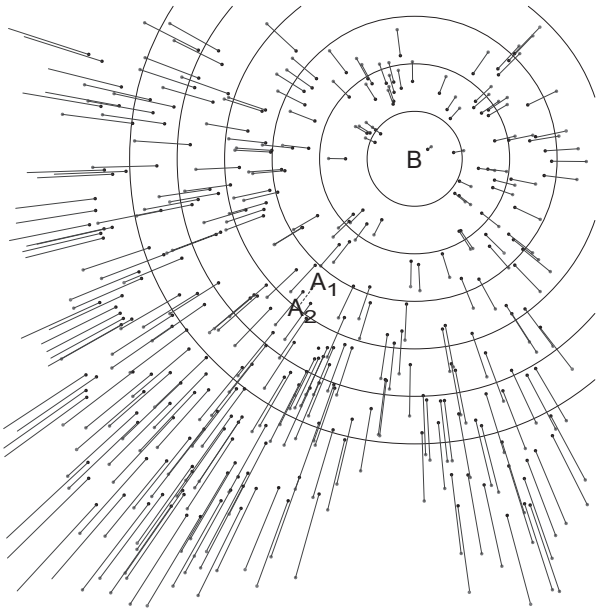


- The Hubble time is *not the age of the Universe* because
- the Hubble parameter can remain constant only in a Universe devoid of matter, fields, and energy.
- The realistic Universe contains all of these and
- expansion is *accelerated* by gravitational interactions.

In later cosmological models we shall see that the age of the Universe may be *substantially longer or shorter than* τ_H, depending on details of the matter, fields, and energy contained in the Universe.



(a) As seen from galaxy A, galaxy B appears to recede from B_1 to B_2 over the time interval shown.



(b) As seen from galaxy B, galaxy A appears to recede from A_1 to A_2 over the time interval shown.

The figures above show the *same uniform two-dimensional Hubble expansion,* as seen from *two different vantage points*.

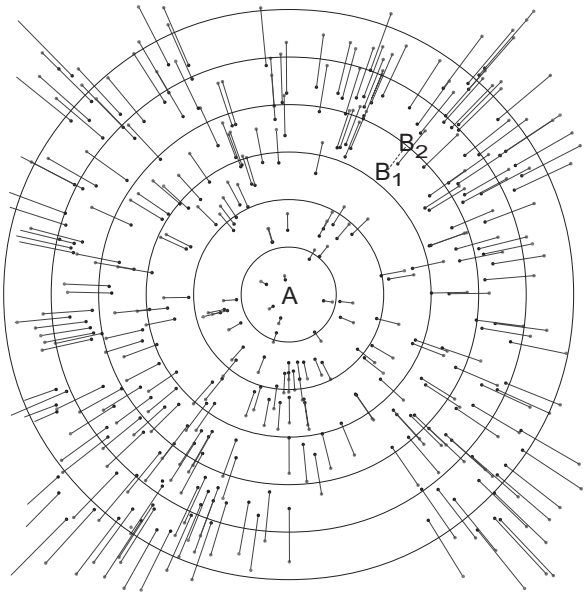
(a) As *observed from galaxy A*.

(b) As *observed from galaxy B*.

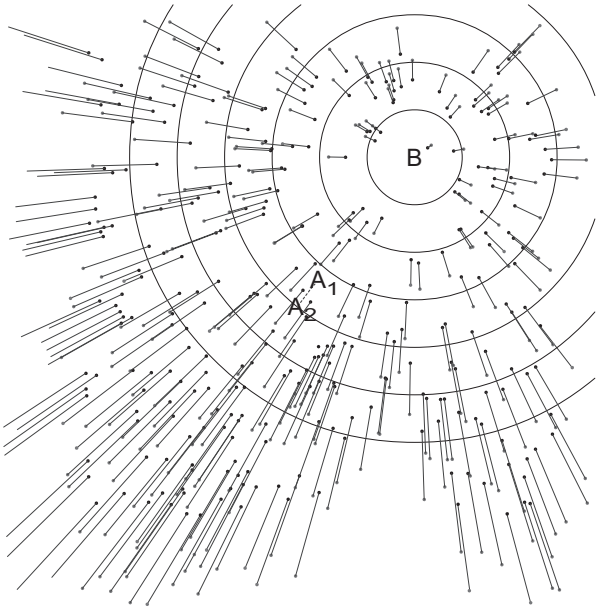
The patterns in (a) and (b) *look very different*, but

- this is *exactly the same expansion*
- seen from two *different perspectives*.

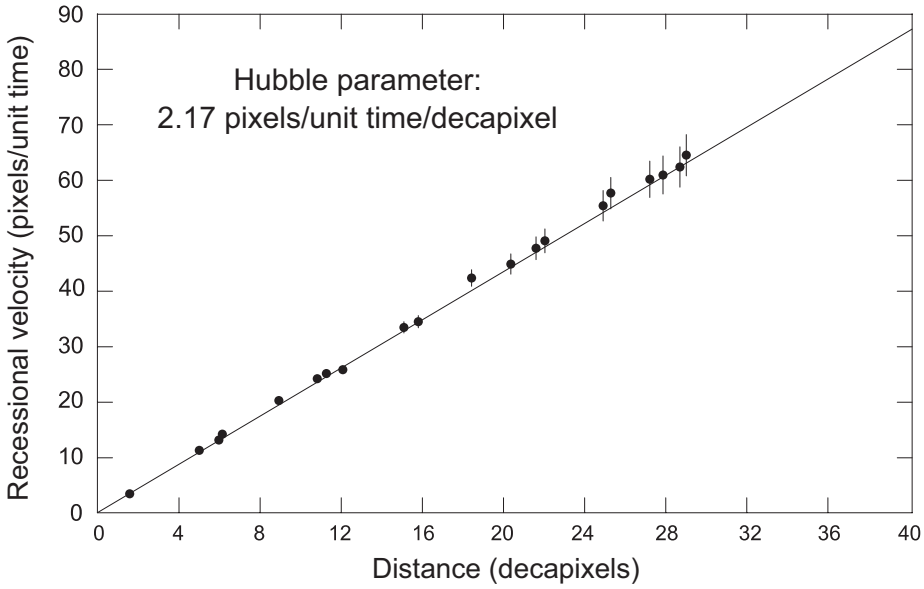
A and B will find *the same Hubble constant*, and each thinks they are the *center of the expansion*.



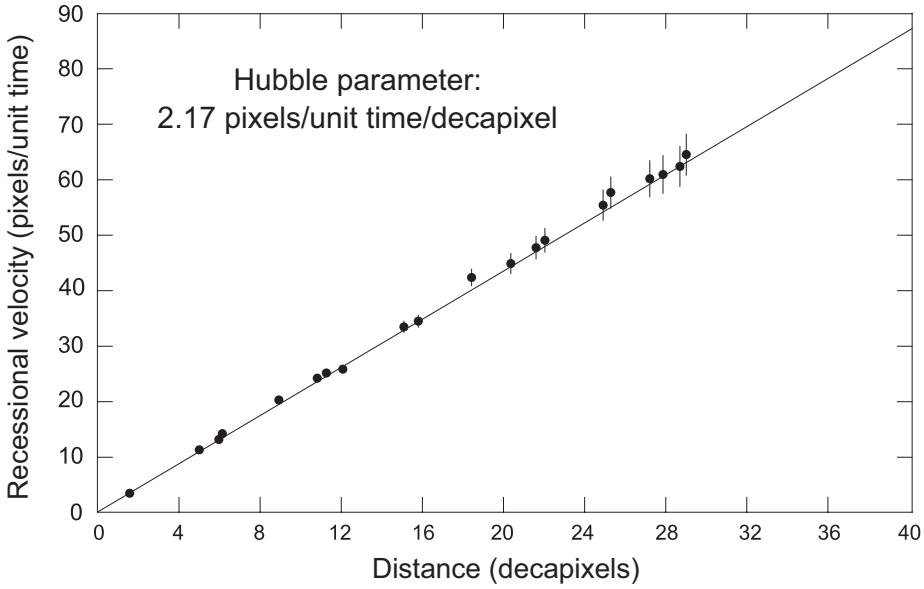
(a) As seen from galaxy A, galaxy B appears to recede from B_1 to B_2 over the time interval shown.



(b) As seen from galaxy B, galaxy A appears to recede from A_1 to A_2 over the time interval shown.



The figures above illustrate *extracting the Hubble law* from the simulated 2D expansion.



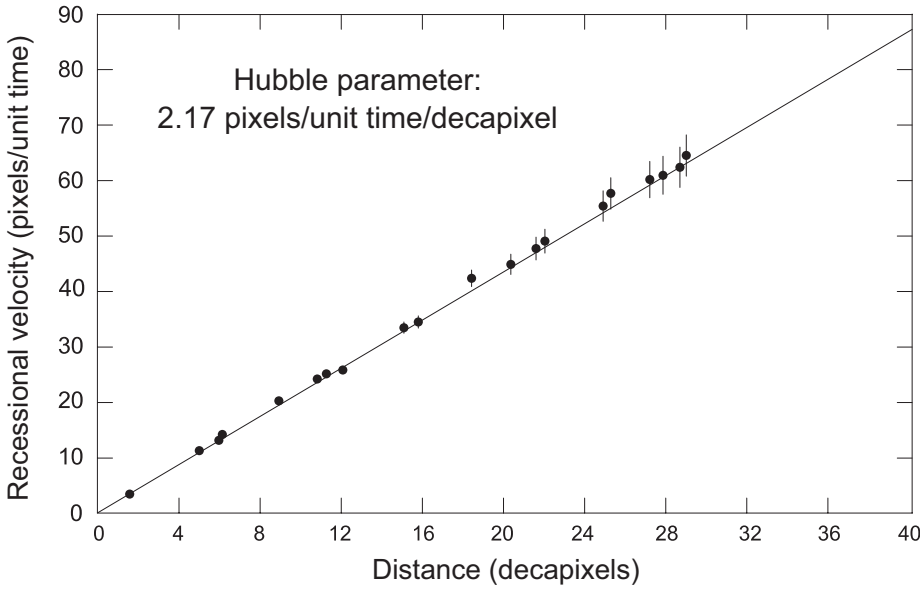
The *Hubble plot* is constructed by

- choosing representative galaxies and
- plotting the apparent recessional velocity versus the distance from the observer,
- with the Hubble constant extracted from the *slope of a linear fit* to the data.

Within a simulated observational uncertainty,

- *Observer A* and *Observer B*
- extract the *same Hubble parameter,*

by measuring recessional velocities and distances *relative to each observer's position*.



In this model

- distances are measured in *pixels* on the computer screen,
- time is measured in units of the *elapsed time for expansion*, and
- velocities are estimated in units of *pixels per unit time* by
- counting the *number of pixels the galaxy moves* in the expansion time.

Thus, the *unit for* H_0 in this simulation of

$$\text{pixels} \times \text{time}^{-1} \times \text{decapixel}^{-1}$$

(where 1 decapixel = 10 pixels) mimics the standard unit of

$$\text{km}\,\text{s}^{-1}\,\text{Mpc}^{-1}$$

used for the *actual Hubble constant*.

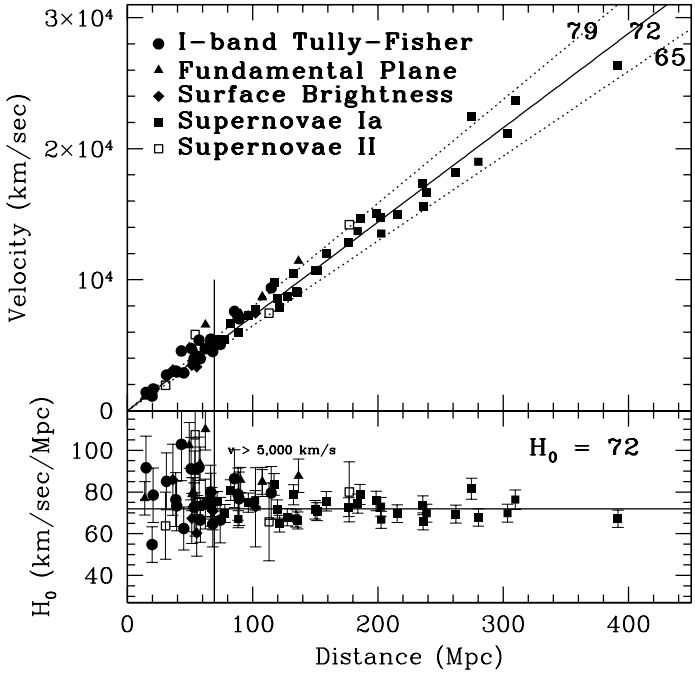


Figure 16.2: Hubble parameter extracted from observations Three possible slopes (Hubble parameter) are indicated in the upper part of the figure.

16.2.3 Measuring the Actual Hubble Constant

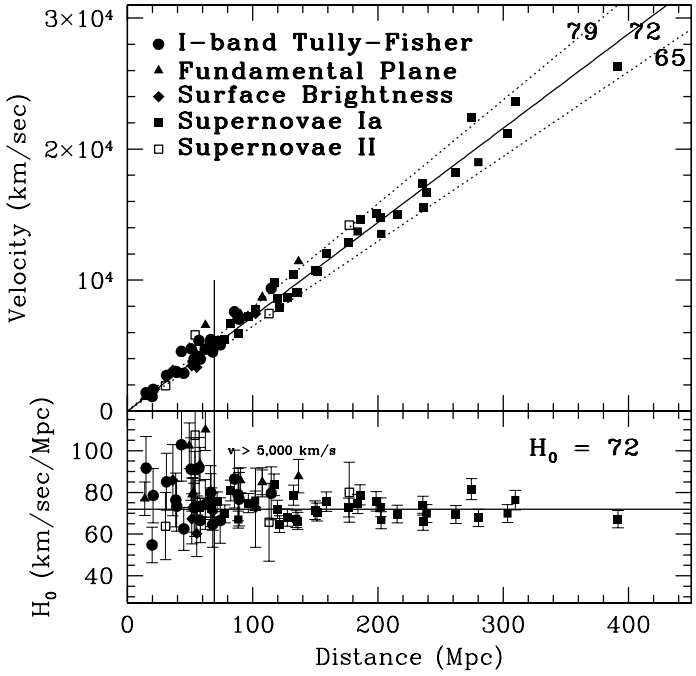
The Hubble constant may be *determined observationally* by

- *measuring the redshift* for spectral lines and
- comparing that with the *distance to objects*,

at large enough distances so *peculiar motion is negligible*. Fig. 16.2 illustrates for various observations, with an adopted value

$$H_0 = 72 \pm 8\,\mathrm{km\,s^{-1}\,Mpc^{-1}},$$

corresponding to $h = 0.72$.



The top part of the above figure shows *velocity vs. distance* for

- five different secondary distance indicators, calibrated by *Cepheid variable observations* out to $\sim 20\,\text{Mpc}$.
- The lower part of the diagram shows the inferred value of H_0 as a function of distance, with
- $H_0 = 72\,\text{km s}^{-1}\,\text{Mpc}^{-1}$ indicated by the horizontal line.
- The *large scatter of points* below $\sim 100\,\text{Mpc}$ is because of *peculiar velocities* due to local gravitational attraction.
- *Type Ia supernovae* provide data at *the largest distances*.

Later, use of *Type Ia supernovae* to probe expansion at even *higher redshifts* will be discussed.

Modern determinations find that

$$H_0 \sim 67 - 73\,\mathrm{km\,s^{-1}\,Mpc^{-1}},$$

with generally

- *"distance ladder" approaches* as just illustrated tending to give values nearer the higher end of this range and
- *cosmic microwave background data* to be discussed later tending to give values closer to the lower end.

Gravitational waves can determine H_0 (see later), but present data give $\pm 10\%$ errors and can't yet add meaningful constraints to the value of H_0.

16.3 Limitations of the Standard World Picture

The standard picture has been *remarkably successful* in describing many features of our Universe. However, two aspects of this picture suggest that *it is (at best) incomplete:*

1. In order to get the big bang to produce the present universe, certain *assumptions about initial conditions* must be taken as given. While not necessarily wrong, some of these assumptions seem *unnatural by various standards*.

2. As the expansion is extrapolated backwards, eventually we would reach a state of sufficient temperature and density that a fully *quantum theory of gravitation would be required*.

   - This is the *Planck era*, and the corresponding scales of distance, energy, and time are called the *Planck scale*.
   - Since we do not yet have a consistent theory of quantum gravity, the presently understood laws of physics may be expected to *break down on the Planck scale*.
   - Thus the standard picture says nothing about the Universe at those very early times.

In later chapters we shall address these issues, to consider whether modifications of the standard picture can alleviate some of these problems.

Chapter 17

Energy and Matter in the Universe

The history and fate of the Universe ultimately turn on how much matter, energy, and pressure it contains:

1. These components of the stress–energy tensor all couple to gravity.
2. This coupling determines how self-gravitation of the Universe influences the Hubble expansion.

In this chapter we begin to address quantitatively the issue of the matter and energy contained in the Universe and how that determines its history.

The Universe is *mostly empty space*.

- This might suggest that *Newtonian gravity* (valid in the weak gravity limit) is adequate for describing its large-scale structure.
- But whether *general relativity* effects are important may be estimated in terms of
  - the ratio of an *actual radius* for a massive object
  - compared with its *radius of gravitational curvature*.
- If we *apply such a criterion to the entire Universe*, reasonable estimates for the mass–energy of the Universe indicate that
  - the *actual radius* of the known Universe and
  - the corresponding *gravitational curvature radius*

  are *likely to be comparable*.
- Thus, a description of the large-scale structure of the Universe (cosmology) *must be built on a covariant gravitational theory*, rather than on Newtonian gravity.

Even so, we can *understand a substantial amount* concerning the expanding Universe simply by *using Newtonian concepts*.

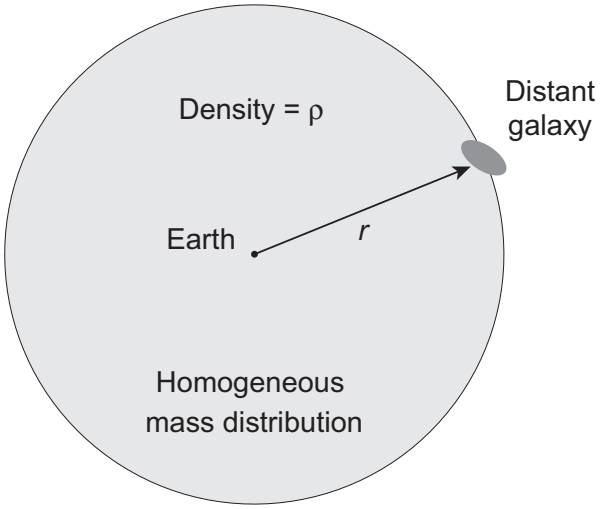


Figure 17.1: Newtonian model of the expanding Universe.

17.1 Expansion and Newtonian Gravity

The *gravitational potential* acting on the galaxy in Fig. 17.1 is

$$U = \frac{-GMm}{r},$$

where m is the *mass of the galaxy* and

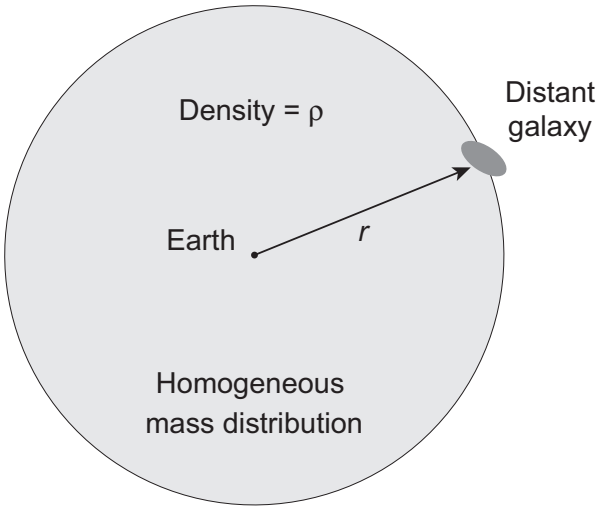
$$\text{Total mass of sphere} = M = \tfrac{4}{3}\pi r^3 \rho,$$

which is *constant* since

- ρ decreases with time and
- r increases with time,

but the product ρr^3 is constant. Thus

$$U = -\tfrac{4}{3}\pi G r^2 \rho m.$$



If the *motion of the galaxy* is caused *entirely by the Hubble expansion*,

- its *radial velocity* relative to the Earth is $v = H_0 r$.
- This implies a *kinetic energy*

$$T = \tfrac{1}{2}mv^2 = \tfrac{1}{2}mH_0^2 r^2,$$

  where m is the inertial mass of the galaxy, assumed equivalent to its gravitational mass.
- The *total energy* of the galaxy is then

$$\begin{aligned} E = T + U &= \tfrac{1}{2}mH_0^2 r^2 - \tfrac{4}{3}\pi G r^2 \rho m \\ &= \tfrac{1}{2}mr^2\left(H_0^2 - \tfrac{8}{3}\pi G\rho\right), \end{aligned}$$

in this *Newtonian approximation*.

17.2 The Critical Density

If the *expansion is to halt*, we must have $E = 0$ and thus

$$E = \tfrac{1}{2}mr^2\left(H_0^2 - \tfrac{8}{3}\pi G\rho\right) \quad \longrightarrow \quad H_0^2 = \tfrac{8}{3}\pi G\rho .$$

Solving for ρ, the *critical density* just halting the expansion is

$$\rho_c = \frac{3H_0^2}{8\pi G} \simeq 1.88 \times 10^{-29} h^2 \,\mathrm{g\,cm^{-3}}.$$

The corresponding *critical energy density* is

$$\begin{aligned} \varepsilon_c = \rho_c c^2 &= 1.05 \times 10^{-2} h^2 \ \mathrm{MeV\,cm^{-3}} \\ &= 1.69 \times 10^{-8} h^2 \,\mathrm{erg\,cm^{-3}}. \end{aligned}$$

> The critical density ρ_c corresponds to an average of only *six hydrogen atoms per cubic meter*, or about *140 $M_\odot$ per cubic kiloparsec*.

We may distinguish *three qualitative regimes* for the *actual density* ρ (in this simple Newtonian picture):

1. If $\rho > \rho_c$ the Universe is *closed* and the *expansion will stop in a finite time*.
2. If $\rho < \rho_c$ the Universe is *open* and the *expansion will never halt*.
3. If $\rho = \rho_c$ the Universe is *flat* (or *euclidean*) and the *expansion will halt*, but only *asymptotically as* $t \to \infty$.

Thus, in this simple Newtonian picture

- the *ultimate fate of the Universe*
- is determined by its *present matter density*.

We shall see that this conclusion is modified—profoundly—by the presence of *dark energy* in the actual Universe.

It will prove convenient to introduce the dimensionless *total density parameter* evaluated at the present time

$$\Omega \equiv \frac{\rho}{\rho_c} = \frac{\varepsilon}{\varepsilon_c} = \frac{8\pi G\rho}{3H_0^2}.$$

where ρ is the current total density that couples to gravity.

- Thus the *closure condition* implies that $\Omega = 1$.
- Note for future reference that many authors use a subscript "0" on Ω and ρ (and other cosmological parameters) to indicate explicitly that *they are evaluated at the present time*.
- Where possible we *suppress these zero subscripts* to avoid notational clutter.

> Unless otherwise noted, you should *always understand* cosmological parameters like Ω to be *evaluated at the present time*, even if we suppress the subscript zero.

17.3 Cosmic Scale Factor

The *Hubble expansion* makes it convenient to

- introduce a *cosmic scale factor* $a(t)$ that *sets the global distance scale* of the Universe.
- If peculiar motion is ignored, the expansion is governed entirely by $a(t)$ and *all distances scale with this factor*.
- *Example:* If
  - the present time is t_0 and
  - the present scale factor is a_0,
  - a wavelength of light λ emitted at time $t < t_0$
  - is scaled to λ_0 at $t = t_0$ by the universal expansion:

$$\frac{\lambda_0}{a_0} = \frac{\lambda}{a(t)}.$$

- Likewise, if r_0 and ρ_0 are the present values of r and ρ,

$$\frac{r(t)}{r_0} = \frac{a(t)}{a_0} \qquad \frac{\rho(t)}{\rho_0} = \left(\frac{a_0}{a(t)}\right)^3,$$

  and so on.

The universal scaling of distances with $a(t)$ permits us to express *all dynamical equations in terms of the scale factor.*

Example:

- The gravitational force acting on a galaxy in Newtonian approximation is

$$F_{\rm G} = -\frac{\partial U}{\partial r} = -G\frac{Mm}{r^2} = -\frac{4}{3}\pi G\rho rm,$$

and the corresponding gravitational acceleration is

$$\ddot{r} = \frac{F_{\rm G}}{m} = -\frac{GM}{r^2} = -\frac{4}{3}\pi G\rho r.$$

- Then from

$$\frac{r(t)}{r_0} = \frac{a(t)}{a_0},$$

we have for the *acceleration of the scale factor,*

$$\ddot{r} = \frac{r_0}{a_0}\ddot{a} = -\frac{4}{3}\pi G\rho_0\frac{a_0^3}{a^3}\frac{r_0}{a_0}a \quad \rightarrow \quad \ddot{a} = -\frac{4}{3}\pi G\rho_0 a_0^3\left(\frac{1}{a^2}\right).$$

Dropping the subscript on ρ_0 and utilizing

$$\Omega \equiv \frac{8\pi G\rho}{3H_0^2},$$

the acceleration of the scale factor also may be expressed in terms of the *density parameter* Ω,

$$\ddot{a} = -\frac{H_0^2 a_0^3\Omega}{2a^2}.$$

From the result

$$\ddot{a} = -\frac{H_0^2 a_0^3 \Omega}{2a^2}.$$

we may show that (Problem)

$$\dot{a}^2 = a_0^2 H_0^2 f(\Omega, t),$$

where we define

$$f(\Omega, t) = 1 + \Omega \frac{a_0}{a(t)} - \Omega,$$

which must obey the condition

$$f(\Omega, t) \geq 0,$$

since $\dot{a}^2$ can never be negative. *NOTE:* $\Omega \equiv \Omega_0$ in these equations.

We may use this condition to enumerate *different possible histories* for the Universe.

17.4 Possible Expansion Histories

Consider as an example *dust-filled universes*, which contain

- only *pressureless, non-relativistic matter* and
- *negligible radiation* and *negligible vacuum energy*.

There are three qualitatively different scenarios, depending on the value of $\Omega \equiv \Omega_0$.

Possible expansion histories:

1. $\Omega < 1$ *(undercritical):* In this case, as $a(t) \to \infty$,

$$f(\Omega,t) = 1 + \Omega\frac{a_0}{a(t)} - \Omega \quad \longrightarrow \quad 1 - \Omega > 0,$$

so $\dot{a}$ never vanishes, implying an *open, ever-expanding universe*.

2. $\Omega = 1$ *(critical):* For this case, as $a(t) \to \infty$,

$$f(\Omega,t) \longrightarrow 0,$$

but it only reaches 0 at $t = \infty$. Hence,

- the universe is ever-expanding *(constraint: it is expanding now)*
- but the rate of expansion approaches zero asymptotically as $t \to \infty$.

3. $\Omega > 1$ *(overcritical):* Now as t increases

$$f(\Omega,t) \longrightarrow 0,$$

but in a *finite* time $t_{\max}$.

- Beyond this time $f(\Omega,t) \geq 0$ must still be satisfied.
- Thus, if $\Omega > 1$ the expansion turns into a *contraction* at time $t_{\max}$,

and the *universe begins to shrink*.

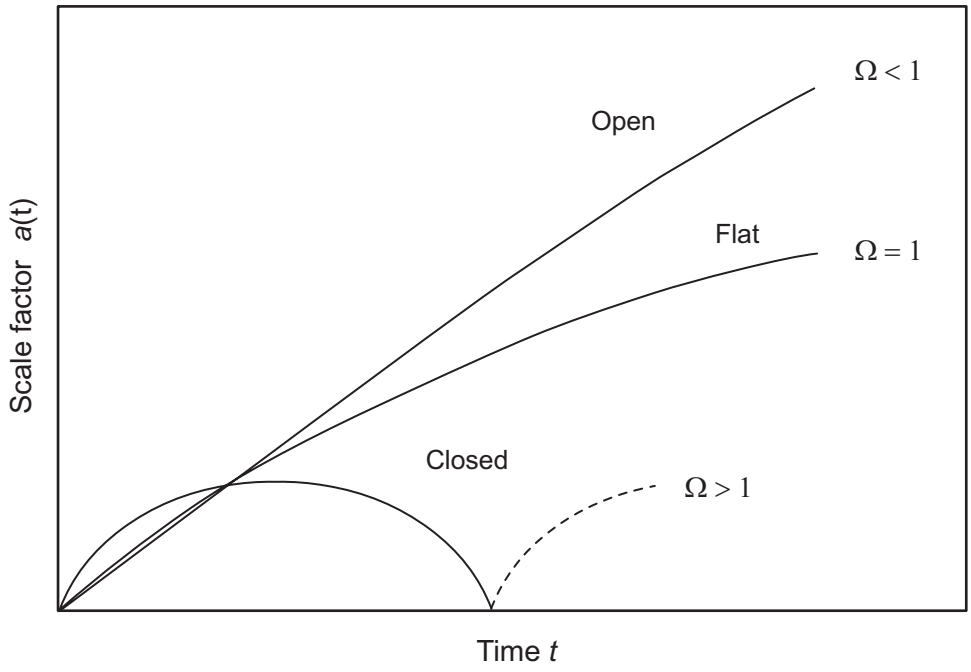


Figure 17.2: Behavior of the scale factor $a(t)$ as a function of time for a dust-filled universe.

By integrating the scale factor equation

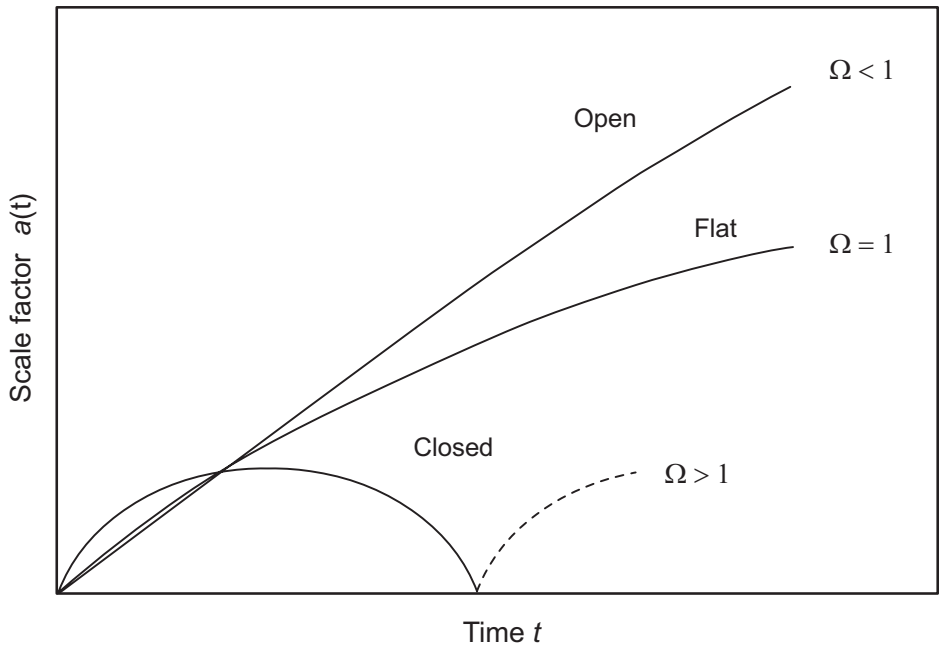
$$\dot{a}^2 = a_0^2 H_0^2 f(\Omega, t),$$

for a dust model (see Problems) we obtain for these three scenarios:

- For a *flat dust universe* with $\Omega = 1$,

$$a(t) = \left(\frac{3t}{2t_{\rm H}}\right)^{2/3}.$$

This behavior is sketched as the $\Omega = 1$ curve in Fig. 17.2.



- For a *closed dust universe* with $\Omega > 1$,

$$a(\psi) = \frac{1}{2}\frac{\Omega}{\Omega - 1}(1 - \cos\psi)$$
$$t(\psi) = \frac{1}{2H_0}\frac{\Omega}{(\Omega - 1)^{3/2}}(\psi - \sin\psi).$$

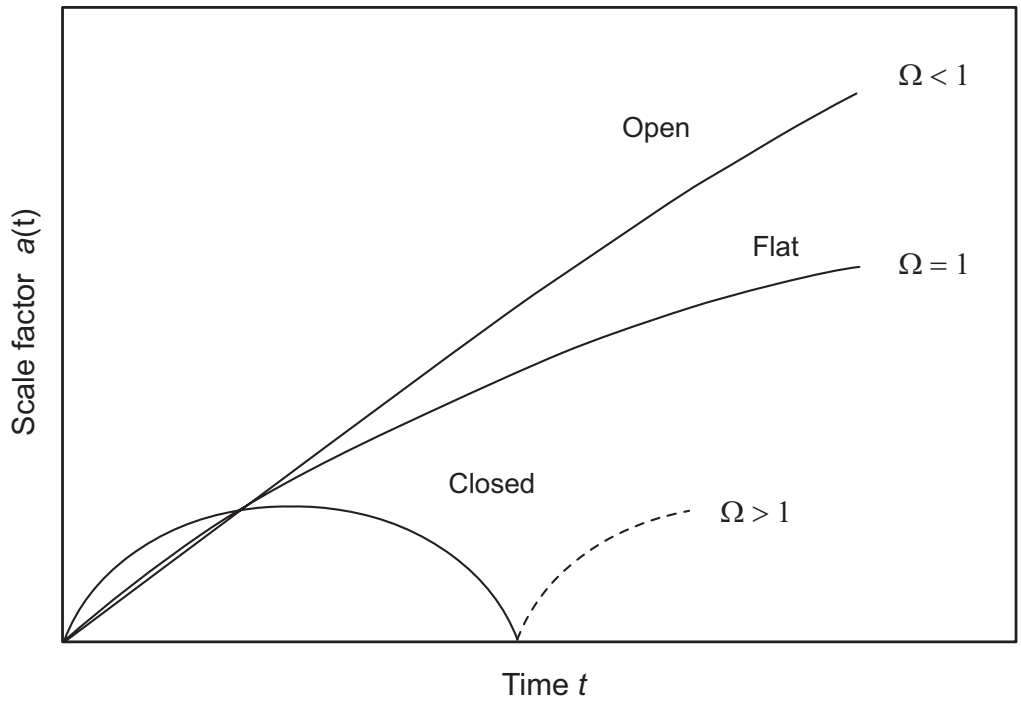
where $\psi \geq 0$ parameterizes the solution. The first maximum for $a(\psi)$ occurs for $\psi = \pi$,

$$\frac{a_{\max}}{a_0} = \frac{\Omega}{2(\Omega - 1)}(1 - \cos\pi) = \frac{\Omega}{\Omega - 1},$$

and the time at which this occurs is

$$t_{\max} = t(\pi) = \frac{\pi\Omega}{2H_0(\Omega - 1)^{3/2}}.$$

This case is sketched as $\Omega > 1$ in the figure above.



- For an *open dust universe* with $\Omega < 1$,

$$a(\psi) = \frac{\Omega}{2(1-\Omega)}(\cosh\psi - 1)$$
$$t(\psi) = \frac{1}{2H_0}\frac{\Omega}{(1-\Omega)^{3/2}}(\sinh\psi - \psi).$$

  where ψ is a parameter. This behavior is sketched as the $\Omega < 1$ curve in the figure above.

17.5 Lookback Times

Telescopes are time machines:

- *Lookback time:* $t_{\rm L}$ how far back in time we are looking when we view an object having a redshift z,

$$t_{\rm L} = t(0) - t(z),$$

where $t(z=0)$ is the present age of the Universe and $t(z)$ is the age when light observed today with redshift z was emitted.

- *Example:* in a flat dust-filled universe (Problem)

$$\frac{t(z)}{\tau_{\rm H}} = \frac{2}{3}(1+z)^{-3/2} \qquad \frac{t(0)}{\tau_{\rm H}} = \frac{2}{3}$$

and the lookback time is

$$\begin{aligned}\frac{t_{\rm L}}{\tau_{\rm H}} &= \frac{2}{3} - \frac{2}{3}(1+z)^{-3/2} \\ &= \frac{2}{3}\left(1 - \frac{1}{(1+z)^{3/2}}\right),\end{aligned}$$

where $\tau_{\rm H} = 1/H_0$ is the Hubble time.

- Thus light from an object that we observe with a redshift $z \sim 6$ was emitted when

1. The Universe was only $\sim 5\%$ *of its present age*.
2. The cosmic scale factor $a(t)$ was $1+z=7$ *times smaller* than it is today.

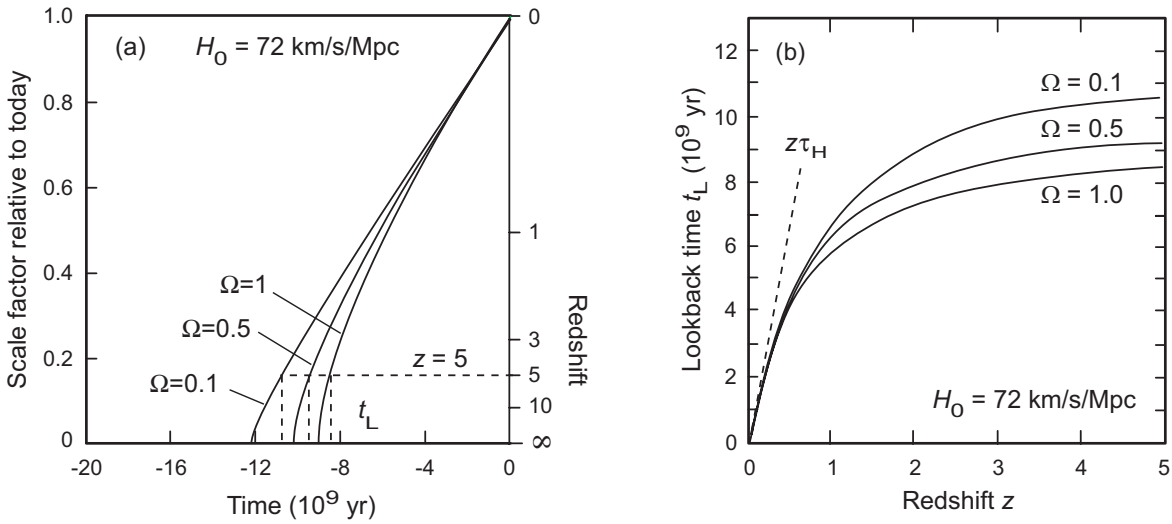


Figure 17.3: (a) Geometrical interpretation of the lookback time t_L for $z = 5$ in a dust universe with three different values of the density parameter Ω. (b) Lookback time as a function of redshift for different assumed density parameters in a dust model. The dashed line gives the result for Hubble's law.

Lookback time as a function of redshift is interpreted graphically for a dust model in Fig. 17.3(a), and is plotted for various values of the density parameter Ω in Fig. 17.3(b).

- For small redshift, $t_L \simeq z\tau_H$, as would be expected from the Hubble law.
- But for larger redshifts t_L differs substantially from this approximation.

17.6 The Inadequacy of Dust Models

The preceding discussion has applied *Newtonian gravity* to a universe containing only *pressureless matter (dust)*.

- Until the last decade of the 20th century, a covariant version of such theories was *the favored cosmology*.
- For example, one common model was the *Einstein–de Sitter universe,* a *covariant version of the* $\Omega = 1$ *solution*
  - with exactly *a closure density of dust* and
  - *zero curvature*

  that will be described later.
- But *observations of that time* indicated that there was
  - not nearly enough visible matter for closure,
  - suggesting an open-Universe cosmology with $\Omega < 1$.

We now know that the actual Universe contains *additional components* that influence its evolution in a highly-nontrivial way.

- As a result *a Newtonian dust cosmology is a poor approximation* to the actual history of the Universe.
- To understand the *new cosmology,* we first take inventory of these additional components.

> Our starting point will be evidence that most matter is *not the visible matter of stars and galaxies*.

17.7 Evidence for Dark Matter

There is strong observational evidence for large amounts of *dark matter* in the Universe

- that *reveals its presence gravitationally*, but
- is *not seen by any other probe*.

Let's review some of this evidence.

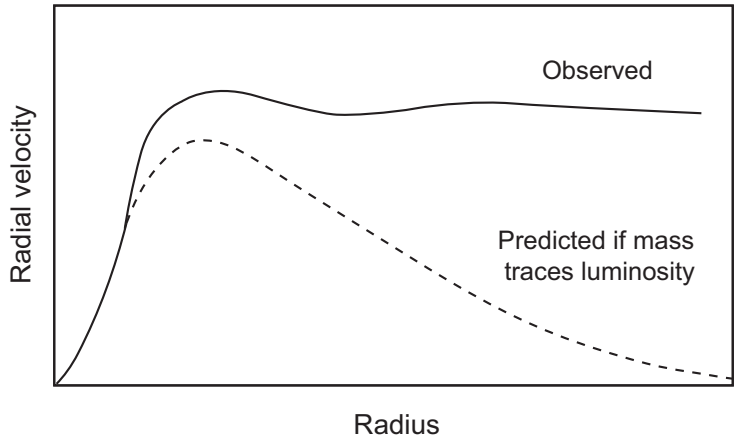


Figure 17.4: Schematic velocity curves for spiral galaxies.

17.7.1 Rotation Curves for Spiral Galaxies

In *spiral galaxies*, if we balance the centrifugal and gravitational forces at a radius R, the *tangential velocity* v should obey

$$v = \sqrt{\frac{GM}{R}},$$

where M is the enclosed mass.

- Well outside the matter distribution, we expect $v \simeq R^{-1/2}$.
- Velocities can be measured using the Doppler effect, both
  - for *visible light* from the luminous matter, and for
  - the *21 cm hydrogen line* for non-luminous hydrogen.
- For many spirals we find not $v \simeq R^{-1/2}$ but *almost constant velocity* well outside the bulk of the luminous matter.

This is illustrated schematically in Fig. 17.4.

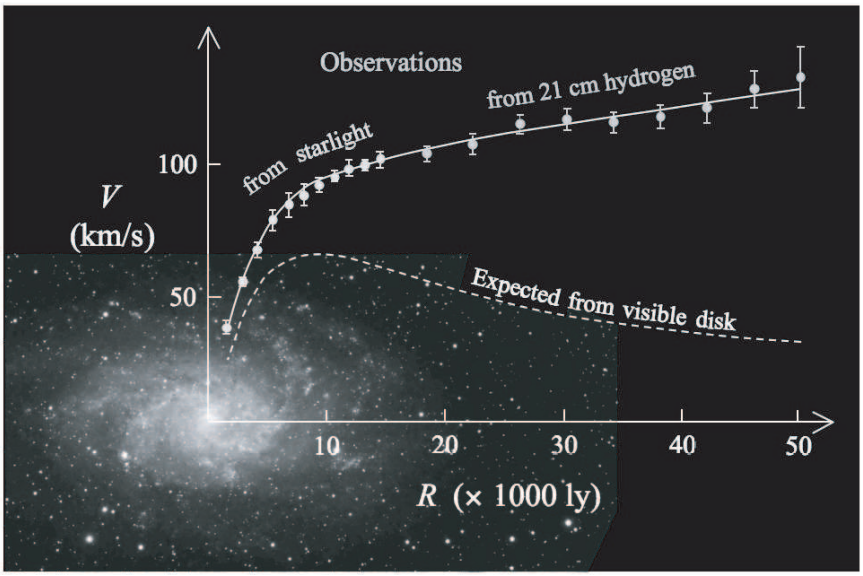


Figure 17.5: Rotation curve for the galaxy M33 out to a distance of about 15 kpc (50,000 ly) from the center. Points inside about 15,000 ly are from visible starlight; points beyond that are from radio frequency (RF) observations.

An example of a *measured rotational curve* is shown in Fig. 17.5 for the galaxy M33.

This rotation curve indicates

- the presence of *substantial gravitating matter*
- distributed in a *halo that extends beyond the visible matter*.

17.7.2 The Mass of Galaxy Clusters

From the *virial theorem*,

- the *mass within a region* can be estimated from
- the *velocity dispersion of objects bound gravitationally* in that region.
- In particular, we may estimate the mass contained within large clusters of galaxies by *studying the motion of galaxies* within the cluster.

Such estimates indicate that clusters of galaxies contain much more mass than their luminosity would suggest.

Example: The *Coma Cluster* contains thousands of galaxies.

- The *measured velocity dispersion* in a region lying within about 3 Mpc of the center is

$$\sigma_{\rm r} \simeq 900\,{\rm km\,s^{-1}}.$$

- Inserting this into the virial relation

$$\sigma_{\rm r}^2 \simeq \frac{GM}{5R}.$$

leads to the conclusion that the Coma Cluster has a mass of

$$M \sim 2.8 \times 10^{15} M_\odot.$$

- On the other hand, the *visual luminosity* of the Coma Cluster is

$$L = 5 \times 10^{12} L_\odot,$$

where $L_\odot$ is the luminosity of the Sun.

- Thus, the *ratio of mass to light* for the Coma Cluster is

$$\frac{\text{Mass}}{\text{Light}} \equiv \frac{M}{L} = \frac{2.8 \times 10^{15} M_\odot}{5 \times 10^{12} L_\odot} \simeq 565 \frac{M_\odot}{L_\odot}.$$

This suggests the presence of *large amounts of unseen mass* within this cluster of galaxies.

The preceding example is representative and in rich clusters (those containing thousands of galaxies)

- the *measured velocity dispersions* typically lie in the range $\sigma_r = 800$–$1000\ \mathrm{km\,s^{-1}}$ and
- the *mass to luminosity ratio* in solar units is typically found to be *several hundred*.
- This should be compared with a value of order one found in the Solar neighborhood.

We conclude that clusters of galaxies contain dark matter halos above and beyond the halos of the individual galaxies comprising the cluster.

17.7.3 Hot Gas in Clusters of Galaxies

Observations by X-ray satellites indicate that

- clusters of galaxies often are *immersed in a hot gas* filling the space between the galaxies.
- The *X-ray intensity* is a measure of the *strength of the gravitational field* because to produce X-rays requires large velocities of the gas particles.

These observations allow the total amount of gravitating matter in the cluster to be estimated.

- It is found that there must be *much more matter than just the luminous matter* to account for the hot gas observed systematically in the clusters.
- Otherwise, the gravitational field would be too weak to trap the gas for extended periods.

Example: X-ray observations of the *Coma Cluster* by ROSAT

- indicated a gas temperature of about 10^8 K, which is
- *much too hot* to be bound by the gravity produced by *visible matter of the cluster*.

Thus, the *hot and luminous X-ray gas bound in many galaxy clusters* signals a large contribution of dark matter to the cluster gravitational field.

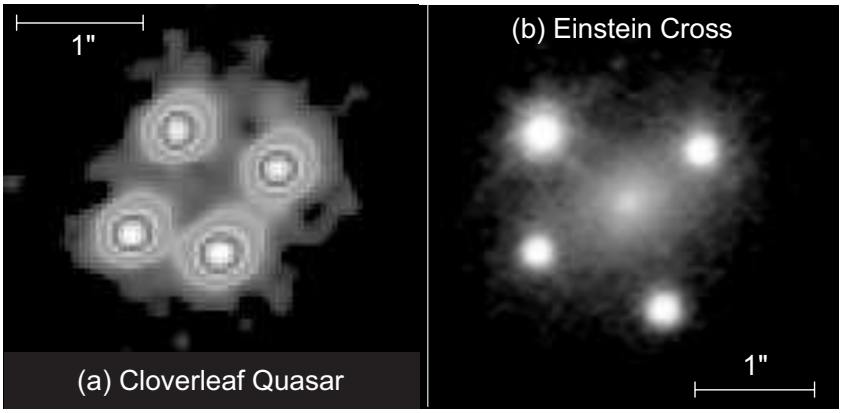


Figure 17.6: Gravitational lensing of quasars: (a) The Cloverleaf Quasar. The four images are of a single quasar lensed by foreground galaxies too faint to see in this image. (b) The Einstein Cross. The four outer images are all of a single quasar lensed by a foreground galaxy near the center of the image. Identical spectra confirm that these are images of a single object.

17.7.4 Gravitational Lensing

The path of light is curved in a gravitational field.

- This can cause *gravitational lensing*, where intervening mass acts as a "lens" to alter the image of distant objects.
- Spectacular examples of lensing are shown in Fig. 17.6.
- In these images, *a single distant object appears as four objects* because of lensing by a foreground galaxy.

The amount of lensing depends on the *total mass causing the lensing*, whether visible or not.

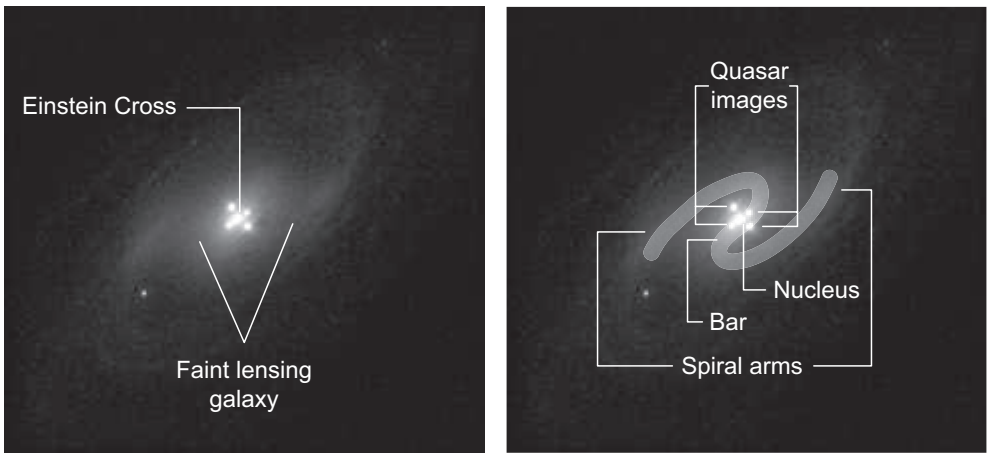


Figure 17.7: The Einstein Cross and its lensing galaxy. Intensity has been displayed on a logarithmic scale so that the very bright quasar images and the extremely faint bar and arms of the lensing galaxy can be seen at the same time (see annotated image on the right).

The lensing interpretation of the *Einstein Cross* from the previous figure is bolstered by Fig. 17.7.

- The left image shows in faint outline the foreground lensing galaxy and the four quasar images.
- The lensing galaxy is a relatively nearby barred spiral.

  Both the spiral arms and the central bar of the foreground galaxy can be seen faintly (see the annotation in the right panel).

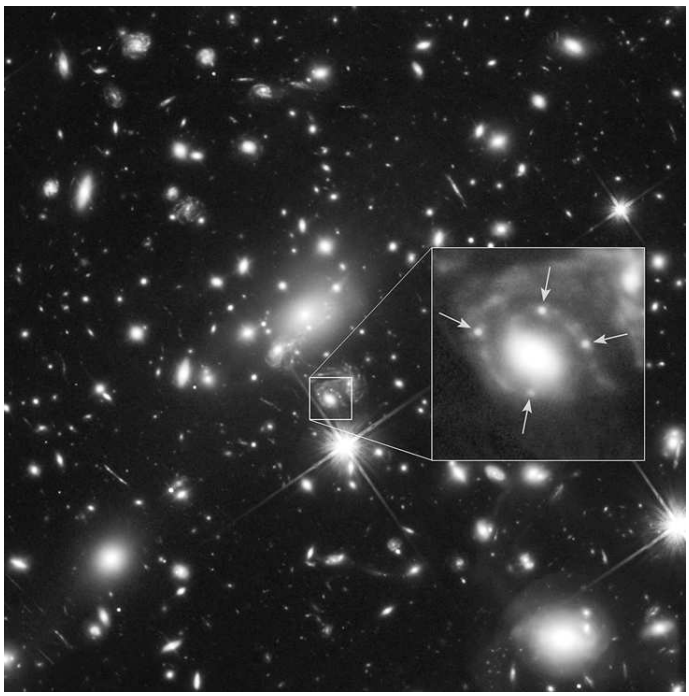


Figure 17.8: Light from a supernova in a galaxy at $z = 1.49$ split into four images by gravitational lensing from a cluster of galaxies at $z = 0.54$. The positions of the four supernova images are indicated by arrows in the inset box. Most of the galaxies in the image are members of the lensing cluster.

Another example of a cross, in this case for lensing of a supernova in a distant galaxy, is displayed in Fig. 17.8.

- The supernova host galaxy is a spiral at redshift $z = 1.49$.
- It is being lensed by a massive elliptical galaxy in a cluster of galaxies at $z = 0.54$.
- The gravitational potential of the entire cluster also produces multiple images of the supernova host galaxy.

Different magnifications and staggered arrival times of the images carry information about the supernova and the lens.

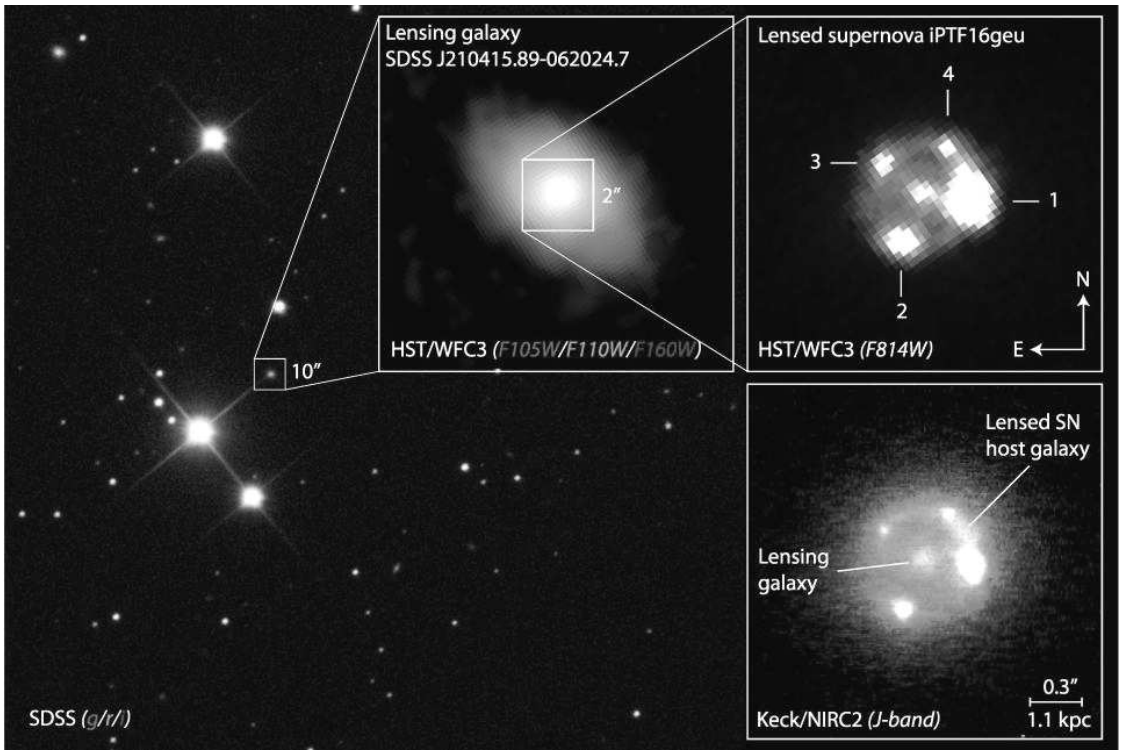
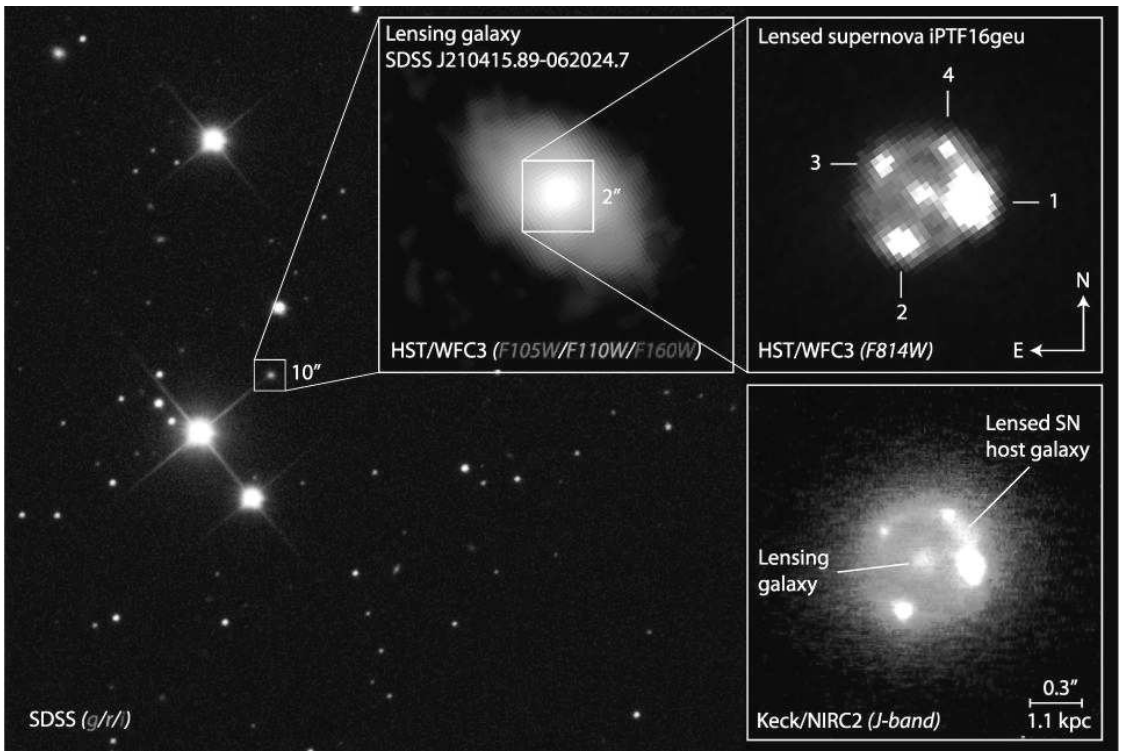


Figure 17.9: Lensed Type Ia supernova. The supernova and lensing galaxy are almost collinear, giving four images of the supernova and a partial Einstein ring from lensing of the supernova host galaxy (lower right panel).

A lensed image of a *Type Ia supernova* is shown in Fig. 17.9.

- Normally, the true brightness of a lensed object is unknown.
- However, because of the *standardizable candle properties of Type Ia supernovae*, the actual brightness of the lensed object and its variation with time are well determined.

The apparent brightness of the supernova has been *amplified by a factor of 50* in this lens.



Furthermore, by *using the known variation of the luminosity with time* (the standardized lightcurve of the supernova),

- the delay times for light in each of the four images can be measured precisely.
- The difference in arrival times for light in the four lensed images is *inversely proportional to the Hubble parameter*.
- Conversely, if a *cosmological model is assumed* the time delays measure directly the *gravitational potential* sensed by the four images.
- This permits reconstruction of the *lens mass distribution*.

Summarizing, the strength of a gravitational lens depends on the *total mass contained within it,* whether that mass is visible or not.

- Gravitational lenses can serve as excellent indicators of how much unseen matter is present in the lens,
- and even of the distribution of mass within the lens.
- Systematic analysis of gravitational lensing leads to conclusions similar to those suggested above by the rotation curves for spiral galaxies and the properties of galaxy clusters:

> More than 90% of the mass contributing to the strength of large gravitational lenses is dark.

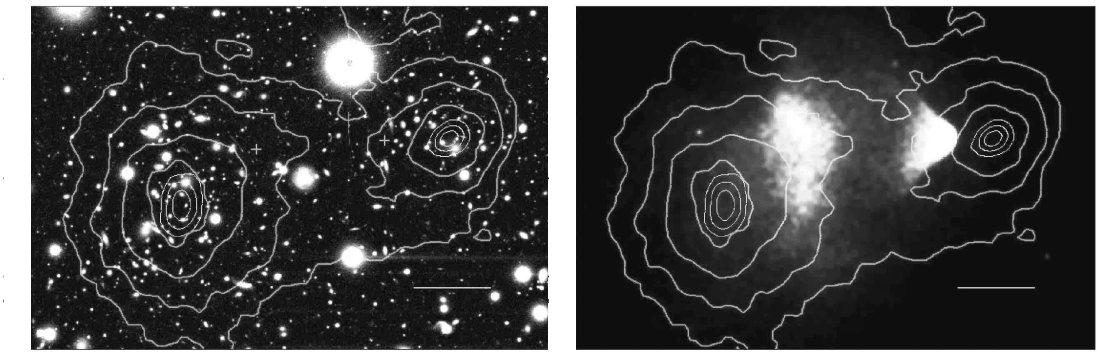
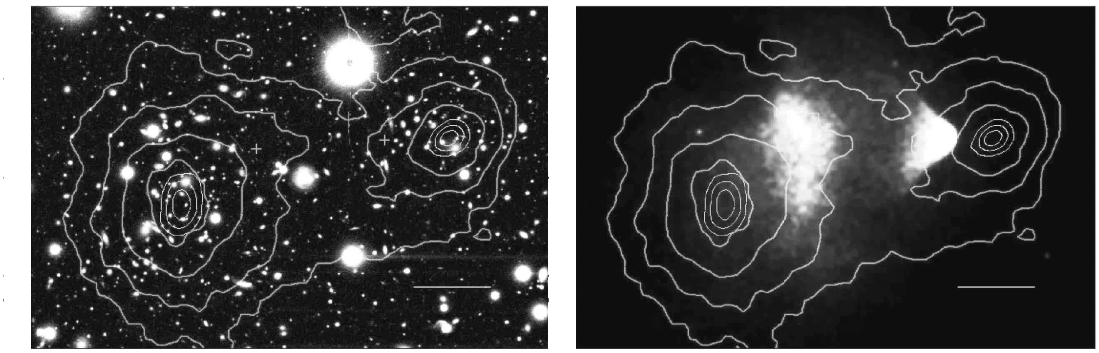


Figure 17.10: Evidence for dark matter in the Bullet Cluster. The left image shows galaxies in the cluster and total mass contours inferred from gravitational lensing. The right image shows X-ray luminosity superposed on mass contours. The simplest explanation for displacement of X-ray luminosity from mass concentrations is that the majority of the mass is dark matter now found at the two mass centers.

17.7.5 Example: The Bullet Cluster

Fig. 17.10 shows *evidence for dark matter* in galaxy clusters.

- The double cluster of galaxies 1E0657-558 (*Bullet Cluster*) has been studied using *gravitational lensing* and *detection of X-rays*.
- The double cluster represents *two galaxy clusters that collided* about 100 million years ago.
- The *star distributions* would have largely passed through each other, but the
- *gas* would have interacted strongly through ram pressure,
- while *dark matter* would have not interacted at all.



After the collision, as shown above,

- The compressed gas,
  - *radiating X-rays* (bright spots), is
  - *displaced from the mass centers* of the two clusters (indicated by contours).
- The collision has
  - separated the dark matter (at the two local maxima of the mass contours)
  - from the regular matter (concentrated at the sources of X-ray luminosity).

A number of other examples similar to the Bullet Cluster of *apparent separation of normal and dark matter by collisions* are now known.

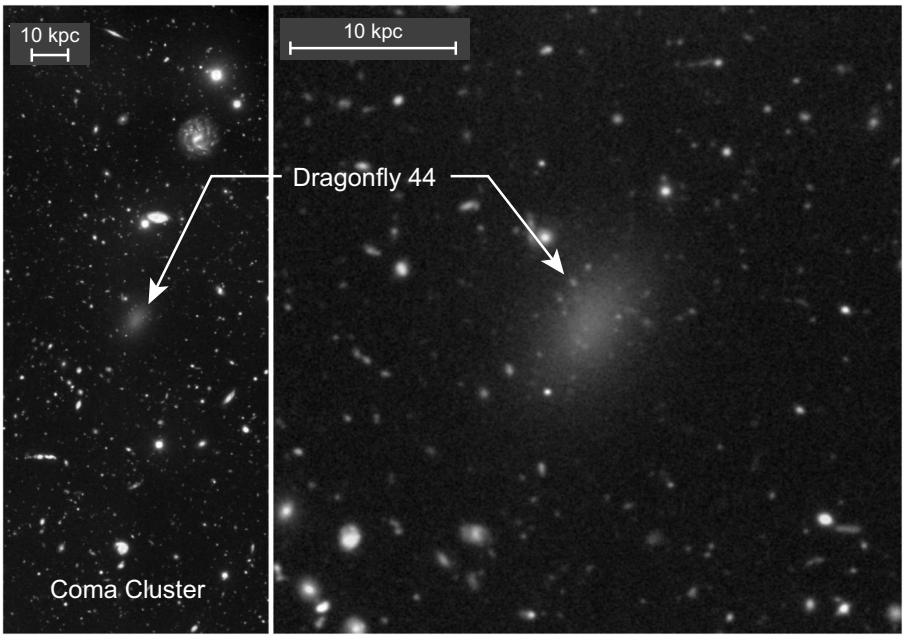


Figure 17.11: Dragonfly 44, a galaxy that may be almost all dark matter. Although it has a very low surface brightness, it contains of order 100 globular clusters and has a mass comparable to that of the Milky Way galaxy.

17.7.6 Dark Matter in Ultra-Diffuse Galaxies

A population of *ultra-diffuse galaxies (UDG)* has been studied in the *Coma Cluster*.

- These are *very faint* but appear to have *large masses*.
- The UDG *Dragonfly 44* is shown in Fig. 17.11.
- From velocity dispersion, the *mass is similar to the Milky Way*, and ~ 100 globular clusters were identified.

From the *large mass but faint light*, it was estimated that *Dragonfly 44 is 98% dark matter*.

17.8 Baryonic and Non-Baryonic Matter

Baryonic matter is "ordinary" stuff made of protons and neutrons. *Non-baryonic matter* consists of particles that do not undergo the strong interactions.

Neutrinos are one example of non-baryonic matter.

There seems to be *a lot of dark matter* in the Universe.

- *How much of it is baryonic,* ("ordinary stuff") and how much is more exotic *(non-baryonic matter)*?
- We can answer that question by considering *how many photons and baryons* there are in the Universe.
- Define the ratio η of baryon number density $n_{\rm B}$ to photon number density n_γ,

$$\eta \equiv \frac{n_{\rm B}}{n_\gamma} \simeq \frac{\Omega_{\rm b}\rho_{\rm c}/m_{\rm B}}{410\,{\rm cm}^{-3}} \simeq 2.76\times 10^{-8}\Omega_{\rm b}h^2,$$

  - where $\Omega_{\rm b} \equiv n_{\rm B}/\rho_{\rm c}$ is the *baryon density parameter*,
  - $\rho_{\rm c}$ is the *closure density*,
  - $m_{\rm B}$ is the average *mass of a baryon*, and
  - the *photon density* has been approximated by the *cosmic microwave background* density, $n_\gamma = 410\,{\rm cm}^{-3}$.

The abundances of light isotopes produced by *nucleosynthesis in the big bang* such as

- ^{4}He,
- ^{3}He,
- ^{2}H,
- ^{7}Li

are very sensitive to the quantity

$$\eta \equiv \frac{n_{\rm B}}{n_\gamma} \simeq \frac{\Omega_{\rm b}\rho_{\rm c}/m_{\rm B}}{410\,{\rm cm}^{-3}} \simeq 2.76\times 10^{-8}\Omega_{\rm b}h^2,$$

- The *measured values of these isotopic abundances* indicate that
$$\eta \sim 6\times 10^{-10}.$$
- Therefore, from this observed value of η
$$\Omega_{\rm b} \simeq (3.6\times 10^7 h^{-2})\eta \simeq 0.04,$$
indicating that baryonic matter contributes *less than 5% of closure density*.
- Hence, *strong nucleosynthesis constraints* indicate that

> Most the the matter producing the gravity of the Universe *cannot be baryonic*.

Extending the trend started by Copernicus:

- We are not the center of the Universe, and
- We aren't even made up of the dominant matter of the Universe.

> Not only are we not the center of the Universe, *we aren't even made of the right stuff!*

If the Universe is full of dark matter invisible to non-gravitational probes, what could it be?

17.9 Baryonic Candidates for Dark Matter

We address first possible *baryonic candidates* for the dark matter. The simple possibilities can be *dismissed quickly*.

- If the dark matter were a *cold baryonic gas*
  - it would not radiate, but a cold gas would be heated by hot gas and radiation,
  - and therefore would eventually become visible.
- Likewise, *cold dust* would be heated and re-radiate light, so it too would become visible.
- *Black holes* could hide large amounts of matter but
  - they would be more likely near the centers of galaxies where densities are higher,
  - not out in halos where dark matter is most required.
- The most promising candidates for dark baryonic matter are the *Massive Compact Halo Objects (MACHOS):*
  - Jupiter-like objects, or
  - more massive *brown dwarfs* with too little mass to form stars.
- If such objects were abundant in the halos of the galaxies, they would be difficult to see because of low luminosities.

A typical *MACHO experiment* looks for

- *gravitational lensing* when the MACHO moves in front of a star being imaged in the Large Magellanic Cloud.
- The expected signature is *brightening over days or weeks* for a star not normally a variable star.

Searches indicate the *presence of MACHOS* in the halo of our galaxy, but *too few to account for most dark matter*.

- In addition to the absence of direct evidence for sufficient baryonic dark matter,

  All baryonic dark-matter solutions violate the *nucleosynthesis constraint* discussed above.

- There is good observational evidence favoring a total density $\Omega = 1$ and baryonic density $\Omega_b \sim 0.04$.

Thus, the

- dominant matter in the Universe is *probably not baryonic* and
- the baryonic matter that we are composed of is but *a minor pollutant* in the Universe.

17.10 Candidates for Non-Baryonic Dark Matter

There are two broad classes of candidates for *non-baryonic dark matter:*

- *Cold Dark Matter (CDM),*
  - which consists of particles that *decoupled very early*,
  - or particles that were *never in thermal equilibrium*.
- *Hot Dark Matter (HDM),* which consists of
  - *low-mass particles* that still had
  - *relativistic velocities* at the time of matter–radiation decoupling.

Each of these corresponds to either *conjectured or known elementary particles*.

17.10.1 Cold Dark Matter

Cold dark matter had $v \ll c$ when galaxy formation started. Candidates may be divided into

- *Weakly Interacting Massive Particles (WIMPS)*, and
- *Superlight particles with superweak interactions* that were never in equilibrium.

WIMPS would decouple from the plasma earlier than normal leptons. Some proposed candidates include

- several particles expected for *supersymmetric theories*,
- and a *neutrino* with mass greater than 45 MeV.

If dark matter is *superlight particles never in equilibrium*

- they would have been *decoupled from the beginning*.
- A prime candidate is the *axion*,
- which is a *conjectured boson* required in some elementary particle physics theories.

There is *no experimental evidence* for

- supersymmetric particles,
- massive neutrinos,
- or axions

at the present time.

17.10.2 Hot Dark Matter

Hot dark matter is *relativistic when galaxy formation begins*.

- It could correspond to as-yet undiscovered particles but *neutrinos are obvious candidates*.
- The *present number density of neutrinos* may be estimated
  - by assuming that most neutrinos are in a uniform background of neutrinos analogous to the
  - cosmic microwave background to be discussed later.
- The number density of neutrinos in this background should be related to the number density of the photons in the microwave background by a factor of $\frac{3}{11}$.
- Thus, the *number density for each of the three neutrino families* is

$$n_\nu \simeq \frac{3}{11} n_\gamma \simeq 112\,\text{neutrinos}\,\text{cm}^{-3}.$$

- Current data indicate that *neutrinos have a tiny mass*.

Thus the contribution to closure density of all neutrinos in the Universe can be no more than several percent.

17.11 Dark Energy

Dark matter may appear exotic by normal standards, since

- we don't know what it is and therefore
- we do not know why it fails to couple strongly

through any force other than gravity.

- However, we shall see later that the evolution of the present Universe is being dominated by something even more exotic: *dark energy*.
- *Dark energy* (also called *vacuum energy*) behaves fundamentally differently from either
  - normal matter and energy,
  - or dark matter.

> *Dark energy* causes gravity to effectively become *repulsive*.

- To understand and to deal adequately with this remarkable notion will require a *covariant formulation of gravity*.

> Therefore, we defer substantial discussion of the evidence for and role played by dark energy until the following chapters.

17.12 Radiation

Astronomers classify massless and nearly massless particles such as photons, gluons, gravitons, and neutrinos as *radiation.*

- The radiation in the Universe influences its evolution
- The energy density of radiation in the present Universe is very small.
- However, it dominated the energy density of the very early Universe.
- Only a small amount of radiation density in the present Universe is found in starlight.
- The bulk (more than 90%) is in the cosmic microwave background (CMB) radiation, which will be discussed in later chapters.

17.13 Density Parameters

We have already introduced the *total density parameter* evaluated at the present time

$$\Omega \equiv \frac{\rho}{\rho_c} = \frac{8\pi G\rho}{3H_0^2}.$$

where ρ is the *current* total density coupled to gravity.

- Thus, *closure* implies that $\Omega = 1$ (*critical density*).
- The subscript "0" is often used on Ω and ρ to indicate explicitly that they are *evaluated at the present time*.
- We will often *suppress that subscript* to avoid notational clutter in later equations.

> Unless otherwise noted, you should *always* understand cosmological parameters to be *evaluated at the present time,* even if there is no subscript zero.

- Anticipating later treatment of the expansion using general relativity, we expect that
- the density parameter gets contributions from *three major sources* in the current Universe:

1. *Matter*, including dark matter (with density ρ_m)
2. *Radiation* (with density ρ_r)
3. *Vacuum or dark energy* (with density ρ_Λ).

These densities may be used to define corresponding *partial density parameters* Ω_i through

$$\rho_{\rm r}(a) = \rho_{\rm c}\Omega_{\rm r} \qquad \rho_{\rm m}(a) = \rho_{\rm c}\Omega_{\rm m} \qquad \rho_\Lambda(a) = \rho_{\rm c}\Omega_\Lambda,$$

and we will see later that the *total density* changes with $a(t)$ as

$$\rho(a) = \rho_{\rm c}\left(\frac{\Omega_{\rm r}}{a^4} + \frac{\Omega_{\rm m}}{a^3} + \Omega_\Lambda\right) \quad (a(t_0) \equiv 1),$$

- we have assumed the *standard convention* of normalizing the current value of the scale parameter $a(t_0)$ to unity.
- We shall make *no explicit distinction* between mass density ρ and the corresponding energy density $\varepsilon = \rho c^2$, since they are numerically the same in $c = 1$ units.
- Note that the *different densities scale differently with* $a(t)$, and thus differently with time.

For *baryonic matter alone*, we have already seen that

$$\Omega_{\rm b} \simeq 0.04. \qquad \text{(baryonic matter)}.$$

where $\Omega_{\rm b}$ is a part of $\Omega_{\rm m}$.

This is well below the critical density ($\Omega = 1$) but

- the effect of *non-baryonic dark matter* and
- the effect of *dark energy*

must be added to give the true value of Ω.

Table 17.1: Density parameters

Source	Value ($\Omega_i = \rho_i/\rho_c$)
Total matter	$\Omega_m = 0.3$
Baryonic matter	$\Omega_B = 0.04$
Total radiation	$\Omega_r \leq 8 \times 10^{-5}$
Total vacuum	$\Omega_\Lambda = 0.7$
Curvature	$\Omega_c \leq 0.01$

Some estimates of the current density parameters for the radiation, matter, baryonic portion of the matter, and the vacuum energy are given in Table 17.1 (the curvature density entry will be explained later).

17.14 The Deceleration Parameter

The density of the Universe is clearly related to the rate at which the Hubble expansion is changing with time.

- If we expand the cosmic scale factor to second order in time,

$$a(t) \simeq a_0 + \dot{a}_0(t-t_0) + \tfrac{1}{2}\ddot{a}_0(t-t_0)^2$$

  where $\dot{a}_0 \equiv (da/dt)_{t=t_0}$, and so on,

- introduce the *deceleration parameter at the present time* $q_0 \equiv q(t_0)$ through

$$q_0 \equiv -a_0\frac{\ddot{a}_0}{\dot{a}_0^2},$$

- and utilize

$$\frac{\dot{a}_0}{a_0} = H_0,$$

- we obtain

$$a(t) = a_0\left(\underbrace{1 + H_0(t-t_0)}_{\text{Hubble}} - \underbrace{\tfrac{1}{2}H_0^2 q_0(t-t_0)^2}_{\text{correction}} + \ldots\right).$$

By expanding a^{-1} to second order in a binomial series the redshift z may be expressed in terms of

$$a(t) = a_0\left(1 + H_0(t-t_0) - \tfrac{1}{2}H_0^2 q_0(t-t_0)^2 + \ldots\right).$$

in the form

$$z = \frac{1}{a} - 1 = H_0(t_0 - t) + H_0^2(1 + \tfrac{1}{2}q_0)(t_0 - t)^2 + \ldots$$

This is a quadratic in $H_0(t_0 - t)$, which may be solved to give

$$H_0(t_0 - t) = z - (1 + \tfrac{1}{2}q_0)z^2.$$

Integrating

$$\frac{cdt}{a(t)} = dr.$$

by using the expansion

$$a(t) = a_0\left(1 + H_0(t-t_0) - \tfrac{1}{2}H_0^2 q_0(t-t_0)^2 + \ldots\right).$$

to first order, the *current proper distance for an object with redshift* z is found to be (Problem)

$$d(t_0) \simeq c(t_0 - t) + \tfrac{1}{2}cH_0(t_0 - t)^2 + \ldots$$

Inserting

$$H_0(t_0 - t) = z - (1 + \tfrac{1}{2}q_0)z^2.$$

in this expression and neglecting higher-order terms gives

$$d(t_0) = \frac{cz}{H_0}\left(1 - \frac{1+q_0}{2}z\right).$$

for the proper distance in terms of z and q_0.

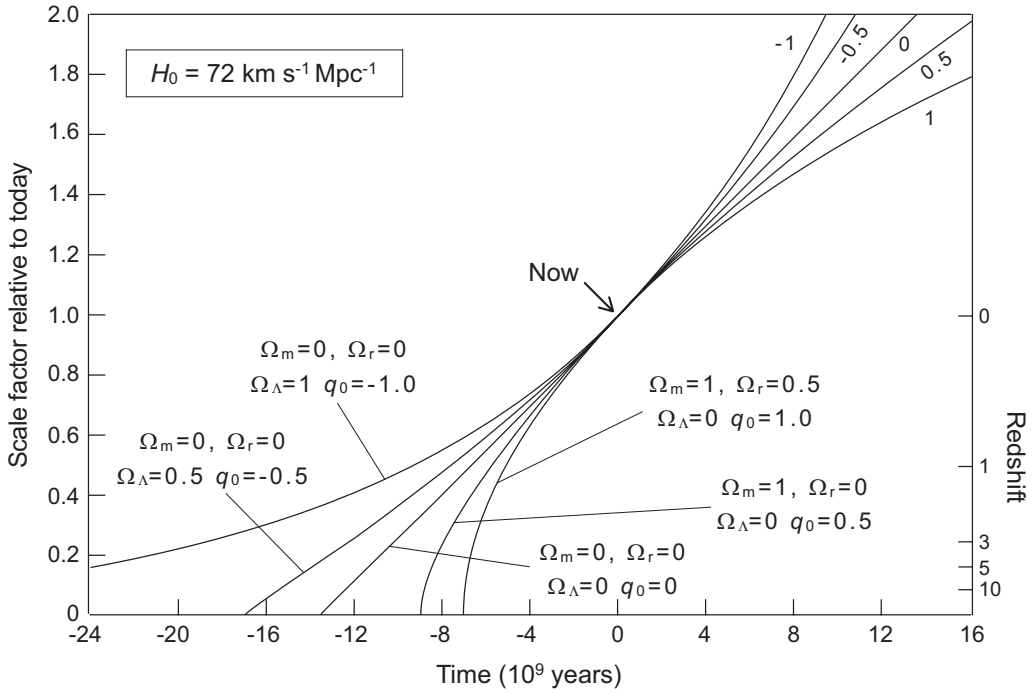


Figure 17.12: Deviations from the Hubble law to second order. The scale factor is shown on the left axis and the corresponding redshift on the right axis. Time is measured from today. Different curves correspond to different assumed values of the density parameters Ω_i. Each curve has the same linear term but a different quadratic (acceleration) term.

In the equation for the proper distance,

$$d(t_0) = \frac{cz}{H_0}\left(1 - \frac{1+q_0}{2}z\right).$$

the leading term is the Hubble law linear in z and the second term gives a correction one order higher in z.

Measurements at large enough z test correction terms to the linear Hubble law (see Fig. 17.12).

17.14.1 Deceleration and Density Parameters

The *deceleration parameter* q_0 is related to the *density parameters* Ω_i through

$$q_0 = \frac{\Omega_{\rm m}}{2} + \Omega_{\rm r} - \Omega_\Lambda .$$

In a *dust-only universe* $\Omega_{\rm r} = \Omega_\Lambda = 0$ and

$$q_0 = \frac{\Omega_{\rm m}}{2} \qquad \text{(dust only)}.$$

For a *flat universe* with radiation, matter, and vacuum energy,

$$\Omega = \Omega_{\rm r} + \Omega_{\rm m} + \Omega_\Lambda = 1 \qquad \text{(flat universe)},$$

which requires that

$$q_0 = \tfrac{3}{2}\Omega_{\rm m} + 2\Omega_{\rm r} - 1 \qquad \text{(flat universe)}.$$

Since $\Omega_{\rm m}$ and $\Omega_{\rm r}$ are non-negative,

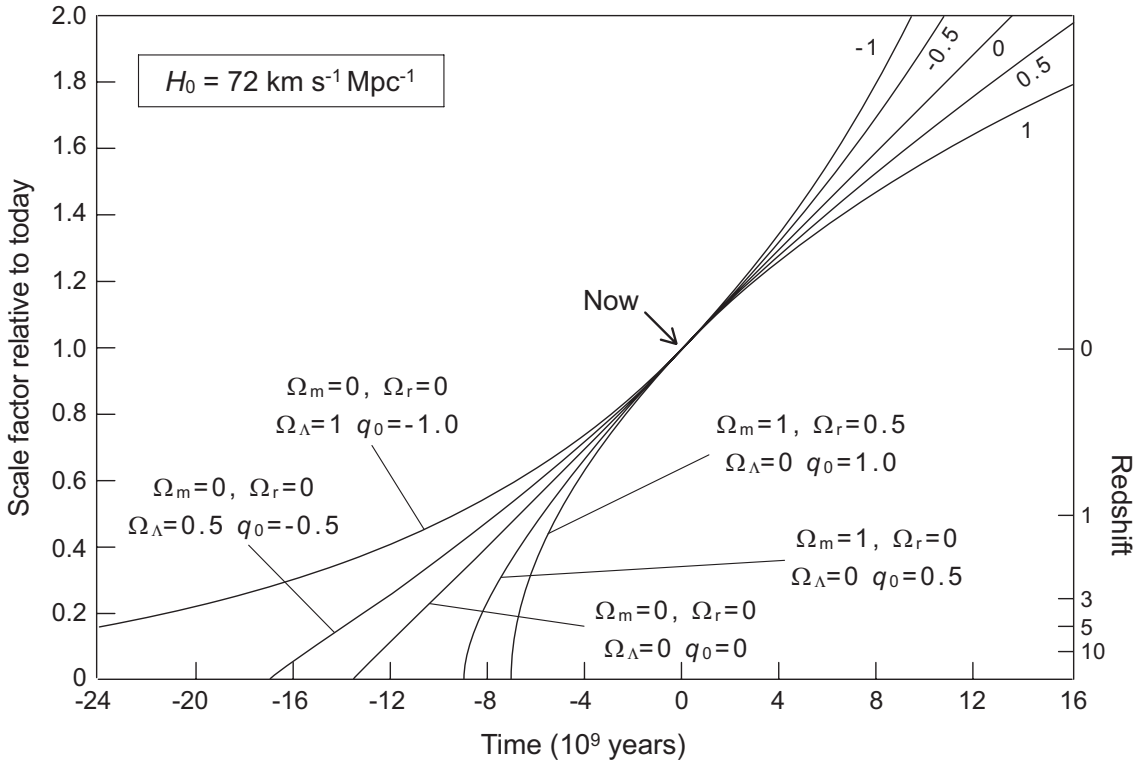
$$q_0 = \frac{\Omega_{\rm m}}{2} + \Omega_{\rm r} - \Omega_\Lambda .$$

implies that a negative q_0 *(acceleration of the expansion)* is possible only if there is a non-zero vacuum energy density.

For the *measured density parameters*, the deceleration parameter for the present Universe is *negative*,

$$q_0 \simeq \frac{\Omega_{\rm m}}{2} - \Omega_\Lambda \simeq -0.55,$$

and the expansion is currently *accelerating*.



Some *quadratic deviations from the Hubble law* for different combinations of density parameters are illustrated in the figure above.

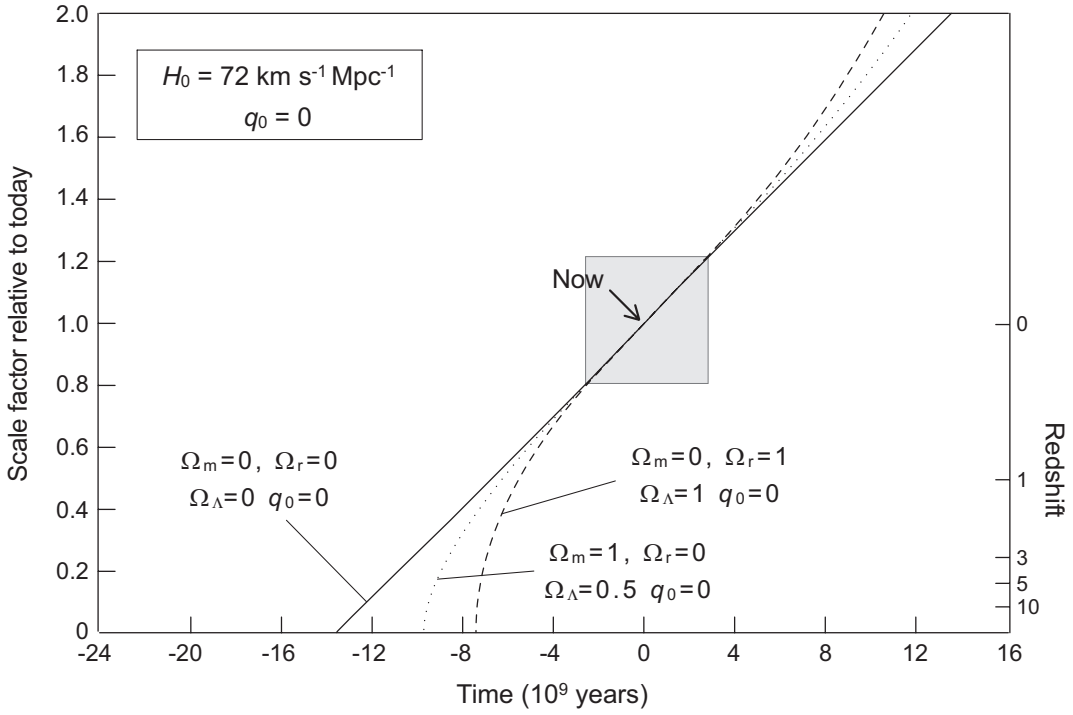


Figure 17.13: Different choices of matter, radiation, and vacuum energy densities giving the same deceleration parameter. Curves all agree near the present time to second order, but have very different long-time behaviors.

17.14.2 Deceleration and Cosmology

Figure 17.13 illustrates that H_0 and q_0 determine the behavior of the Universe *only near the present time*.

- The curves have the same H_0 and $q_0 = 0$, but different mixtures of matter, radiation, and vacuum energy.
- Within the gray box the curves are essentially indistinguishable but for $z \geq 1$ they are *very different*.
- For example, the curves predict ages of the Universe (intercepts with lower axis) differing by a factor of ~ 2.

Until recently the *primary quest in cosmology* was to determine with precision

- the *Hubble constant* H_0 and
- the *deceleration parameter* q_0.

Acquisition of *precise cosmology data* through

- The study of *high-redshift Type Ia supernovae* and
- Detailed analysis of fluctuations in the *cosmic microwave background radiation*

has changed that:

> Cosmological data now constrain a *broader range of parameters* than just H_0 and q_0, as we discuss in later chapters.

17.15 Problems with Newtonian Cosmology

We have made some headway in understanding cosmology using *Newtonian gravitational concepts*. However, a Newtonian approach leads to *problems and inconsistencies*. For example,

1. At large distance the expansion implies *recessional velocities* $v > c$. How are we to interpret this?
2. *Newtonian gravity acts instantaneously* but signals obey $v \leq c$, so there should be a *delay in the action of gravity*.
3. In the Newtonian picture we had
   - a uniform, finite, isotropic sphere expanding into nothing, which causes conceptual problems in interpreting the expansion.
   - Alternatively, if the sphere is of infinite extent, there are formal difficulties with even defining a potential.
4. The *radius of gravitational curvature* appears to be comparable to the actual "radius" of the known Universe, undermining the validity of Newtonian gravitation.
5. The mass–energy of the Universe receives a large contribution from *dark energy*. How do we deal with that?

These and other difficulties suggest that we need a *better theory of gravity* to describe cosmologies of expanding universes. *We know of such a theory!*

Chapter 18

Friedmann Cosmologies

At the end of the last chapter we motivated the necessity of a *covariant theory of gravity* to describe cosmology.

- Let us now consider *solutions of the Einstein equations* that may be relevant for describing the large-scale structure of the Universe and its time evolution.
- To do so, we must first make a choice for the *form of the spacetime metric* governing our Universe.

18.1 The Cosmological Principle

Possible forms for the *metric of spacetime* are strongly constrained by the *cosmological principle:*

> The Universe on large scales is *homogeneous* (no preferred place) and *isotropic* (no preferred direction).

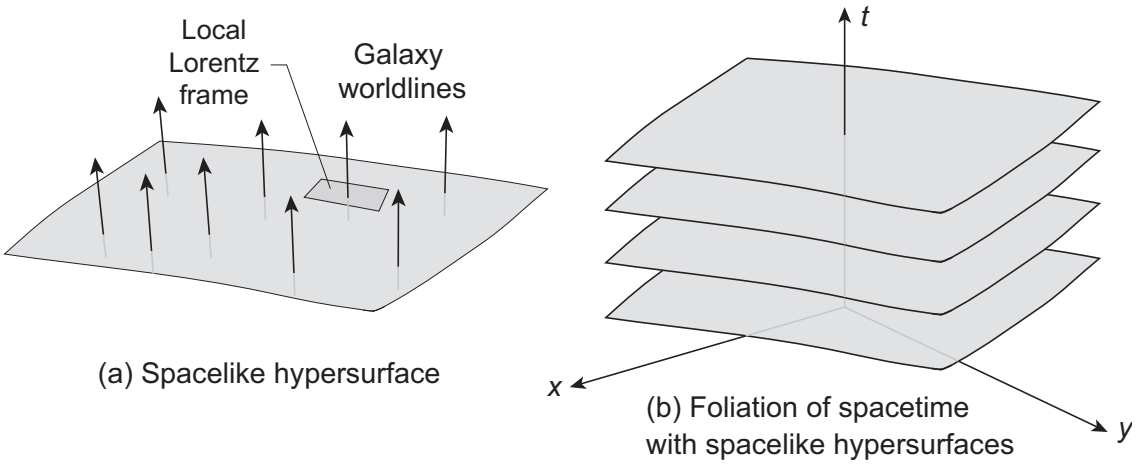
- This implies a proper time such that at any instant the *3D spatial line element* of the Universe,

$$d\ell^2 = g_{ij}dx^1 dx^j \qquad (i,j=1,2,3)$$

(where g_{ij} is the spatial part of the metric tensor) is *the same in all places and all directions*, with

$$\begin{aligned} ds^2 &= -dt^2 + a(t)^2 d\ell^2 \\ &= -dt^2 + a(t)^2 g_{ij}dx^i dx^j. \end{aligned}$$

- The *scale parameter* $a(t)$ describes *expansion or contraction of the spatial metric*.
- Cosmological principle $\rightarrow$ no reason for time to pass at different rates for different locations in an isotropic and homogeneous Universe.
- If time depended on coordinates, measurement of time could distinguish one place from another *(contradiction)*.
- Thus the *time term is simply* dt, and not a more complicated expression as in the Schwarzschild metric.



(a) Spacelike hypersurface

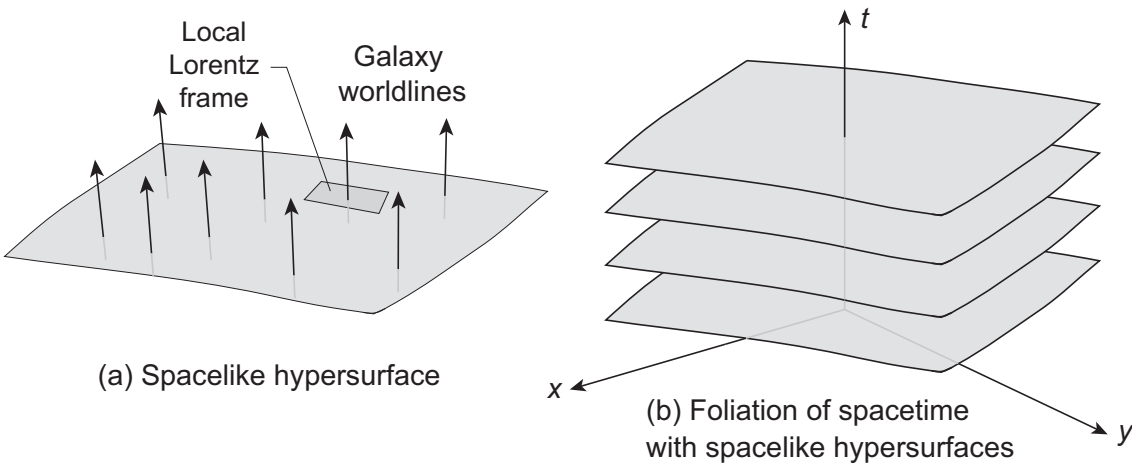
(b) Foliation of spacetime with spacelike hypersurfaces

18.1.1 Cosmological Proper Time

A global time has meaning if specific requirements are met.

- A metric embodying the cosmological principle *meets those requirements*.
- Then a global time termed the *cosmological proper time* or *cosmic time* can be introduced by a *foliation* of spacetime in terms of a sequence of non-intersecting spacelike 3D surfaces, as illustrated in the figure above.
- We assume that all galaxies lie on such a hypersurface such that the surface of simultaneity for the Lorentz frame of each galaxy coincides locally with the hypersurface.

Thus a hypersurface contains the *smoothly-meshed Lorentz frames of all galaxies*, with the 4-velocity of each galaxy orthogonal to the local hypersurface.



(a) Spacelike hypersurface

(b) Foliation of spacetime with spacelike hypersurfaces

This series of hypersurfaces may be *labeled by a parameter* that can be viewed as a *global cosmic time* for all galaxies on the hypersurface,

- but only if the space is *homogeneous and isotropic*,
- which implies further that *spatial curvature is constant*.

Operationally, *cosmic time* is the time

- measured by any observer who sees the Universe *expanding uniformly around her*.
- Such an observer is said to be a
  - *comoving observer* or a
  - *fundamental observer*.

All frames are *equally valid* but the Universe is *particularly simple* when viewed from comoving frames.

Thus, we must investigate *3D curved manifolds*

- that are both *homogeneous and isotropic*, and
- parameterized by a *cosmic time*.

Let's first consider this question in two dimensions, where visualization is easier, and then generalize to three spatial dimensions.

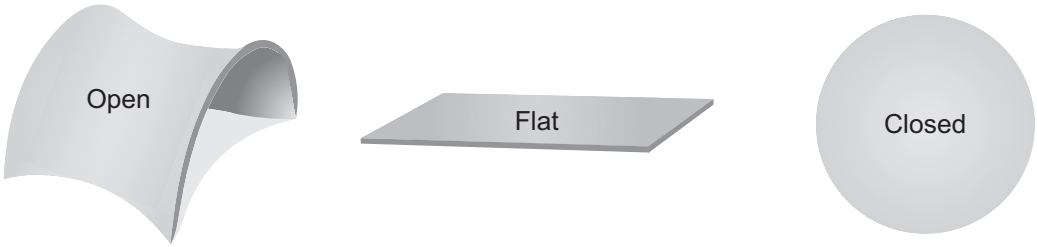


Figure 18.1: Isotropic, homogeneous 2-dimensional spaces.

18.2 Homogeneous and Isotropic 2D Spaces

In 2D there are *three independent possibilities* for homogeneous, isotropic spaces:

1. *Flat* euclidean space.
2. A *sphere* of constant (positive) curvature.
3. An *hyperboloid of* constant (negative) curvature.

These three possibilities are illustrated in Fig. 18.1.

> Constant negative curvature surfaces can't be fully embedded in 3D euclidean space. The saddle-like open surface only approximates constant negative curvature near its center.

- In each case the corresponding space has neither a special point nor a special direction.
- Thus, these are 2D spaces with metrics that are consistent with the cosmological principle.

Let's examine the *2-sphere* as a representative example.

- We are used to thinking of 2-spheres in terms of a *2D surface embedded in a 3D space*.
- But a metric defines *intrinsic properties of a manifold* that should be independent of any additional embedding dimensions (recall *Gaussian curvature*).
- Therefore, it should be possible to express the metric of the 2-sphere in terms of *only two coordinates*.

From

$$x^2 + y^2 + z^2 = S^2,$$

we may deduce that

$$dz^2 = \frac{(xdx + ydy)^2}{S^2 - x^2 - y^2}$$

- Thus, we may write the *metric for the 2-sphere* in the form

$$\begin{aligned} d\ell^2 &= dx^2 + dy^2 + dz^2 \\ &= dx^2 + dy^2 + \frac{(xdx + ydy)^2}{S^2 - x^2 - y^2} \end{aligned}$$

- This metric describes distances on a 2D surface and *depends on only two coordinates* (S is a constant).

Distances may be specified *entirely by coordinates intrinsic to the 2D surface,* independent of 3rd embedding dimension.

18.3 Homogeneous and Isotropic 3D Spaces

Let us now generalize the discussion of the preceding section.

- Our spacetime appears to have three rather than two spatial dimensions.
- Thus we consider the embedding of 3D spaces in four Euclidean dimensions.

Mathematically this will be very similar to the 2D case just considered.

18.3.1 Constant Positive Curvature

For a *3-sphere* the generalization is obvious:

$$x^2 + y^2 + z^2 + w^2 = S^2,$$

where w is now a fourth dimension.

- The metric of the 3-sphere is independent of the embedding and expressible in terms of only 3 of the coordinates.
- By analogy with the 2-sphere, we may write

$$d\ell^2 = dx^2 + dy^2 + dz^2 + \frac{(xdx + ydy + zdz)^2}{1 - x^2 - y^2 - z^2},$$

  where we've *specialized to a unit 3-sphere* ($S = 1$) because the overall spatial scale will be set by $a(t)$.

Introduce spherical polar coordinates (r, θ, φ) through

$$x = r\sin\theta\cos\varphi \qquad y = r\sin\theta\sin\varphi \qquad z = r\cos\theta.$$

- The *metric of the unit 3-sphere* then takes the form

$$\begin{aligned} d\ell^2 &= \frac{dr^2}{1 - r^2} + r^2 d\theta^2 + r^2\sin^2\theta d\varphi^2 \\ &= \frac{dr^2}{1 - r^2} + r^2 d\Omega^2 \end{aligned}$$

  where we have defined

$$d\Omega^2 \equiv d\theta^2 + \sin^2\theta d\varphi^2$$

The *3-sphere*, by analogy with the 2-sphere,

- Corresponds to a *homogeneous, isotropic space* that is *closed and bounded*.
- It has *great circles as geodesics*.
- It is a space of *constant positive curvature*.
- An ant dropped onto the surface of an otherwise featureless 3-sphere would find that
  - No point or direction appears any different from any other (*homogeneous and isotropic*).
  - The shortest distance between any two points corresponds to a segment of a great circle.
  - A sufficiently long journey in a fixed direction would return one to the starting point.

From these observations the ant concludes that its universe has *no boundary* but that its total *volume is finite*.

18.3.2 Constant Negative Curvature

The generalization of the 2-hyperboloid to three dimensions is given by

$$x^2 + y^2 + z^2 + w^2 = -S^2.$$

By a similar argument as above, the metric in this case can be expressed as

$$d\ell^2 = \frac{dr^2}{1+r^2} + r^2 d\Omega^2,$$

- This describes a space that is *homogeneous and isotropic*, but now *unbounded and infinite*.
- It has *constant negative curvature*, with *hyperbolas as geodesics*.
- Our hypothetical ant explorer would find
  - No preferred direction or location.
  - The shortest distances between any two points would now be segments of hyperbolas.
  - The ant would never return to the starting point by continuing an infinite distance in a fixed direction.
  - The ant would conclude that the volume of the space is infinite.

18.3.3 Zero Curvature

Finally, for a Euclidean 3-space the metric may be expressed in the form

$$d\ell^2 = dr^2 + r^2 d\Omega^2,$$

This space

- *is homogeneous and isotropic*, and
- is of *infinite extent*, with *straight lines as geodesics*.
- Obviously, this space corresponds to the limit of *no spatial curvature*.
- The *volume is infinite*, and a straight path in one direction will never return to the starting point.

18.4 The Robertson–Walker Metric

Combining preceding results, the most general *3D spatial metric* that incorporates *isotropy* and *homogeneity* is

$$d\ell^2 = \frac{dr^2}{1-kr^2} + r^2 d\Omega^2,$$

where the parameter k determines the nature of the curvature:

$$k = \begin{cases} +1 & \text{hypersphere of positive curvature} \\ 0 & \text{flat Euclidean space} \\ -1 & \text{hyperboloid of negative curvature} \end{cases}$$

The parameter k can have other normalizations but one is free to rescale equations so that it takes only these values.

Finally, combining

$$d\ell^2 = \frac{dr^2}{1-kr^2} + r^2 d\Omega^2 \qquad ds^2 = -dt^2 + a(t)^2 d\ell^2$$

we arrive at the most general spacetime metric consistent with homogeneity and isotropy *(cosmological principle)*,

$$ds^2 = -dt^2 + a(t)^2 \left(\frac{dr^2}{1-kr^2} + r^2 d\theta^2 + r^2 \sin^2\theta d\varphi^2 \right),$$

where $k = 0, \pm 1$.

The metric specified by the line element

$$ds^2 = -dt^2 + a(t)^2 \left(\frac{dr^2}{1-kr^2} + r^2 d\theta^2 + r^2 \sin^2\theta d\varphi^2 \right),$$

is commonly called the *Robertson–Walker (RW) metric*.

- It is the starting point for any description of our Universe on scales sufficiently large that the cosmological principle applies.
- The time variable t appearing in the Robertson–Walker metric is the time that would be measured by an observer who sees *uniform expansion of the surrounding Universe*.

The Robertson–Walker time t is termed the *cosmological proper time* or the *cosmic time.*

In *matrix form* the Robertson–Walker line element is

$$
\begin{aligned}
ds^2 &= g_{\mu\nu} dx^\mu dx^\nu \\
&= (dt\ dr\ d\theta\ d\varphi)
\begin{pmatrix}
-1 & 0 & 0 & 0 \\
0 & \dfrac{a^2}{1-kr^2} & 0 & 0 \\
0 & 0 & a^2 r^2 & 0 \\
0 & 0 & 0 & a^2 r^2 \sin^2\theta
\end{pmatrix}
\begin{pmatrix} dt \\ dr \\ d\theta \\ d\varphi \end{pmatrix},
\end{aligned}
$$

with nonvanishing covariant metric tensor components

$$
g_{00} = -1 \quad g_{11} = \frac{a^2}{1-kr^2} \quad g_{22} = a^2 r^2 \quad g_{33} = a^2 r^2 \sin^2\theta,
$$

and corresponding contravariant components

$$
g^{00} = -1 \quad g^{11} = \frac{1-kr^2}{a^2} \quad g^{22} = \frac{1}{a^2 r^2} \quad g^{33} = \frac{1}{a^2 r^2 \sin^2\theta}.
$$

The metric is generally a *rank-2 symmetric tensor* with

- *10 independent components*.
- The symmetries assumed in the RW metric have reduced this to only *two independent parameters:*
  - the scale factor $a(t)$ and
  - the curvature parameter k.

> The cosmological principle has greatly simplified our description of cosmology.

The RW metric may be expressed in an *alternative form* by

- introducing *4D polar angles* (Problem)

$$w = \cos\chi \qquad x = \sin\chi\sin\theta\cos\varphi$$
$$y = \sin\chi\sin\theta\sin\varphi \qquad z = \sin\chi\cos\theta$$

  into the equations for spherical geometry, and

- these variables with the substitutions

$$w \to iw \qquad \chi \to -i\chi \qquad S \to iS,$$

  into the equations for hyperbolic geometry,

- and choosing $S = 1$ in both cases.

Then the RW metric may be written as

$$ds^2 = -dt^2 + a(t)^2 \begin{cases} d\chi^2 + \sin^2\chi(d\theta^2 + \sin^2\theta d\varphi^2) & (k=+1) \\ d\chi^2 + \chi^2(d\theta^2 + \sin^2\theta d\varphi^2) & (k=0) \\ d\chi^2 + \sinh^2\chi(d\theta^2 + \sin^2\theta d\varphi^2) & (k=-1) \end{cases}$$

which is related to

$$ds^2 = -dt^2 + a(t)^2\left(\frac{dr^2}{1-kr^2} + r^2 d\theta^2 + r^2\sin^2\theta d\varphi^2\right),$$

by the change of variables

$$r = \begin{cases} \sin\chi & (k=+1) \\ \chi & (k=0) \\ \sinh\chi & (k=-1) \end{cases}$$

Notice that

- The derivation of the Robertson–Walker metric was *purely geometrical*,
- subject to the constraints of *isotropy* and *homogeneity*.
- *No dynamical considerations* enter explicitly into its formulation.
- Dynamics will come later,
  - from the *Einstein field equations,*
  - with a metric of the RW form
  - which must be solved for the *time dependence of the scale factor* $a(t)$.

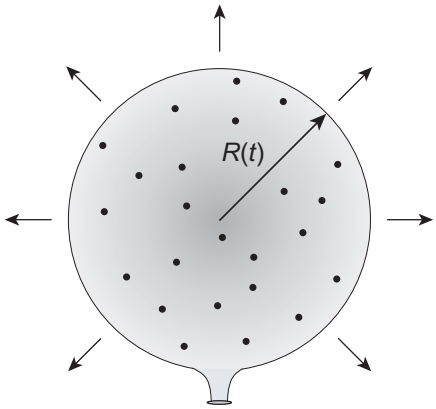


Figure 18.2: Hubble expansion in two spatial dimensions.

18.5 Comoving Coordinates

Homogeneous, isotropic expansion can be exemplified in 2D by placing dots on a balloon and blowing it up (Fig. 18.2).

- The spherical coordinates (θ, φ) are unchanged but distance between points changes with the scale factor $R(t)$.
- *Example:* 2 positions on the surface of the Earth.
  - Expand the size of the globe by a factor of 2.
  - Distance between 1 and 2 *doubles*,

  but the *coordinates* of the two cities *are unchanged*.
- Termed *comoving coordinates* or a *comoving frame*.
- Observer attached to a comoving coordinate sees all other points recede and sees a *homogeneous, isotropic universe.*
- The receding points *maintain their comoving coordinates.*
- An *observer not comoving* does *not* see isotropy.

The balloon analogy is useful, but one must guard against *misconceptions that it can generate*.

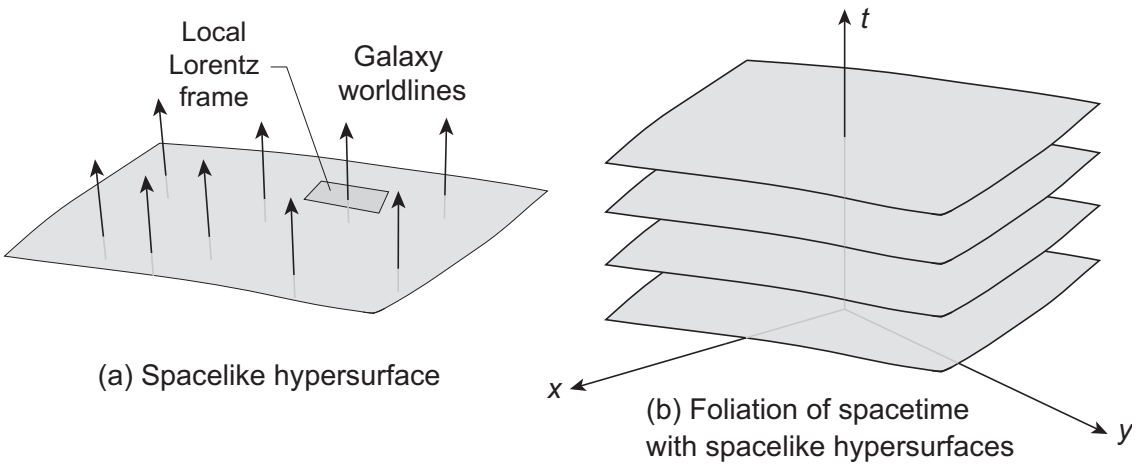
- First, the surface that is expanding is 2D.
  - The "center" of the balloon is in the third dimension.
  - It is not part of the 2D surface, which *has no center*.
- Second, the Universe is *not* being expanded by a pressure, nor is the balloon.
  - The expansion of the balloon is generated by a *difference in pressure.*
  - But in a homogeneous, isotropic universe there can be *no pressure differences on global scales*.
  - Furthermore, because pressure couples to gravity in the Einstein equation,
  - addition of (positive) pressure to the Universe would *slow,* not increase, the expansion rate.
- Finally, if the dots on the balloon represent galaxies, they too will expand.
- But *galaxies don't partake of the general Hubble expansion* because they are *gravitationally bound*.

A better analogy is to *glue solid objects* to the surface of the balloon to represent galaxies, so that they *don't expand when the balloon expands*.

If we now generalize this 2D comoving-dots-on-a-balloon idea to 4D spacetime,

- The coordinates (r, θ, φ) or (χ, θ, φ) of the RW metric are *comoving coordinates*.
- In the RW metric, as the Universe expands
  - the galaxies *keep the same comoving coordinates* (r, θ, φ) or (χ, θ, φ), and
  - only the RW scale factor $a(t)$ changes with time.
- Just as in the 2D analogy, the galaxies recede from us but,
  - if we are comoving observers the receding galaxies maintain their comoving coordinates and
  - the recession is described entirely by the *time dependence of the scale factor* $a(t)$.

Peculiar velocities change the comoving coordinates, but these are small on the large scales where the RW metric is valid.



(a) Spacelike hypersurface

(b) Foliation of spacetime with spacelike hypersurfaces

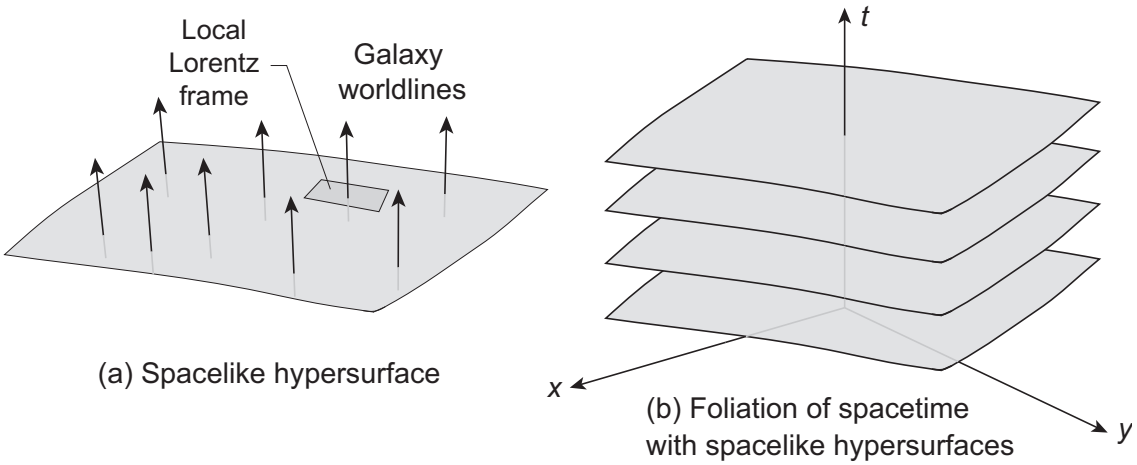
18.6 Proper Distances

Let's consider *measurement of distances* in the RW metric.

- Measuring distances on a cosmological scale, or even defining them, is a *nontrivial task*
- in a spacetime that is *expanding* and *possibly curved*.
- As a consequence, we shall discuss *several notions of distance* before we are finished.

The first is suggested by the diagram above, where

- we imagine cosmic time to be held fixed and
- use the metric to compute the distance between galaxies on the spacelike hypersurface corresponding to that fixed time.



(a) Spacelike hypersurface

(b) Foliation of spacetime with spacelike hypersurfaces

- We take one galaxy to be at comoving coordinates $(r,\theta,\varphi) = (0,0,0)$ and the other to be at $(r,0,0)$, at fixed time t.
- Thus, the portions of the line element depending on time and the angles θ and φ make no contribution and we obtain from

$$ds^2 = -dt^2 + a(t)^2\left(\frac{dr^2}{1-kr^2} + r^2 d\theta^2 + r^2 \sin^2\theta d\varphi^2\right)$$

$$\longrightarrow \quad ds^2 = a(t)^2\left(\frac{dr^2}{1-kr^2}\right)$$

the *proper distance*

$$\ell = a(t)\int_0^r \frac{dr}{\sqrt{1-kr^2}}.$$

Let's evaluate this for the three curvature parameter values.

1. *Flat Euclidean space ($k = 0$):* the integral is trivial and

$$\ell = a(t)\int_0^r \frac{dr}{\sqrt{1-kr^2}} \quad \longrightarrow \quad r = \frac{\ell}{a}$$

Thus, r increases without limit as ℓ increases at fixed $a(t)$.

- This implies a universe with *no boundary or curvature*, that is of *infinite extent*.
- Such a universe is said to be *flat*.

2. *Positive curvature ($k = +1$):* Solution gives

$$\ell = a\int_0^r \frac{dr}{\sqrt{1-r^2}} = a\sin^{-1} r \quad \longrightarrow \quad r = \sin\left(\frac{\ell}{a}\right).$$

In a universe of constant positive spatial curvature, r returns to the origin whenever $\ell = \pi a$ and

- this space has *no boundary* but is of *finite volume*.
- Such a universe is said to be *closed*.

3. *Negative curvature ($k = -1$):* Solution gives

$$\ell = a\int_0^r \frac{dr}{\sqrt{1+r^2}} = a\sinh^{-1} r \quad \longrightarrow \quad r = \sinh\left(\frac{\ell}{a}\right).$$

Therefore, r grows without limit in a universe of constant negative curvature as ℓ is increased at fixed $a(t)$.

- This space has *no boundary* and has *infinite extent*.
- Such a universe is said to be *open*.

The *proper distance*

$$\ell = a(t) \int_0^r \frac{dr}{\sqrt{1-kr^2}}.$$

- is the *"rulers end–to–end" distance* that would be
- measured by a set of observers with rulers distributed between the two objects at fixed cosmic time.
- This notion of distance is conceptually well-grounded.
- However, it is *impractical to implement in an astrophysics context* where instead
- essentially all distance information comes from data carried by signals propagating on null geodesics (light).

Therefore, we shall later have to consider more extensively

- the *meaning of distance* and
- how practically to *specify and measure it*

in observational astronomy.

18.7 The Hubble Law and the RW Metric

For a galaxy participating in the Hubble flow

- r is a *comoving coordinate* and so is *constant in time*.
- Therefore, illustrating for $k = 0$,

$$\ell = a(t)\int_0^r \frac{dr}{\sqrt{1-kr^2}} \quad \rightarrow \quad \dot{\ell} = \dot{a}(t)\int_0^r dr = \frac{\dot{a}(t)}{a(t)}\ell,$$

  where $\ell = a(t)r$ was used.

- This may be recognized as a *generalized form of Hubble's law*, with

$$v \equiv \dot{\ell} = H\ell \qquad H = \frac{\dot{a}(t)}{a(t)}.$$

Similar results follow from the equations for positive and negative curvature at small r.

- The RW metric, and the cosmological principle upon which it rests, imply the Hubble law.
- Conversely, observation of a Hubble law is an indicator of homogeneous, isotropic spacetime.

Hubble law $\longleftrightarrow$ RW metric.

18.8 Particle and Event Horizons

An important consequence of

- the metric structure of spacetime and
- the finite speed of light

is the possibility that regions of spacetime may be *intrinsically unknowable for a fixed observer*,

- either *at the present time*, or
- perhaps *for all time*.

Such limitations are termed *horizons*. We shall distinguish two related concepts:

- A *particle horizon*
- An *event horizon*.

> The former is of particular importance in cosmology and one often finds there the generic term *"horizon"* used to mean *"particle horizon"*.

18.8.1 Particle Horizons in the RW Metric

A *particle horizon* is

- the largest distance from which a light signal could have reached us at time t
- if it were emitted at time $t = 0$.

Imagine a spherical light wave emitted by us at the time of the big bang. (Anything is possible in a thought experiment!)

- Over time it sweeps out over more and more galaxies.
- By symmetry, at the same instant that those galaxies can see us we can see them.

So this spherical light front divides the galaxies into two groups:

1. Those *inside our particle horizon*, for which their light has had time to reach us since the big bang.
2. Those *outside our particle horizon*, for which their light has not yet had time to reach us since the bang.

In the RW metric, light travels along a geodesic

$$ds^2 = 0.$$

Choosing $\theta = \varphi = 0$, and inserting $ds^2 = d\theta = d\varphi = 0$ in

$$ds^2 = -dt^2 + a(t)^2 \left(\frac{dr^2}{1-kr^2} + r^2 d\theta^2 + r^2 \sin^2\theta d\varphi^2 \right)$$

we have

$$dt = \frac{a dr}{\sqrt{1-kr^2}}.$$

Integrating both sides gives

$$\int_0^t \frac{dt'}{a(t')} = \int_0^{r_\mathrm{p}} \frac{dr}{\sqrt{1-kr^2}},$$

where r_p is the comoving distance to the particle horizon. The proper distance to the particle horizon is

$$\ell_\mathrm{h} = a(t) \int_0^{r_\mathrm{p}} \frac{dr}{\sqrt{1-kr^2}}.$$

Combining the previous two expressions we obtain

$$\ell_\mathrm{h} = a(t) \int_0^t \frac{dt'}{a(t')}.$$

- Whether a particle horizon exists depends on the behavior of this integral.
- Convergence at the lower limit is not ensured because the scale factor in the denominator often tends to zero at the lower limit in RW cosmologies.

18.8.2 Example of Particle Horizon: Flat, Static Universe

As a simple example of a particle horizon, assume a *flat, static Universe*.

- Then the RW metric reduces to the *Minkowski metric* expressed in spherical coordinates with *a equal to a constant* and

$$\ell_{\rm h} = a(t)\int_0^{t_0}\frac{dt'}{a(t')} \quad \longrightarrow \quad \ell_{\rm h} = a\frac{ct_0}{a} = ct_0.$$

- This is just the distance light would travel in a time t_0 if the Universe were *flat and not expanding*.

Later we will see that this result can be modified substantially by expansion of the Universe.

18.8.3 Conformal Time and Horizons

- It is sometimes useful to introduce a new time coordinate η called the *conformal time* through

$$dt = a(t)d\eta.$$

- The flat Robertson–Walker metric then may be expressed as

$$\begin{aligned} ds^2 &= -dt^2 + a^2(dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\varphi^2) \\ &= -a^2 d\eta^2 + a^2(dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\varphi^2) \\ &= a^2(-d\eta^2 + dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\varphi^2), \end{aligned}$$

- This is the same form as the metric for a uniformly expanding Minkowski space.
- For radial light rays $d\theta = d\varphi = ds = 0$, so this becomes

$$a(t)^2(d\eta^2 + dr^2) = 0 \quad \longrightarrow \quad d\eta = \pm dr.$$

- Thus, in the η–r plane *light rays move at 45 degree angles* at all times, which simplifies discussion of horizon and causality issues.

The transformation to conformal time is a special case of a *conformal transformation*.

- A *conformal transformation* is a transformation on the metric of the form

$$g^{\mu\nu} \to f g^{\mu\nu},$$

  where f is an arbitrary spacetime function called the *conformal factor*.

- Generally, if $g^{\mu\nu}$ is a solution of the Einstein equation, $f g^{\mu\nu}$ is not, except for the trivial case where f is a constant.

- Thus, a conformal transformation is not a coordinate change; it gives a new metric.

Null geodesics are conformally invariant. Thus conformal transformations

- Give a new metric not equivalent to the original metric, but

- the new metric *preserves the lightcone structure* of the original metric.

As a result, conformal transformations are often useful in *analyzing the causal structure* of a spacetime.

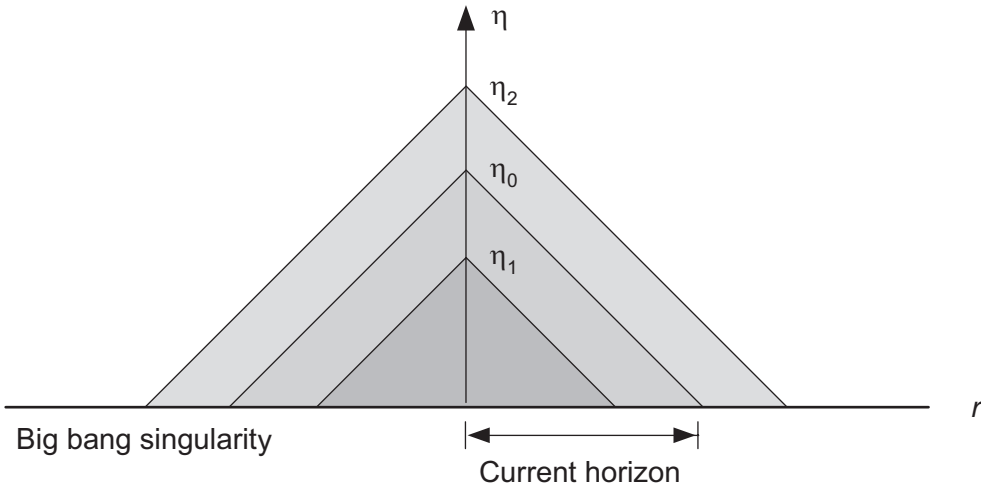


Figure 18.3: Particle horizons in conformal time.

Fig. 18.3 displays particle horizons plotted in conformal time.

- The current horizon at the present conformal time,

$$\eta_0 = \int_0^{t_0} \frac{dt}{a(t)},$$

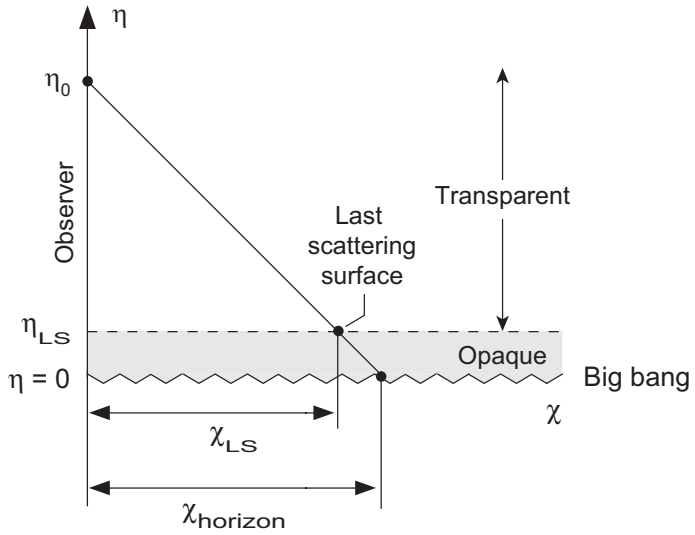
  is indicated.

- Clearly the horizon was smaller at the earlier time η_1, and will be larger at the future time η_2.

The $k = 0$ Robertson–Walker universe is related to Minkowski spacetime by a conformal transformation.

- It is called "flat" for this reason, even though its spacetime is not flat (*spacelike slices of its spacetime are flat*).

- The RW metric is said to be *conformally flat* for all k, meaning that, for all k, transformations exist that permit the metric to be cast in the Minkowski form.



The lower limit of zero on the integral for the horizon,

$$\ell_{\rm h} = a(t) \int_0^t \frac{dt'}{a(t')},$$

assumes that we can see all the way back to the big bang.

- Practically, the earliest visible time is *later* because of the *opacity of the early Universe* to various probes.
- The Universe was *opaque to photons* until about $400,000$ years after the big bang.
- Thus photons from earlier times are *not directly observable*, even if light has had time to reach us.
- The earliest photons that can be seen are from the *last scattering surface* illustrated in the figure above.

> *Neutrinos* and *gravitational waves* could be seen from earlier times, as will be discussed later.

18.8.4 Event Horizons

An event horizon is the most distant present event from which a world line could *ever* reach our world line.

- The *comoving distance* r_e to an event horizon can be defined through the integral

$$\int_{t_0}^{t_{max}} \frac{dt'}{a(t')} = \int_0^{r_e} \frac{dr}{\sqrt{1-kr^2}}.$$

- Then the *proper distance* to the event horizon is

$$\ell_e(t) = a(t) \int_0^{r_e} \frac{dr}{\sqrt{1-kr^2}} = a(t) \int_{t_0}^{t_{max}} \frac{dt'}{a(t')},$$

  which differs from the expression for the particle horizon ℓ_h only in the *limits of the integral*.

- As for particle horizons, whether an event horizon exists depends on the behavior of the integral on the right side of

$$\ell_e(t) = a(t) \int_{t_0}^{t_{max}} \frac{dt'}{a(t')},$$

If this integral *converges as* $t_{max} \to \infty$ (which depends on the detailed behavior of $a(t)$ in this limit), an event horizon exists.

Some metrics have event horizons and some don't. The expanding balloon analogy illustrates this qualitatively.

- Suppose inhabitants of galaxies on the surface of the balloon can exchange *signals of constant local speed*.
- Physical distance between galaxies increases with time, so exchanged signals must cover a *greater distance* in going from one galaxy to the next than in static spacetime.
- If space expands fast enough, the distance to a galaxy
  - may increase so rapidly that
  - the signal will *never reach that galaxy*.
- Then a spherical radius may be imagined centered on each galaxy that divides other galaxies into two groups:
  1. Those that have
     - *already been reached* by a signal sent from the galaxy, or
     - *will be reached* by the signal at some point in the finite future.
  2. Those galaxies that will *never be reached* by the signal, even after infinite time has elapsed.

If this radius exists, it defines an *event horizon* for the observer, for by symmetry no signals from beyond this radius will ever reach the observer.

If such event horizons exist, they have some similarity to event horizons associated with black holes.

- One important difference is that cosmological event horizons are *defined relative to an observer*.
- Each observer in a universe containing event horizons has her own event horizon.
- The event horizon associated with a Schwarzschild or Kerr black hole, on the other hand, is associated with a particular region of spacetime.

Particle horizons and event horizons are defined by the *same integrals*, but with *different limits*.

- A particle horizon represents the largest distance from which light could have reached us *today,* if it had traveled since the beginning of time.
- An event horizon is the largest distance from which light *emitted today* could reach us at *any future time*.
- Thus particle and event horizons are distinct:
  - Cosmological event horizons, as for black hole event horizons, separate regions of spacetime according to the *causal properties* of the spacetime.
  - Particle horizons separate spacetime events according to whether objects in the spacetime can be seen by a particular observer at a particular time and place.
- Hence the meaning of particle horizons is similar to our normal meaning of horizons on the Earth.
- Horizons on the Earth also illustrate clearly the dependence on location of the observer.

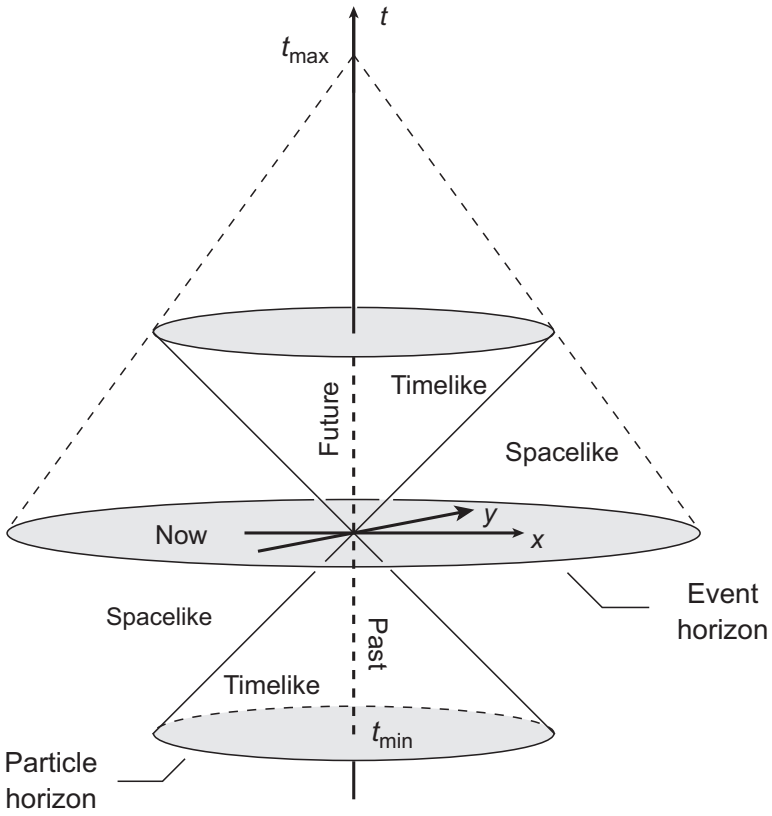


Figure 18.4: Schematic representation of particle and event horizons in cosmology. The maximum possible time coordinate is t_{max}.

The relationship between particle and event horizons is illustrated in Fig. 18.4. From this we see that an event horizon, if it exists, may be interpreted as the ultimate particle horizon.

18.9 The Einstein Equations for the RW Metric

Let's now solve the *Einstein equations* that may be relevant for a description of our Universe. A simple set of such solutions

- Assumes the content of the Universe to be a *perfect fluid* characterized by
  - an *energy density* ε and
  - a *pressure* P.
- Assumes this perfect fluid to be distributed
  - *homogeneously* and
  - *isotropically*,

so that the *Robertson–Walker metric* is valid.

> The corresponding solutions of the Einstein equations are termed *Friedmann Cosmologies*.

18.9.1 Metric and Stress–Energy Tensor

The nonvanishing covariant metric tensor components are

$$g_{00} = -1 \quad g_{11} = \frac{a^2}{1-kr^2} \quad g_{22} = a^2 r^2 \quad g_{33} = a^2 r^2 \sin^2\theta,$$

and corresponding contravariant components are

$$g^{00} = -1 \quad g^{11} = \frac{1-kr^2}{a^2} \quad g^{22} = \frac{1}{a^2 r^2} \quad g^{33} = \frac{1}{a^2 r^2 \sin^2\theta}.$$

For a perfect fluid the most general form of the stress–energy tensor is

$$T_{\mu\nu} = (\varepsilon + P) u_\mu u_\nu + P g_{\mu\nu},$$

and for a comoving observer the fluid elements are at rest, so

$$u^\mu = (1, 0, 0, 0)$$

Thus from the 4-velocity and the metric, the non-vanishing covariant components of the stress–energy tensor are

$$T_{00} = \varepsilon \quad T_{11} = \frac{Pa^2}{1-kr^2} \quad T_{22} = Pr^2 a^2 \quad T_{33} = Pr^2 a^2 \sin^2\theta.$$

18.9.2 The Connection Coefficients

The required connection coefficients are given by inserting the metric tensor components

$$g_{00} = -1 \quad g_{11} = \frac{a^2}{1-kr^2} \quad g_{22} = a^2 r^2 \quad g_{33} = a^2 r^2 \sin^2\theta$$

$$g^{00} = -1 \quad g^{11} = \frac{1-kr^2}{a^2} \quad g^{22} = \frac{1}{a^2 r^2} \quad g^{33} = \frac{1}{a^2 r^2 \sin^2\theta}$$

into

$$\Gamma^{\sigma}_{\lambda\mu} = \tfrac{1}{2} g^{\nu\sigma}\left(\frac{\partial g_{\mu\nu}}{\partial x^{\lambda}} + \frac{\partial g_{\lambda\nu}}{\partial x^{\mu}} - \frac{\partial g_{\mu\lambda}}{\partial x^{\nu}}\right)$$

For example

$$\begin{aligned}
\Gamma^{2}_{12} = \tfrac{1}{2} g^{02}\left(\frac{\partial g_{20}}{\partial x^1} + \frac{\partial g_{10}}{\partial x^2} - \frac{\partial g_{21}}{\partial x^0}\right) &\qquad (\nu = 0 \text{ term}) \\
+\tfrac{1}{2} g^{12}\left(\frac{\partial g_{21}}{\partial x^1} + \frac{\partial g_{11}}{\partial x^2} - \frac{\partial g_{21}}{\partial x^1}\right) &\qquad (\nu = 1 \text{ term}) \\
+\tfrac{1}{2} g^{22}\left(\frac{\partial g_{22}}{\partial x^1} + \frac{\partial g_{12}}{\partial x^2} - \frac{\partial g_{21}}{\partial x^2}\right) &\qquad (\nu = 2 \text{ term}) \\
+\tfrac{1}{2} g^{32}\left(\frac{\partial g_{23}}{\partial x^1} + \frac{\partial g_{13}}{\partial x^2} - \frac{\partial g_{21}}{\partial x^3}\right) &\qquad (\nu = 3 \text{ term}) \\
= \tfrac{1}{2} g^{22}\left(\frac{\partial g_{22}}{\partial x^1} + \frac{\partial g_{12}}{\partial x^2} + \frac{\partial g_{21}}{\partial x^2}\right) = \tfrac{1}{2} g^{22}\left(\frac{\partial g_{22}}{\partial x^1}\right) & \\
= \frac{1}{2}\left(\frac{-1}{r^2 a^2}\right)\frac{\partial}{\partial r}\left(-r^2 a^2\right) = \frac{1}{r}. &
\end{aligned}$$

Table 18.1: Non-vanishing Friedmann connection coefficients

$$\Gamma^0_{11} = a\dot{a}/(1-kr^2) \qquad \Gamma^0_{22} = r^2 a\dot{a} \qquad \Gamma^0_{33} = r^2\sin^2\theta a\dot{a}$$

$$\Gamma^1_{01} = \dot{a}/a \qquad \Gamma^1_{11} = kr/(1-kr^2) \qquad \Gamma^1_{22} = -r(1-kr^2)$$

$$\Gamma^1_{33} = -r(1-kr^2)\sin^2\theta \qquad \Gamma^2_{02} = \dot{a}/a \qquad \Gamma^2_{12} = 1/r$$

$$\Gamma^2_{33} = -\sin\theta\cos\theta \qquad \Gamma^3_{03} = \dot{a}/a \qquad \Gamma^3_{13} = 1/r \qquad \Gamma^3_{23} = \cot\theta$$

The coefficients are symmetric in the lower indices: $\Gamma^\mu_{\alpha\beta} = \Gamma^\mu_{\beta\alpha}$

The nonvanishing connections coefficients are summarized in Table 18.1.

18.9.3 The Ricci Tensor and Ricci Scalar

The *Ricci tensor* may now be constructed from the connection coefficients

$$R_{\mu\nu} = \Gamma^{\lambda}_{\mu\nu,\lambda} - \Gamma^{\lambda}_{\mu\lambda,\nu} + \Gamma^{\lambda}_{\mu\nu}\Gamma^{\sigma}_{\lambda\sigma} - \Gamma^{\sigma}_{\mu\lambda}\Gamma^{\lambda}_{\nu\sigma}.$$

Utilizing the connection coefficients from Table 18.1, the non-vanishing components of the Ricci tensor are

$$R_{00} = -\frac{3\ddot{a}}{a} \qquad R_{11} = \frac{a\ddot{a} + 2\dot{a}^2 + 2k}{1 - kr^2}$$

$$R_{22} = r^2(a\ddot{a} + 2\dot{a}^2 + 2k) \qquad R_{33} = R_{22}\sin^2\theta,$$

and the *Ricci scalar* is obtained by contraction with the metric tensor,

$$R = g^{\mu\nu}R_{\mu\nu} = \frac{-6(a\ddot{a} + \dot{a}^2 + k)}{a^2}.$$

18.9.4 The Friedmann Equations

We now have the necessary ingredients to construct the Einstein equations

$$R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R = 8\pi G T_{\mu\nu}.$$

From the 00 and 11 components, utilizing the previous results for $R_{\mu\nu}$, R, $g_{\mu\nu}$ and $T_{\mu\nu}$,

$$H^2 \equiv \left(\frac{\dot{a}}{a}\right)^2 = \left(\frac{8\pi G}{3}\right)\varepsilon - \frac{k}{a^2}$$

$$\frac{2\ddot{a}}{a} + \frac{\dot{a}^2}{a^2} + \frac{k}{a^2} = -8\pi G P,$$

where $H = \dot{a}/a$ has been used.

- These are termed the *Friedmann equations*.
- They represent the solution of the covariant gravitational equations with the conditions that we have imposed.

(The 22 and 33 components don't give any new results).

18.9.5 Static Solutions and the Cosmological Term

Let's first ask if the Friedmann equations have a *static solution* (solution with a constant scale factor).

- Setting $\ddot{a} = \dot{a} = 0$, the Friedmann equations become

$$3k = 8\pi G\varepsilon a^2 \qquad k = -8\pi G P a^2,$$

from which we find that

$$\frac{k}{a^2} = \frac{8\pi G}{3}\varepsilon = -8\pi G P.$$

- But from this we conclude that

1. For the present energy density ε to be positive, we must have $k = +1$ *(positive curvature)*.
2. If $\varepsilon > 0$, the *pressure must be negative*, $P < 0$!

- Thus, we find that the Friedmann universe *cannot be static*: it is unstable against either expansion or contraction.
- When Einstein first realized this, Hubble had not yet discovered the expansion of the Universe and
- the natural assumption was that a correct cosmology should give a static solution.

Thus, Einstein was led to make what he reportedly thought was the *greatest mistake of his career*.

Einstein modified the field equations by

- adding a term $\Lambda g_{\mu\nu}$ to the left side,

$$R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R + \Lambda g_{\mu\nu} = 8\pi G T_{\mu\nu}.$$

  where the constant Λ is called the *cosmological constant*.

- Unlike the other terms on the left side, the Λ term does not vanish in the limit of vanishing mass and curvature.

- The cosmological term $\Lambda g_{\mu\nu}$

  - is a *rank-2 tensor* since Λ is a scalar, and
  - it has *vanishing covariant divergence* since

$$\nabla_\alpha g_{\mu\nu} = g_{\mu\nu;\alpha} = 0,$$

  Thus it satisfies all the properties that we expect for a term in the Einstein equations.

In the modified Friedmann equations

- Positive Λ gives a *repulsion that counters gravity* and
- Negative Λ gives an *attraction that adds to gravity*.

> By proper adjustment of Λ, it is then possible to obtain a *static Friedmann universe* (for a time; the solution was shown later to be unstable).

When Hubble discovered the expansion of the Universe, Einstein realized the opportunity that had been missed.

- Had he more confidence in his original field equations, he could have *predicted* that the Universe had to be either expanding or contracting.
- Once Hubble demonstrated that the Universe was expanding, Einstein discarded the cosmological term.
- But in modern cosmology there may still be a need for the cosmological term, although for reasons very different from Einstein's original motivation.
- Because the data require Λ to be very small if it exists, it can play a role only over volumes of space that are cosmological in dimension; this is why it is commonly termed the *cosmological constant*.
- Because it is equivalent to an energy density associated with the ground state of the Universe, Λ is also often termed the *vacuum energy density*.

Because of the interpretation of Λ as a *vacuum energy density*, it is convenient in modern applications to absorb the effect of the cosmological constant on the left side of

$$R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R + \Lambda g_{\mu\nu} = 8\pi G T_{\mu\nu}.$$

into the terms arising from the stress–energy tensor on the right side.

Therefore, in the following development we shall

- Omit the explicit Λ terms from the left sides of the Friedmann equations and
- Include the possibility of a finite vacuum energy by a redefinition of the stress energy tensor $T_{\mu\nu}$ on the right side.

18.10 Resolution of Difficulties with Newtonian View

Finally, let us comment that the covariant theory of gravity implies conceptual differences relative to Newtonian cosmology.

1. Expanding space in general relativity alleviates inconsistencies associated with apparent recessional velocities that would exceed the speed of light at large distances.
2. Since the "recessional velocities" are generated by the expansion of space itself, not by motion within space, there is no conceptual difficulty with recessional velocities larger than light velocity.
3. Because of the finite speed of (massless) gravitons implied by Lorentz invariance, gravity is no longer felt instantaneously at a distance. It propagates at the speed of light.
4. In general relativity spacetime is generated by matter (and energy and pressure), so the idea of a boundary between a the universe and an empty space "outside" does not arise.
5. We shall see that the cosmological constant term suggests a way to deal with dark energy.

Therefore, general relativity solves, at least in principle, several conceptual difficulties with the Newtonian approach to cosmology, in addition to providing a more solid quantitative basis.

Chapter 19

Evolution of the Universe

In the preceding Chapter we demonstrated that the Friedmann equations correspond to the Einstein equations when the metric takes the Robertson–Walker form. In this chapter we consider solutions of the Friedmann equations and the history of the Universe that those solutions imply.

19.1 Friedmann Cosmologies

Let us return to an examination of the Friedmann equations and the cosmology that they imply.

- The *Friedmann equations*

$$H^2 \equiv \left(\frac{\dot{a}}{a}\right)^2 = \left(\frac{8\pi G}{3}\right)\varepsilon - \frac{k}{a^2}$$
$$\frac{2\ddot{a}}{a} + \frac{\dot{a}^2}{a^2} + \frac{k}{a^2} = -8\pi G P.$$

  describe the evolution of *perfect-fluid* universes having *Robertson–Walker metrics*,

- when they are supplemented by an appropriate *equation of state*.

However, it is convenient to express these equations in a *different form* before proceeding.

19.1.1 The Friedmann Equations

Consider the Friedmann equations

$$H^2 \equiv \left(\frac{\dot{a}}{a}\right)^2 = \left(\frac{8\pi G}{3}\right)\varepsilon - \frac{k}{a^2}$$
$$\frac{2\ddot{a}}{a} + \frac{\dot{a}^2}{a^2} + \frac{k}{a^2} = -8\pi G P.$$

where a cosmological term is omitted because we shall include it by a modification of the density.

- From the first of these equations we may write

$$\dot{a}^2 + k = \frac{8\pi G}{3}\varepsilon a^2,$$

and differentiating with respect to time gives

$$2\dot{a}\ddot{a} = \frac{8\pi G}{3}\left(2\varepsilon a\dot{a} + \dot{\varepsilon}a^2\right).$$

- Subtracting the Friedmann equations and solving for $\ddot{a}$,

$$2\ddot{a} = -\frac{8\pi G}{3}a(\varepsilon + 3P),$$

and using this to eliminate $\ddot{a}$ from

$$2\dot{a}\ddot{a} = \frac{8\pi G}{3}\left(2\varepsilon a\dot{a} + \dot{\varepsilon}a^2\right)$$

leads to

$$\dot{\varepsilon} + 3(\varepsilon + P)\frac{\dot{a}}{a} = 0 \quad \longleftrightarrow \quad \frac{\dot{\rho}}{\rho} + 3\left(1 + \frac{P}{\rho}\right)\frac{\dot{a}}{a} = 0$$

where $\varepsilon = \rho c^2 = \rho$ in $c = 1$ units.

The result

$$\dot{\varepsilon} + 3(\varepsilon + P)\frac{\dot{a}}{a} = 0 \quad \longleftrightarrow \quad \frac{\dot{\rho}}{\rho} + 3\left(1 + \frac{P}{\rho}\right)\frac{\dot{a}}{a} = 0$$

is in fact

- a *continuity equation* for the *conservation of mass–energy* in the cosmic fluid,
- since it can also be *derived from the requirement*

$$T^{\mu\nu}_{;\nu} = 0,$$

as shown in a Problem.

The equation

$$\dot{\varepsilon} + 3(\varepsilon + P)\frac{\dot{a}}{a} = 0$$

or its equivalent form

$$\frac{\dot{\rho}}{\rho} + 3\left(1 + \frac{P}{\rho}\right)\frac{\dot{a}}{a} = 0$$

is called the *fluid equation*.

Thus, we may solve for *evolution of the Universe* either by

- solving the two original Friedmann equations

$$\left(\frac{\dot{a}}{a}\right)^2 = \left(\frac{8\pi G}{3}\right)\varepsilon - \frac{k}{a^2} \quad \text{and} \quad \frac{2\ddot{a}}{a} + \frac{\dot{a}^2}{a^2} + \frac{k}{a^2} = -8\pi G P$$

- or by solving the two equations

$$\underbrace{\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\varepsilon - \frac{k}{a^2}}_{\text{Friedmann equation}} \quad \text{and} \quad \underbrace{\begin{cases} \dfrac{\dot{\rho}}{\rho} + 3\left(1+\dfrac{P}{\rho}\right)\dfrac{\dot{a}}{a} = 0 \\ \text{or} \\ \dot{\varepsilon} + 3(\varepsilon + P)\dfrac{\dot{a}}{a} = 0 \end{cases}}_{\text{Fluid or continuity equation}}$$

But examination of these equations indicates that (if the integer k is fixed at one of its three possible values)

- There are *3 unknowns*: ρ (or ε), P, and the scale factor a.
- Thus, we require an *additional equation*.
- This is provided by an *equation of state*

$$P = P(\varepsilon) = P(\rho),$$

which relates the thermodynamical variables P and either ε or ρ.

Therefore, the behavior of the Friedmann universe can be determined by *simultaneous solution* of the equations

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\varepsilon - \frac{k}{a^2} \quad \text{and} \quad \begin{cases} \dfrac{\dot{\rho}}{\rho} + 3\left(1+\dfrac{P}{\rho}\right)\dfrac{\dot{a}}{a} = 0 \\ \text{or} \\ \dot{\varepsilon} + 3(\varepsilon + P)\dfrac{\dot{a}}{a} = 0 \end{cases}$$

supplemented by an *equation of state*,

$$P = P(\rho) = P(\varepsilon).$$

To proceed we must examine possible *equations of state for the Universe*.

19.1.2 Equations of State

Cosmology deals with *very dilute gases*.

- Hence we may expect the equations of state to be *linear*,

$$P = w\varepsilon,$$

  where w is a *dimensionless constant.*

- The *adiabatic sound speed* in a dilute gas is given by

$$c_s^2 = \frac{dP}{d\rho} = \frac{dP}{d\varepsilon}c^2 = wc^2.$$

  Therefore,

  - If w is not negative, $c_s = \sqrt{w}c$,
  - so we must require that $w \leq 1$ if the sound speed is not to exceed the speed of light.

Let us consider several examples of possible equations of state that satisfy the condition $w \leq 1$ and could be important in the evolution of the Universe.

Note: If $w < 0$ then c_s is imaginary and disturbances don't propagate as soundwaves but instead decay exponentially.

Example: *Nonrelativistic, Low-Density Gas*

A *low-density, nonrelativistic gas* obeys the *ideal gas law*

$$P = \frac{\rho kT}{\mu},$$

where μ is the mean mass of gas molecules.

- For a nonrelativistic gas $\varepsilon \sim \rho c^2$ and

$$P = \frac{\rho kT}{\mu} = \frac{kT}{\mu c^2}\varepsilon = w\varepsilon \qquad w \equiv \frac{kT}{\mu c^2} \ll 1.$$

- Generally, nonrelativistic gases have
  - large energy densities but
  - low pressure.

For example, $w \simeq 10^{-12}$ for *air molecules* at room temperature.

> For *non-relativistic gases* we shall assume an equation of state $P = 0$, implying that $w = 0$.

Example: *Ultrarelativistic, Low-Density Gas*

For a *low-density ultrarelativistic gas* the equation of state is the familiar

$$P \equiv w\varepsilon = \tfrac{1}{3}\varepsilon.$$

Therefore massless (photons, gravitons) or nearly massless particles (e.g., neutrinos)

- have equations of state with $w \simeq \frac{1}{3}$, and
- they *exert significant pressure* in comparison with non-relativistic gases.

Example: *Dark Energy and the Cosmological Constant*

Using the earlier result for acceleration of the scale factor,

$$\ddot{a} = -\frac{4\pi G}{3} a(\varepsilon + 3P) \quad \rightarrow \quad \ddot{a} = -\frac{4\pi G}{3} a\varepsilon(1+3w)$$

where $P = w\varepsilon$ has been used. Therefore,

- If $w > -\frac{1}{3}$ the acceleration $\ddot{a}$ is *negative* and the expansion *decelerates* with time.
- If $w < -\frac{1}{3}$ then $\ddot{a}$ is *positive* and the expansion of the Universe *accelerates*.
- Any gas with $w < -\frac{1}{3}$ is termed *dark energy*.
- We shall also refer to dark energy as *vacuum energy*, because one possible source of dark energy is *quantum-mechanical vacuum fluctuations*.

As we shall now demonstrate,

1. A *cosmological constant* Λ in the Einstein equation corresponds to a *component of the Universe with* $w = -1$.
2. This represents a form of *dark energy with* $P = -\varepsilon$.
3. This implies a gas having *negative pressure* if the energy density is positive.

The Einstein equation with cosmological constant can be written as

$$\begin{aligned} R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R &= 8\pi G T_{\mu\nu} - \Lambda g_{\mu\nu} \\ &= 8\pi G (T_{\mu\nu} - \varepsilon_\Lambda g_{\mu\nu}) \end{aligned}$$

by utilizing the definition

$$\varepsilon_\Lambda \equiv \frac{\Lambda}{8\pi G}.$$

Thus, addition of a cosmological constant Λ to the Friedmann equations is *equivalent to* adding a component with energy density ε_Λ to the fluid of the Universe.

From the *fluid equation* applied to this component,

$$\dot{\varepsilon} + 3(\varepsilon + P)\frac{\dot{a}}{a} = 0 \quad \rightarrow \quad \dot{\varepsilon}_\Lambda = -3(\varepsilon_\Lambda + P_\Lambda)\frac{\dot{a}}{a}.$$

- By hypothesis Λ is constant in time so ε_Λ is constant too, requiring that $\dot{\varepsilon}_\Lambda = 0$.
- But from the preceding equation this is possible only if $P_\Lambda = -\varepsilon_\Lambda$.

Therefore, we obtain

$$P = w\varepsilon \qquad w = -1$$

for the equation of state corresponding to a *non-zero cosmological constant*.

Although the *cosmological constant* is an *example of dark energy*, it is not the only possibility.

- Any gas having $w < -\frac{1}{3}$ will cause the Universe to accelerate, and hence qualifies as *dark energy*.
- Equations of state having
  - positive energy density but
  - negative pressure

  are not unknown.
- For example, the *tension in a stretched rubber band* corresponds to a negative pressure since work must be done to stretch the band.
- What *is unusual* is to find these properties in a dilute gas, as is required for any source of dark energy.

  > Normal gases require work to compress, not to expand!

- At this stage we have strong evidence for the presence of dark energy in the Universe, but we have *no solid evidence about its source*.

19.2 Evolution and Scaling of Density Components

The preceding discussion of the equation of state permits a compact formulation of the Friedmann equations.

- Assuming a set of *independent components having density parameters* Ω_i, the Friedmann equation

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{k}{a^2}$$

may be expressed for a scale parameter $a \equiv a(t)$ as

$$\frac{\dot{a}^2}{a^2} = H_0^2 \sum_i \Omega_i \left(\frac{a_0}{a}\right)^{3(1+w_i)} - \frac{k}{a^2},$$

where a_0 is the current value of the scale parameter and the equation of state for each component is given by

$$P_i = w_i \varepsilon_i.$$

- This implies, by virtue of

$$\dot{\varepsilon} + 3(\varepsilon + P)\frac{\dot{a}}{a} = 0,$$

that the *energy density scales as*

$$\varepsilon_i = \varepsilon_i(0)\left(\frac{a}{a_0}\right)^{-3(1+w_i)},$$

where $\varepsilon_i(0)$ is the energy density of component i today, $a = a(t)$ is the scale factor at time t, and a_0 is the scale factor today.

We may give a physical interpretation of the scaling of densities for different components implied by

$$\varepsilon_i = \varepsilon_i(0)\left(\frac{a}{a_0}\right)^{-3(1+w_i)},$$

- For *radiation* (ultrarelativistic particles like photons having negligible mass), $w = \frac{1}{3}$ and energy density scales as

$$\varepsilon_{\rm r} \simeq \frac{h\nu}{V} = \frac{hc}{V\lambda} \simeq a^{-4},$$

because the

  - volume V scales as a^3 and
  - the wavelength λ scales as a.

- For *nonrelativistic matter* $w = 0$ and density scales as

$$\varepsilon_{\rm m} = \rho c^2 = \frac{mc^2}{V} \simeq a^{-3}.$$

- The difference in scaling for radiation dominated and matter dominated universes is ultimately the *additional length-scale factor associated with the redshift of photon wavelengths* caused by the expansion.
- The vacuum energy ($w = -1$ in the simplest models) is associated with "empty" spacetime itself.
- The *vacuum energy density* is not changed as new space is created in the expansion of the Universe.
- Hence a $w = -1$ component has a *density independent of the scale factor.*

19.2.1 A Standard Model

If we assume the Universe to be composed of

1. A *radiation component* with $w = \frac{1}{3}$,
2. A *matter component* with $w = 0$,
3. A *vacuum (dark energy) component* with $w = -1$,

then the equation

$$\frac{\dot{a}^2}{a^2} = H_0^2 \sum_i \Omega_i \left(\frac{a_0}{a}\right)^{3(1+w_i)} - \frac{k}{a^2},$$

reduces to

$$\frac{\dot{a}^2}{a^2} = H_0^2 \left(\underbrace{\Omega_{\rm r} \left(\frac{a_0}{a}\right)^4}_{\text{radiation}} + \underbrace{\Omega_{\rm m} \left(\frac{a_0}{a}\right)^3}_{\text{matter}} + \underbrace{\Omega_\Lambda}_{\text{vacuum}} + \underbrace{\Omega_k}_{\text{curvature}} \right),$$

where we have introduced formally a *curvature density parameter* Ω_k through

$$\Omega_k \equiv \frac{-k}{a^2 H_0^2} \qquad k = \{-1, 0, +1\}.$$

Note though that Ω_k can be *negative* (if $k = +1$), unlike the other Ω_i, which are never negative.

If we evaluate

$$\frac{\dot{a}^2}{a^2} = H_0^2 \left(\Omega_{\rm r} \left(\frac{a_0}{a} \right)^4 + \Omega_{\rm m} \left(\frac{a_0}{a} \right)^3 + \Omega_\Lambda + \Omega_k \right),$$

at the present time

$$a \to a_0 \qquad \dot{a}^2/a^2 \to \dot{a}_0^2/a_0^2 = H_0^2$$

we see that the density parameters are constrained by

$$\Omega_{\rm r} + \Omega_{\rm m} + \Omega_\Lambda + \Omega_k = 1$$

Recall that the density parameters Ω_i are the ratios of the component densities to the critical density *evaluated at the present time.*

Defining a *total density parameter* Ω through

$$\Omega \equiv \Omega_{\rm r} + \Omega_{\rm m} + \Omega_\Lambda,$$

we have that

$$\Omega_{\rm r} + \Omega_{\rm m} + \Omega_\Lambda + \Omega_k = 1 \quad \to \quad \Omega = 1 - \Omega_k.$$

Thus $\Omega = 1$ for a flat universe ($\Omega_k = 0$), irrespective of the values of the individual components $\Omega_{\rm r}$, $\Omega_{\rm m}$, and Ω_Λ.

The preceding equations represent a solution for the evolution of the Universe in terms of *four independent parameters:*

1. The *current value* of the Hubble parameter H_0
2. The *current value* of the radiation energy density parameter Ω_{r}
3. The *current value* of the matter energy density parameter Ω_{m}
4. The *current value* of the vacuum energy density parameter Ω_{Λ}.

The *curvature density parameter* Ω_k is then defined by

$$\Omega_k \equiv \frac{-k}{a^2 H_0^2},$$

given a choice of $k = \{-1, 0, +1\}$.

19.3 Flat, Single-Component Universes

The differential equation governing the cosmological evolution of the Universe,

$$\frac{\dot{a}^2}{a^2} = H_0^2 \sum_i \Omega_i \left(\frac{a_0}{a}\right)^{3(1+w_i)} - \frac{k}{a^2},$$

generally must be *solved numerically*. However, if

- *curvature is neglected* and
- the summation is restricted to a *single component* having an equation of state parameter w_i,

it is possible to find *analytical solutions*.

- These solutions provide insight into how various components contribute to evolution of the Universe.
- However, they are also of considerable practical importance because evidence indicates that
  1. The Universe is *very flat* and was *even flatter at earlier times*.
  2. At various stages in its evolution the Universe has been *dominated by a single component*.

So let's investigate solutions of the Friedmann equations for *flat universes having a single component* described by an equation of state $P = w\rho$.

Restricting the Friedmann equation

$$\frac{\dot{a}^2}{a^2} = H_0^2 \sum_i \Omega_i \left(\frac{a_0}{a}\right)^{3(1+w_i)} - \frac{k}{a^2},$$

to a single component described by a density parameter Ω and equation of state parameter w (and setting $a_0 \equiv 1$) gives

$$\dot{a}^2 = H_0^2 \Omega a^{-(1+3w)} - k.$$

- If $1+3w$ is positive, the curvature term k becomes relatively more important as the Universe expands.
- The recent Universe has been dominated by a component having $1+3w$ negative (dark energy).
- However, the early Universe was dominated by components for which $1+3w$ was positive.
- The curvature term is observed to be *small today*, so it was *even less important in the early Universe*.
- Therefore, as a first approximation we *neglect curvature*, giving

  $$\dot{a}^2 = H_0^2 \Omega a^{-(1+3w)} - k \quad \rightarrow \quad \dot{a}^2 = H_0^2 a^{-(1+3w)},$$

  where $\Omega = 1$ because we assume a flat Universe.

We now consider possible solutions to this equation for the single component defined by the equation of state parameter w.

19.3.1 Vacuum Energy Domination ($w=-1$)

The equation

$$\dot{a}^2 = H_0^2 a^{-(1+3w)},$$

has a special solution for the case $w = -1$:

$$\dot{a} = H_0 a \quad \rightarrow \quad a(t) = e^{Ht}.$$

- Because the exponential curve is self-similar, there is no preferred point to normalize the solution.
- We have chosen to fix the integration constant by requiring that $a(t) = 1$ at $t = 0$.
- More generally, one could choose a solution $\exp H(t - t_0)$, with the normalization $a(t) = 1$ at $t = t_0$.
- We have also replaced H_0 by H, because $H(t) \equiv \dot{a}/a = H_0$, so for this special solution *the Hubble parameter is constant in time*.
- As we have seen, $w = -1$ corresponds to a *pure vacuum-energy universe* (no matter or radiation).
- Such a universe with no curvature is termed a *de Sitter universe.*

At first glance, the de Sitter solution may appear to be academic since the Universe clearly does contain matter and radiation in addition to any vacuum energy. However,

- If the Universe continues to expand, all matter and radiation will eventually be sufficiently diluted that effectively only vacuum energy will remain.
- Therefore, the de Sitter solution may be viewed as a universe in which the density of matter and radiation is negligible compared with vacuum energy.
- This is the ultimate fate of any more standard cosmology that expands forever.
- Furthermore, there is substantial evidence that in its first instants our Universe underwent an inflationary period in which the Universe expanded for a short but extremely important period in a de Sitter phase.

We may use the de Sitter solution to compute quantities of cosmological interest:

- The *age* of a de Sitter universe is infinite,

$$\text{Age} = \infty,$$

  since only at $t \to -\infty$ does the scale factor $a(t)$ go to zero.

- The *energy density* in de Sitter space *is constant*,

$$\varepsilon = \varepsilon_0 a^{-3(1+w)} = \varepsilon_0.$$

- The *redshift* for light emitted at time t_e and detected at the present time $t = 0$ is

$$z = \frac{a(t=0)}{a(t_\text{e})} - 1 = e^{-Ht_\text{e}} - 1,$$

  in terms of which the *time of emission* t_e is

$$t_\text{e} = -\frac{\ln(1+z)}{H}.$$

- The *lookback time* is

$$t_\text{L} = t_0 - t_\text{e} = \frac{\ln(1+z)}{H}.$$

- The *proper distance of a light source at the time of observation* of a light signal is

$$\ell(t=0) = \int_{t_e}^{0} \frac{dt}{a(t)} = \int_{t_e}^{0} e^{-Ht} dt$$
$$= \frac{1}{H}\left(e^{-Ht_e} - 1\right) = \frac{z}{H}.$$

- The *proper distance at the time of emission* of this same light is rescaled by $a(t_e)/a(t_0) = 1/(1+z)$, so

$$\ell(t_e) = \left(\frac{1}{1+z}\right)\frac{z}{H}.$$

- The *particle horizon* ℓ_h is the same integral as for proper distance of a light source at the time of observation, but with $t_e \to -\infty$ at the lower limit; thus

$$\ell_h = \lim_{t_e \to -\infty} \ell(t_0) = \lim_{t_e \to -\infty} \left(\frac{z}{H}\right) = \infty.$$

A de Sitter spacetime is infinitely old *with* no particle horizon.

1. An observer could in principle see every point in a de Sitter universe.
2. However, distant objects would be very strongly redshifted images of their appearance at much earlier times.

- The large redshift limit for the proper distance *at the time of detection* is

$$\lim_{z\to\infty} \ell(t=0) = \lim_{z\to\infty} \frac{z}{H} = \infty.$$

- The corresponding large-redshift limit for the proper distance *at the time of light emission* t_e is

$$\lim_{z\to\infty} \ell(t_e) = \lim_{z\to\infty} \left(\frac{1}{1+z}\right) \frac{z}{H} = \frac{1}{H}.$$

- In a de Sitter universe, objects with high redshift are at very large proper distance *when observed*.
- But the light that we see was *emitted from these objects* near a proper distance of c/H *(Hubble distance)*.

1. Once the source is at greater than the Hubble distance, the expansion (of space) carries it away from the observer at greater than light speed.
2. Thus its light can no longer reach the observer.
3. This constitutes an *event horizon* for the observer.

Table 19.1: Single-component flat universe (vacuum only)

Property	Vacuum only ($\Omega = \Omega_\Lambda = 1$)
Age	∞
Scale factor $a(t)$	e^{Ht}
Time of emission $t_e(z)$	$-\dfrac{\ln(1+z)}{H_0}$
Redshift $z(t_e)$	$e^{-Ht_e} - 1$
Proper distance at detection $\ell(t=0)$	$\dfrac{z}{H}$
Proper distance at emission $\ell(t_e)$	$\left(\dfrac{1}{1+z}\right)\dfrac{z}{H}$
Lookback time t_L	$\dfrac{\ln(1+z)}{H}$
Particle horizon distance ℓ_h	∞
Energy density ε	ε_0 (constant)

The results just obtained for a

- flat universe that contains
- only vacuum energy

are summarized in Table 19.1.

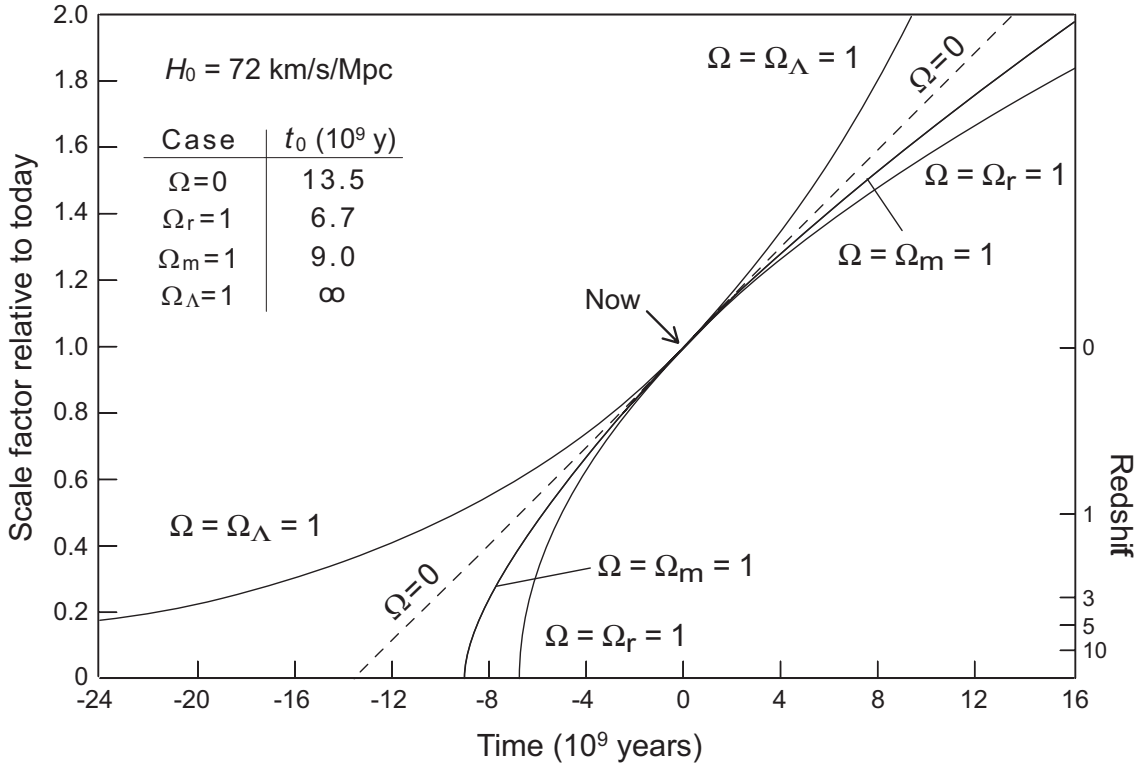


Figure 19.1: Scale factor $a(t)$ versus time for a universe obeying the Hubble law ($\Omega = 0$), and for single-component critical densities of radiation ($\Omega = \Omega_r = 1$), matter ($\Omega = \Omega_m = 1$), and vacuum energy ($\Omega = \Omega_\Lambda = 1$). The left axis gives the ratio of the scale factor to its value today and the right axis gives the redshift. The corresponding age of the Universe t_0 corresponds to the intercept with the bottom axis and is given in the inset table.

Solution of the Friedmann equations for a pure vacuum energy (de Sitter) universe are displayed in Fig. 19.1 as the curve labeled $\Omega = \Omega_\Lambda = 1$.

19.3.2 Solutions for $w \neq -1$

In the *general case where* $w \neq -1$, the form of

$$\dot{a}^2 = H_0^2 a^{-(1+3w)},$$

suggests a *power law solution* of the form

$$a(t) = \left(\frac{t}{t_0}\right)^n.$$

Substitution and comparison of exponents on the two sides indicates that

$$n = \frac{2}{3+3w},$$

while comparison of the non-exponent factors indicates that

$$t_0 = \frac{2}{3H_0(1+w)}.$$

Thus, the general solution for $w \neq -1$ is

$$\dot{a}^2 = H_0^2 a^{-(1+3w)} \quad \rightarrow \quad a(t) = \left(\frac{t}{t_0}\right)^{2/(3+3w)},$$

with the *Hubble constant* H_0 related to the *age of the universe* t_0 through

$$H_0 = \frac{2}{3t_0(1+w)}.$$

From

$$\varepsilon_i = \varepsilon_i(0)\left(\frac{a}{a_0}\right)^{-3(1+w_i)} \qquad a(t) = \left(\frac{t}{t_0}\right)^{2/(3+3w)}$$

we obtain for the *energy density*

$$\varepsilon = \varepsilon_0\left[\left(\frac{t}{t_0}\right)^{2/(3+3w)}\right]^{-3(1+w)} = \varepsilon_0\left(\frac{t}{t_0}\right)^{-2}$$
$$= \frac{t^{-2}}{6\pi(1+w)^2 G},$$

For light emitted at time t_e and detected today at time t_0, the *redshift* is

$$z = \frac{1}{a(t_e)} - 1 = \left(\frac{t_0}{t_e}\right)^{2/(3+3w)} - 1,$$

from which the *time of emission* is

$$t_e = t_0(1+z)^{-3(1+w)/2}.$$

The *current proper distance* is

$$\ell(t_0) = \int_{t_e}^{t_0} \left(\frac{t}{t_0}\right)^{-2/(3+3w)} dt$$
$$= t_0 \frac{3+3w}{1+3w}\left[1 - \left(\frac{t_e}{t_0}\right)^{(1+3w)/(3+3w)}\right],$$

where $w \neq -\frac{1}{3}$. We may use

$$t_0 = \frac{2}{3H_0(1+w)} \qquad z = -1 + \left(\frac{t_0}{t_e}\right)^{2/(3+3w)}$$

to express the current proper distance in terms of z and H_0,

$$\ell(t_0) = \frac{2}{(1+3w)H_0}\left(1 - (1+z)^{-(1+3w)/2}\right).$$

Proper distance at photon emission $\ell(t_e)$ is scaled by a factor

$$\frac{a(t_e)}{a(t_0)} = \frac{1}{(1+z)}$$

and thus

$$\ell(t_e) = \frac{1}{1+z}\ell(t_0).$$

The lookback time is

$$t_L = t_0 - t_e = t_0\left(1 - (1+z)^{-3(1+w)/2}\right),$$

which reduces to the form expected from the Hubble law,

$$t_L \simeq \tau_H z = \frac{z}{H_0} \qquad \text{(small } z\text{)},$$

only for small redshift.

The *particle horizon distance* is the same integral as for the proper distance at the current time but with the lower limit $t_e \to 0$, or equivalently, $z \to \infty$.

- Thus, *if the integral is convergent,*

$$\ell_h = \lim_{z\to\infty} \ell(t_0) = \lim_{z\to\infty} \left[\frac{2}{(1+3w)H_0} \left(1 - (1+z)^{-(1+3w)/2} \right) \right] = \frac{2}{(1+3w)H_0}.$$

- However, this integral is *not convergent for all values of* w. Generally,

$$\ell_h = \begin{cases} \dfrac{2}{(1+3w)H_0} & w > -\frac{1}{3} \\ \infty & w \le -\frac{1}{3} \end{cases}.$$

From these results, we deduce concerning horizons:

- Flat spacetime with a single fluid component has a particle horizon only if $w > -\frac{1}{3}$.
- In a flat universe with one component having $w \le -\frac{1}{3}$,
  - there is *no particle horizon* and
  - an observer can (in principle) see all points in space,
  - but more distant points will be *highly redshifted*.

- In a flat universe with a single component having an equation of state with $w > -\frac{1}{3}$,
  - the horizon distance is finite and
  - an observer sees only a portion of an infinite volume (the *visible universe.*).
- Since the horizon distance is proportional to $1/H_0 = \tau_{\rm H} \sim$ *age of the universe,* it grows with time.

> The horizon grows with time because signals propagating at light speed have had time to reach the observer from increasingly distant points as time goes on.

19.3.3 Flat Universes with Radiation or Matter

Let's apply these results to two specific cases:

- a *radiation-only* flat universe ($w = \frac{1}{3}$), and
- a *matter-only* flat universe ($w = 0$.

This requires only substitution into equations already derived.

- For example, the current age of the Universe is

$$\text{Radiation dominated}: \quad t_0 = \frac{1}{2H_0} \qquad (\Omega = 1, w = \tfrac{1}{3}).$$

$$\text{Matter dominated}: \quad t_0 = \frac{2}{3H_0} \qquad (\Omega = 1, w = 0).$$

- For a *flat radiation-dominated universe* the scale factor is

$$a(t) \simeq \text{constant} \times t^{1/2}.$$

and the *proper distance to the particle horizon* is given by

$$\ell_{\rm h} = \frac{2}{(1+3w)H_0} = \frac{1}{H_0} \qquad (\text{radiation; } w = \tfrac{1}{3}),$$

- For a *flat universe dominated by non-relativistic matter*

$$a(t) \simeq \text{constant} \times t^{2/3}.$$

and the *proper distance to the particle horizon* is

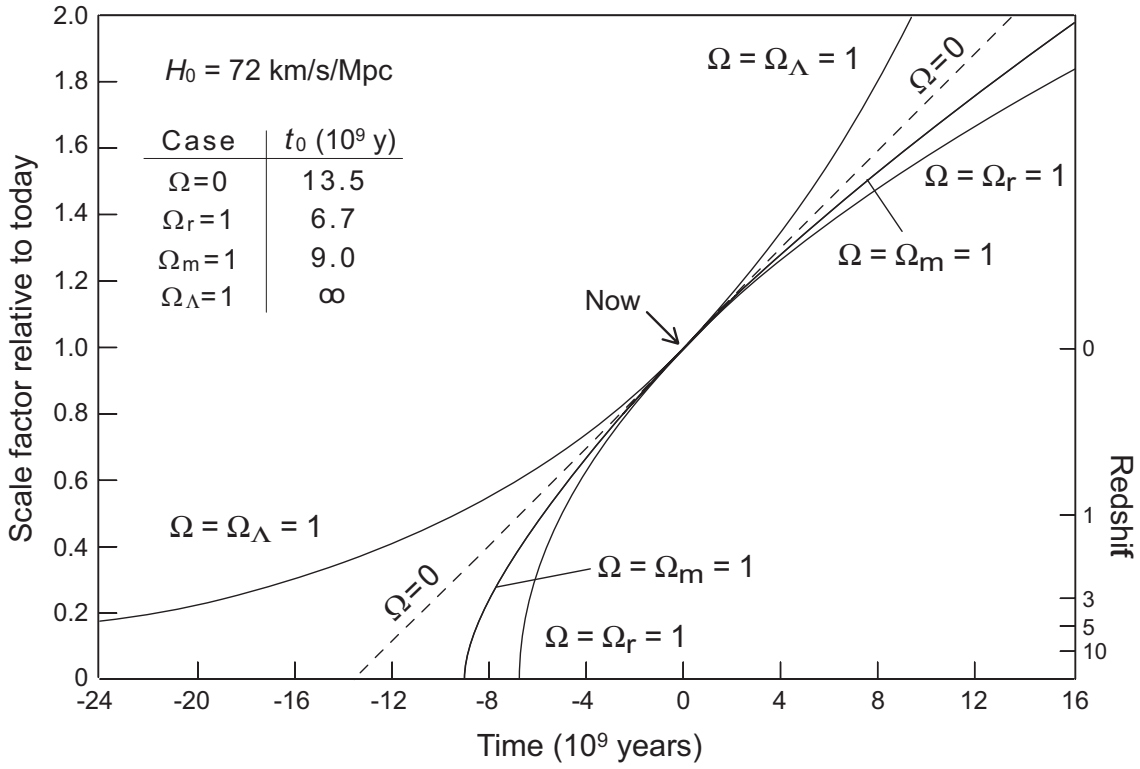
$$\ell_{\rm h} = \frac{2}{(1+3w)H_0} = \frac{2}{H_0} \qquad (\text{matter; } w = 0)$$

and so on.

Table 19.2: Cosmological quantities for single-component flat universes

Property	Vacuum only ($\Omega = \Omega_\Lambda = 1$)	Radiation only ($\Omega = \Omega_r = 1$)	Matter only ($\Omega = \Omega_m = 1$)
Age t_0	∞	$\dfrac{1}{2H_0}$	$\dfrac{2}{3H_0}$
Scale factor $a(t)$	e^{Ht}	$\left(\dfrac{t}{t_0}\right)^{1/2}$	$\left(\dfrac{t}{t_0}\right)^{2/3}$
Time of emission $t_e(z,t_0)$	$-\dfrac{\ln(1+z)}{H}$	$t_0(1+z)^{-2}$	$t_0(1+z)^{-3/2}$
Redshift $z(t_e,t_0)$	$e^{-Ht_e}-1$	$\left(\dfrac{t_0}{t_e}\right)^{1/2}-1$	$\left(\dfrac{t_0}{t_e}\right)^{2/3}-1$
Proper distance at detection $\ell(t_0)$	$\dfrac{z}{H}$	$\dfrac{1}{H_0}\left(\dfrac{z}{1+z}\right)$	$\dfrac{2}{H_0}\left[1-(1+z)^{-1/2}\right]$
Proper distance at emission $\ell(t_e)$	$\left(\dfrac{1}{1+z}\right)\dfrac{z}{H}$	$\dfrac{z}{H_0(1+z)^2}$	$\dfrac{2}{(1+z)H_0}\left[1-(1+z)^{-1/2}\right]$
Lookback time t_L	$\dfrac{\ln(1+z)}{H}$	$\dfrac{1}{2H_0}[1-(1+z)^{-2}]$	$\dfrac{2}{3H_0}\left[1-(1+z)^{-3/2}\right]$
Particle horizon distance $\ell(t_h)$	∞	$\dfrac{1}{H_0}$	$\dfrac{2}{H_0}$
Energy density $\varepsilon(t)$	ε_0 (constant)	$\dfrac{3}{32\pi G}t^{-2}$	$\dfrac{1}{6\pi G}t^{-2}$

Table 19.2 summarizes the preceding cosmological quantities computed for vacuum-only, radiation-only, and matter-only flat universes.



The figure above illustrates the scale factor versus time for a universe obeying the Hubble law (dashed line with $\Omega = 0$) and for single-component critical densities of

- radiation ($\Omega = \Omega_r = 1$),
- matter ($\Omega = \Omega_m = 1$), and
- vacuum energy ($\Omega = \Omega_\Lambda = 1$).

19.4 Full Solution of the Friedmann Equations

In the general case, corresponding to

- arbitrary values of H_0, $\Omega_{\rm r}$, $\Omega_{\rm m}$, and Ω_Λ, (and therefore possible non-zero curvature),
- it is convenient to *solve the equations numerically*.

If we introduce

- a *dimensionless measure* $q(t)$ *of scale* and a
- *dimensionless measure* τ *of time* through

$$q \equiv \frac{a}{a_0} \qquad \tau \equiv H_0 t,$$

then

$$\frac{\dot{a}^2}{a^2} = H_0^2 \left(\Omega_{\rm r} \left(\frac{a_0}{a}\right)^4 + \Omega_{\rm m} \left(\frac{a_0}{a}\right)^3 + \Omega_\Lambda + \Omega_k \right),$$

may be written as

$$\frac{1}{2}\left(\frac{dq}{d\tau}\right)^2 + U(q) = E,$$

where we have defined

$$U(q) \equiv -\frac{1}{2}\left(\frac{\Omega_{\rm r}}{q^2} + \frac{\Omega_{\rm m}}{q} + \Omega_\Lambda q^2\right) \qquad E \equiv \frac{1}{2}(1-\Omega) = \frac{1}{2}\Omega_k$$

and where the redshift z is related to the variable q through

$$z = \frac{1-q}{q}.$$

The preceding equations represent a solution for the evolution of the Universe in terms of *four independent parameters:*

1. The *current value* of the Hubble parameter H_0
2. The *current value* of the radiation energy density parameter $\Omega_{\rm r}$
3. The *current value* of the matter energy density parameter $\Omega_{\rm m}$
4. The *current value* of the vacuum energy density parameter Ω_Λ.

These parameters then fix the *current curvature density parameter* Ω_k through

$$\Omega_{\rm r} + \Omega_{\rm m} + \Omega_\Lambda + \Omega_k = 1.$$

A general solution for evolution of the Universe may be obtained through the following algorithm:

1. Specify values of the cosmic parameters H_0, $\Omega_{\rm r}$, $\Omega_{\rm m}$, and Ω_Λ, which fixes the curvature density parameter Ω_k

$$\underbrace{\Omega_{\rm r}+\Omega_{\rm m}+\Omega_\Lambda}_{\Omega}+\Omega_k = 1 \quad \rightarrow \quad \Omega_k = 1-\Omega.$$

2. Compute the age of the Universe in units of the dimensionless time τ by evaluating numerically the integral

$$\tau_0 = \int_0^1 (2(E-U(q))^{-1/2}\, dq.$$

3. Integrate numerically the differential equation

$$\frac{1}{2}\left(\frac{dq}{d\tau}\right)^2 + U(q) = E \quad \rightarrow \quad dq = (2(E-U(q))^{1/2} d\tau,$$

from the current time ($q = 1$)

- backward to the beginning and
- forward to arbitrary times,

to determine the time evolution of $q(\tau)$.

4. Convert to standard variables t and $a(t)$ using

$$q \equiv \frac{a}{a_0} \qquad \tau \equiv H_0 t,$$

and use the results to compute quantities such as redshifts, lookback times, and horizons of cosmological interest.

19.4.1 Examples: Single Component with Curvature

Figures 19.2–19.4 (following) illustrate some numerical calculations that have been carried out according to this prescription for single-component Universes, but for which *the density parameter* Ω *is not necessarily equal to one*.

- In general, this implies that these spaces have *non-zero curvature* if $\Omega \neq 1$.
- The numerical integrations have been done using a *standard Runge–Kutta algorithm*.
- Each of these plots shows the variation of the scale factor as a function of time, with the current value normalized to one.
- The time axis has been shifted such that it measures the number of billions of years relative to today.
- In all cases, a Hubble constant of $72\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ has been assumed.
- The age of the Universe corresponds to the intercept of the scale factor curve with the bottom axis (that is, $a(t) \to 0$).
- In each case, a corresponding redshift is shown on the right axis.

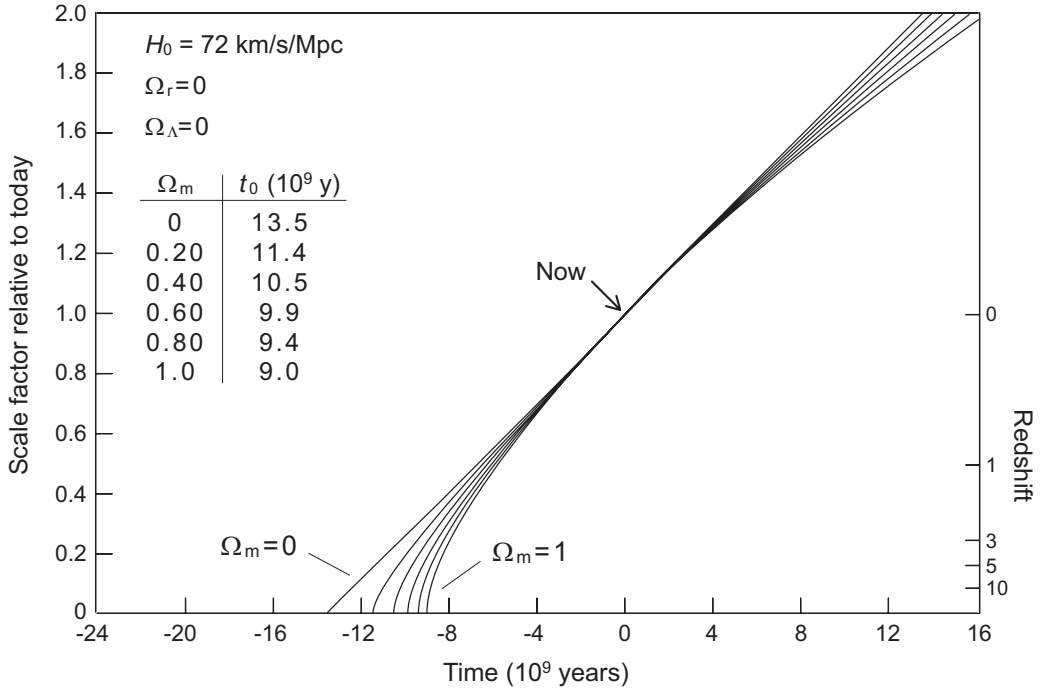


Figure 19.2: Scale factor vs. time for a universe with no vacuum or radiation energy density, and matter density Ω_m with values of 0, 0.2, 0.4, 0.6, 0.8, and 1. Left axis: ratio of scale factor to its value today; Right axis: redshift. The corresponding age of the Universe is given in the inset table.

In Fig. 19.2, evolution of the scale factor for a *matter-only universe* with densities from 0% to 100% of critical is shown.

- All cases agree in at the current time ("Now").
- Age of the Universe ranges from 13.5 to 9.0 billion years.
- For monotonic curves redshift, time, or scale factor may be used as time parameters.
- The case $\Omega_m = 1$ corresponds to a closed, flat universe. All others are open, curved universes.

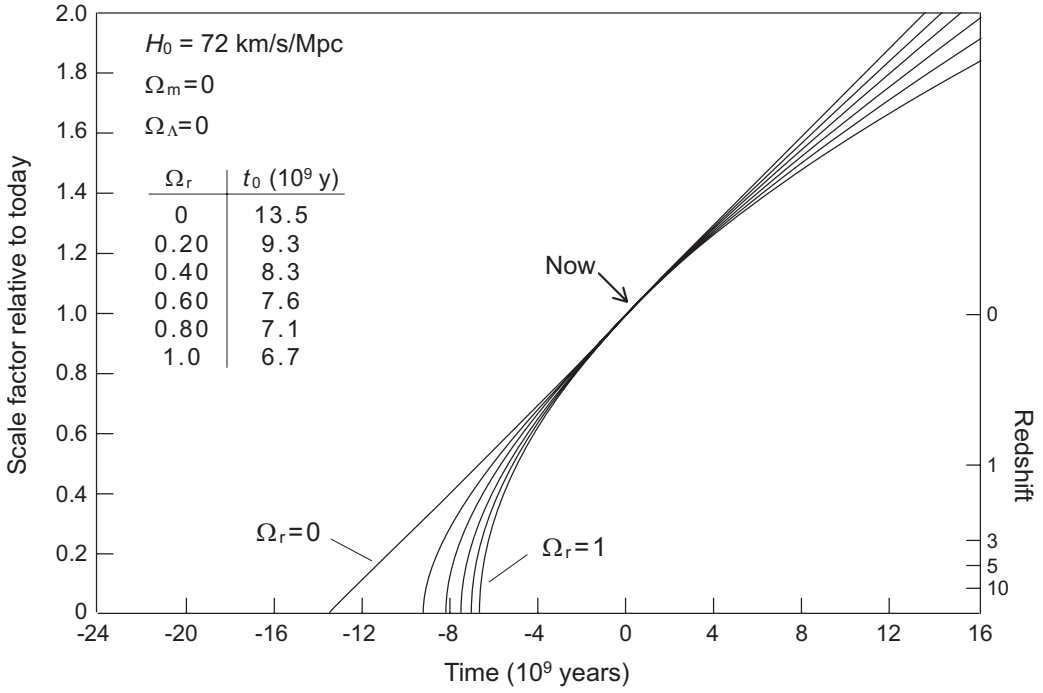


Figure 19.3: Scale factor for a universe with no vacuum or matter energy density and radiation density Ω_r with values of 0, 0.2, 0.4, 0.6, 0.8, and 1.

In Fig. 19.3, we show an example similar to that of Fig. 19.2, except that now the *single component is radiation*.

- Again $\Omega_r = 1$ is a closed, flat universe and the other curves correspond to open, curved universes.
- The predicted lifetimes vary over an even larger range than in the matter-only case, from 6.7 billion years for the $\Omega_r = 1$ case, up to 13.5 billion years for the $\Omega_r = 0$ case.
- Notice that in both Fig. 19.2 and Fig. 19.3, increasing the density causes the age of the universe to decrease relative to the Hubble time of 13.5 billion years (intercept of the $\Omega_r = 0$ curve with the bottom axis).

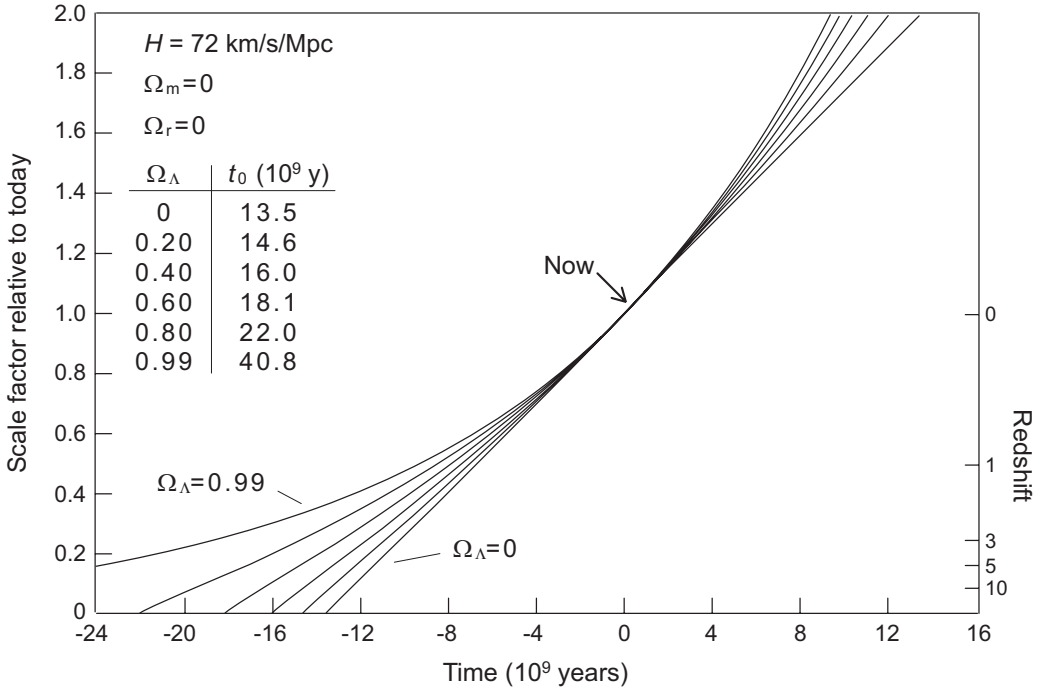


Figure 19.4: Scale factor versus time for a universe with no radiation or matter energy density and vacuum energy density Ω_v with values of 0, 0.2, 0.4, 0.6, 0.8, and 0.99.

In Fig. 19.4, we show an example similar to Figs. 19.2 and 19.3, but now for varying amounts of *pure vacuum energy*.

- Vacuum energy gives opposite curvature than that produced by matter or radiation, so the lifetimes for a pure vacuum energy universe are longer than the Hubble time.
- For these examples the lifetime ranges from 13.5 billion years for $\Omega_v = 0$ to 40.8 billion years for $\Omega_v = 0.99$.
- For $\Omega_v = 1$ the universe has no initial singularity and *an infinite lifetime*.

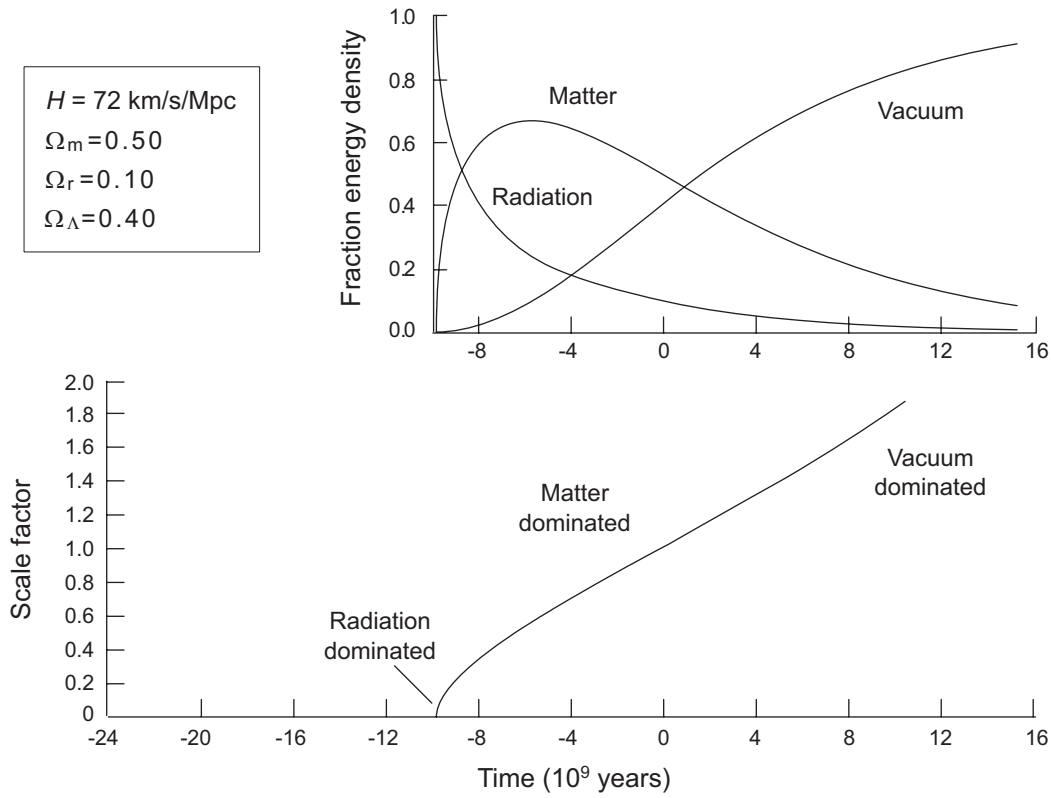
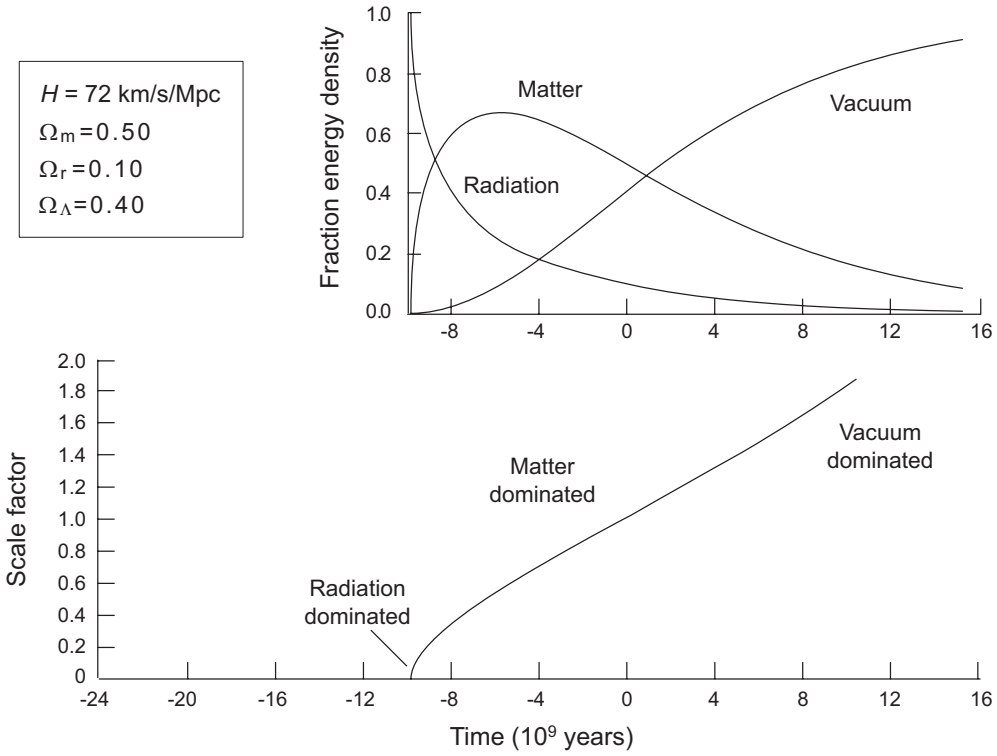


Figure 19.5: Scale factor versus time for a flat universe with 50% matter, 40% vacuum energy density, and 10% radiation. Energy densities of matter, radiation, and vacuum energy as a function of time are indicated in the inset.

19.4.2 Examples: Multiple Components

Solutions of the Friedmann equations with more than one component reflect a superposition of the behaviors illustrated in earlier figures.

- In Fig. 19.5 we show a representative calculation having $\Omega_m = 0.5$, $\Omega_r = 0.10$, and $\Omega_v = 0.40$.
- This is a *closed universe* since $\Omega = \Omega_m + \Omega_r + \Omega_v = 1$.

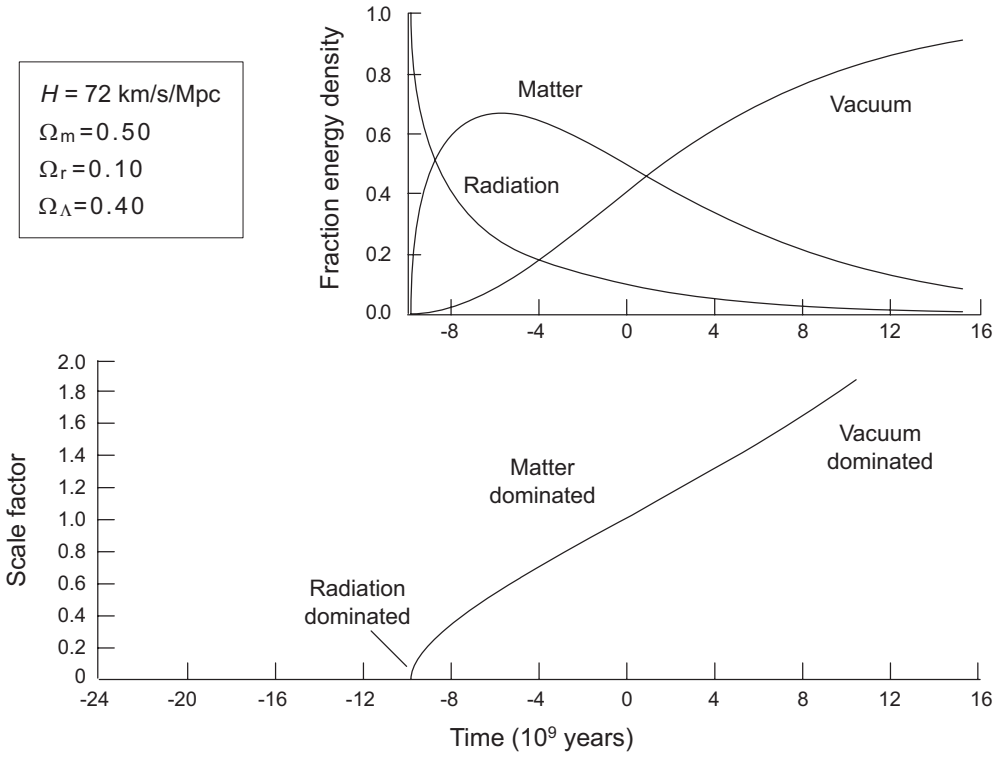


By comparing the figure above with Figs. 19.2–19.4, we can identify parts of the curve dominated by different components.

- Radiation density scales as $a(t)^{-4}$ and falls off most rapidly, followed by matter density scaling as $a(t)^{-3}$.
- The vacuum energy density *is unchanged by expansion* and it makes increasingly larger relative contribution.

Therefore, the scale-factor curve is

- initially dominated by *radiation density*,
- then by *matter density*, and finally by
- *vacuum energy density* at large times.



The relative contribution of each component (total normalized to one) is clear in the inset to the above figure.

- Radiation dominates until about 8 billion years ago.
- Then matter becomes dominant from about 8 billion years ago until the present.
- In another 1-2 billion years vacuum energy will overtake the contribution of matter and increasingly dominate.
- Ultimately this universe will become an
  - exponentially expanding,
  - spatially flat

  de Sitter space, with *negligible matter and radiation*.

19.4.3 Parameters for a Realistic Model

The preceding discussion has introduced the relative contributions of radiation, matter, and vacuum energy to the evolution of the Universe.

- However, these have been toy models that have explored ranges of cosmological parameters.
- What about the real Universe?
- Although the evolution of the Universe is certainly more complicated than the 4-parameter model that we are discussing,
- we have reason to believe that these four parameters are the most critical ones in determining the overall behavior of the Universe.
- Let us assume that to be the case.

Then, we need *observational information* to fix the *current values of*

1. The *Hubble constant* H_0.
2. The *radiation density parameter* $\Omega_{\rm r}$.
3. The *matter density parameter* $\Omega_{\rm m}$.
4. The *vacuum energy density parameter* Ω_Λ.

Hubble Constant: Let's consider the Hubble constant first.

- Different methods of determining H_0 disagree at the $5-10\%$ level among themselves (which is often outside the estimated error limits on individual determinations),
- but generally agree that it lies in the range $67-73\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$.
- Figure 19.6 (next page) illustrates in its top part one determination of the Hubble parameter.
- We will assume the Hubble constant to have the value $H_0 = 72\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ for the representative calculations in this chapter.

In some later calculations we will use somewhat smaller values of H_0. It is unclear at the time of this writing whether the $67-73\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ range for determination of H_0 by different methods indicates a real inconsistency, or an incomplete error analysis.

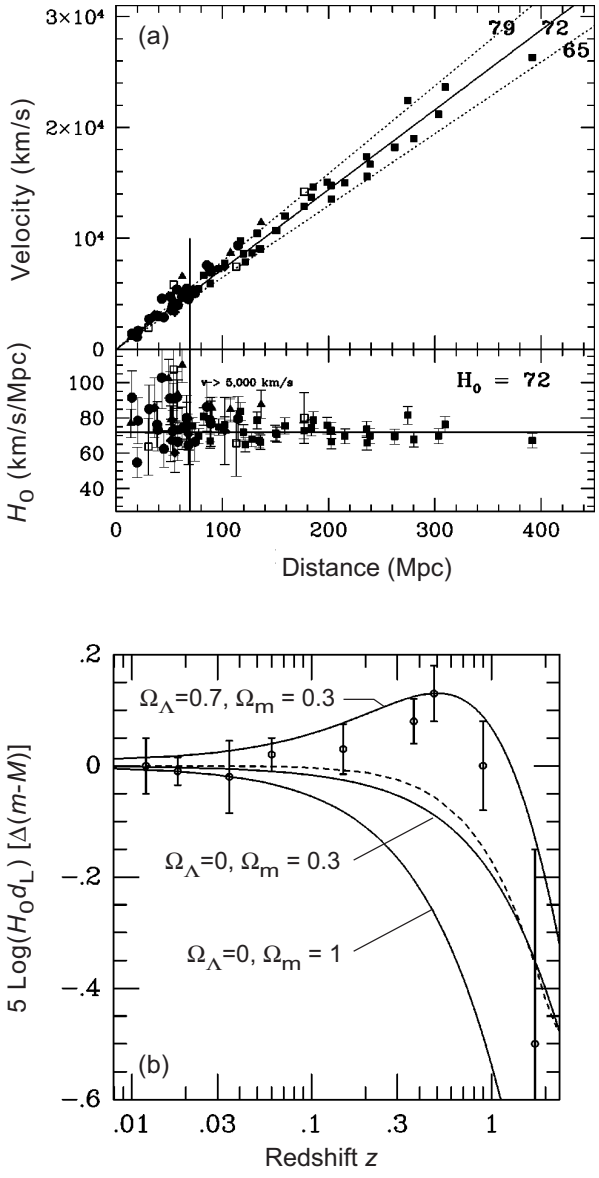


Figure 19.6: (a) Hubble constant determined from low-redshift data. (b) Differential Hubble diagram showing distance modulus versus that for an empty universe ($\Omega = 0$) plotted versus redshift. Data points represent the binning of data from more than 200 high-redshift, Type Ia supernovae. The dashed curve corresponds to a non-accelerating flat universe; points above it indicate acceleration of the expansion. The three solid curves correspond to theoretical models with $\Omega_\Lambda = 0.7$ and $\Omega_m = 0.3$ (top curve), $\Omega_\Lambda = 0.0$ and $\Omega_m = 0.3$ (middle curve), and $\Omega_\Lambda = 0.0$ and $\Omega_m = 1.0$ (bottom curve).

Radiation Density: We may dispose quickly of the radiation energy density.

- Various observations indicate that the radiation density is a factor of more than 10,000 less dense than matter or vacuum energy in the current Universe.
- Therefore radiation density can safely be set to zero for the purposes of cosmological calculations.

This was not always true; as we shall see later, in the early Universe the radiation dominated, but its contribution fell off quickly.

The values of the remaining parameters are best determined by

1. The use of *Type Ia supernovae* as standard candles to constrain the evolution of the scale factor at high redshifts.
2. Quantitative analysis of the fluctuations in the *cosmic microwave background* to constrain a variety of cosmological parameters.
3. Detailed *observations of galaxy clusters* to constrain the amount of matter in the Universe.
4. Quantitative analysis of statistical correlations in the large-scale distribution of galaxies *(baryon acoustic oscillations)* to constrain various cosmological parameters.

We shall discuss all of these further in this and following chapters.

19.4.4 Type Ia Supernovae as Standardizable Candles

Type Ia supernovae are valuable in cosmology because they are *standardizable candles*.

- A *standard candle* is a light source that always has the same intrinsic brightness under some specified conditions.
- A *standardizable candle* may vary in brightness but can be *normalized to a common brightness* by a reliable method.
- Thus, once normalized a standardizable candle becomes effectively a standard candle.

Standard candles enable distance measurement by comparing observed brightness with the standard brightness using

$$F = \frac{L}{4\pi r^2}$$

where F is flux, L is luminosity, and r is distance.

- In *flat, static space*, for a standard candle L is known and measurement of the flux F yields the distance r.
- A distance determined by measuring flux from a standard candle is called a *luminosity distance*.

> The above relationship assumes light propagating in static euclidean space and *requires modification in an expanding, curved space*.

19.4.5 Luminosity Distances in Curved, Expanding Space

One way to proceed in a non-euclidean space is to *define the luminosity distance* $d_{\rm L}$ by

$$F = \frac{L}{4\pi d_{\rm L}^2} \quad \longrightarrow \quad d_{\rm L}^2 \equiv \frac{L}{4\pi F}$$

where F is the observed flux (energy/area/time) and L is the (assumed known) luminosity (energy/time).

- For a flat, static universe, the luminosity distance $d_{\rm L}$ equals the proper distance ℓ.
- However, three effects alter this relationship:
  - The geometry may be curved.
  - Expansion redshifts the energy of photons.
  - Expansion lengthens the time between detection of successive photons.

To investigate, we express the RW metric as,

$$ds^2 = -dt^2 + a(t)^2[dr^2 + S_k(r)^2\, d\Omega^2]$$

$$S_k(r) = \begin{cases} \sin r & (k=+1) \\ r & (k=0) \\ \sinh r & (k=-1) \end{cases}$$

and consider light propagation in this metric.

Let photons be emitted from coordinates (r,θ,φ) at time $t_{\rm e}$ and *spread over a sphere when detected* at $r=0$ and time t_0.

- This sphere at detection has a *proper area* (Problem)

$$A_{\rm p}(t_0) = 4\pi S_k(r)^2.$$

- Next, Hubble expansion *redshifts photon wavelengths* by

$$\lambda_0 = \frac{\lambda_{\rm e}}{a(t_{\rm e})} = (1+z)\lambda_{\rm e},$$

so *emitted energy* $E_{\rm e}$ and *detected energy* E_0 obey

$$E(t_0) = \frac{E(t_{\rm e})}{1+z}.$$

- Finally, time between successive photons is increased by expansion, so the *time interval* Δt_0 *at detection* and the *time interval* $\Delta t_{\rm e}$ *at emission* are related by

$$\Delta t_0 = (1+z)\Delta t_{\rm e}.$$

- Collecting these effects, the *observed flux* is

$$F = \frac{L}{4\pi S_k(r)^2(1+z)^2},$$

and comparing with the definition $d_{\rm L}^2 \equiv L/4\pi F$ gives

$$d_{\rm L} = S_k(r)(1+z) \simeq (1+z)r = (1+z)\ell(t_0),$$

where $\ell(t_0)$ is the current proper distance and the *actual Universe has been assumed flat* with $S_k(r) \sim r$.

As we have just shown, the luminosity distance $d_{\rm L}$ and the proper distance ℓ are related by

$$d_{\rm L} = S_k(r)(1+z) \simeq (1+z)r = (1+z)\ell(t_0),$$

Only for small z does the *luminosity distance* $d_{\rm L}$ equal the *proper distance* ℓ in an expanding, possibly curved, universe.

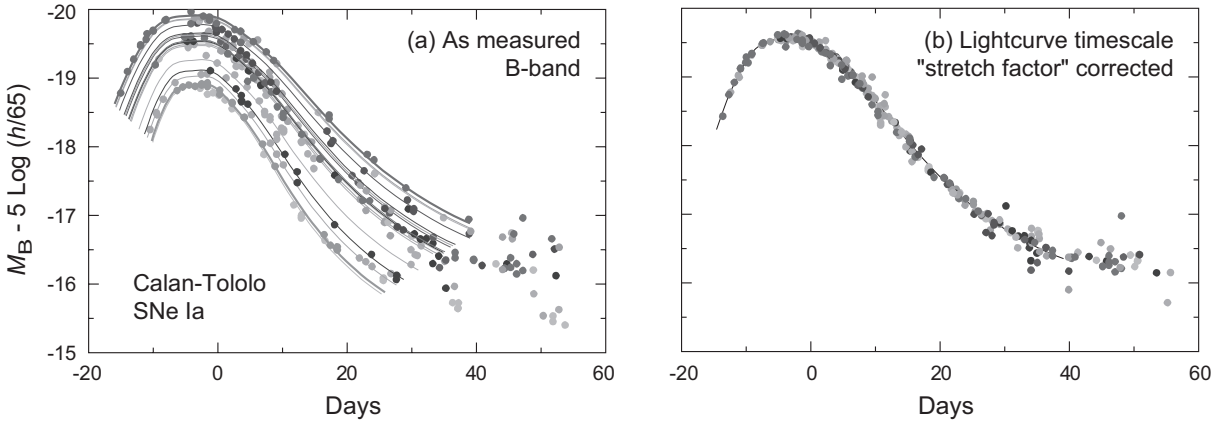


Figure 19.7: Empirical rescaling of Type Ia supernova lightcurves to make them approximate standard candles. (a) Lightcurves for low-redshift supernovae (Calan-Tololo survey). The intrinsic scatter is 0.3 magnitudes in peak luminosity. (b) After a one-parameter correction the dispersion is reduced to 0.15 magnitudes.

The standardizable candle properties of Type Ia supernovae are illustrated in Fig. 19.7.

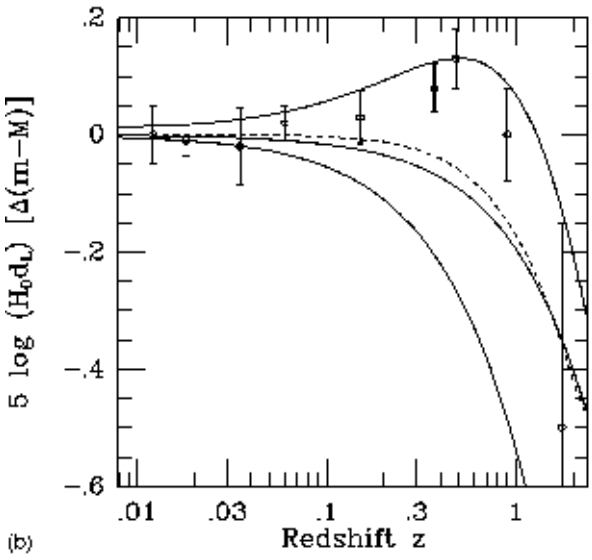


Figure 19.8: Differential Hubble diagram showing distance modulus vs. that for an empty universe ($\Omega = 0$) plotted versus redshift. Data points represent binning of data from more than 200 high-redshift, Type Ia supernovae. The dashed curve corresponds to a non-accelerating flat universe; points above it indicate acceleration of the expansion. The three solid curves correspond to theoretical models with $\Omega_\Lambda = 0.7$ and $\Omega_\mathrm{m} = 0.3$ (top curve), $\Omega_\Lambda = 0.0$ and $\Omega_\mathrm{m} = 0.3$ (middle curve), and $\Omega_\Lambda = 0.0$ and $\Omega_\mathrm{m} = 1.0$ (bottom curve).

The figure above shows data from

- *200 high-z Type Ia supernovae* binned into a few points and compared with theoretical calculations.
- At the highest redshift there is a clear preference for the solution with nontrivial vacuum energy.

These data imply *acceleration of the expansion* and the *presence of vacuum energy* in the Universe.

CMB Fluctuations: In the next chapter it will be shown that

- Measurements of anisotropies in the cosmic microwave background (CMB) provide a precise and independent determination of cosmological parameters.
- These analyses of the CMB probe the relative contribution of matter and vacuum energy, but in a different way than for the Type Ia supernova data.

Galaxy Clusters: Traditional observational astronomy,

- augmented by newer techniques like gravitational lensing,
- has constrained the amount of matter (visible and dark) contained in clusters of galaxies rather tightly.
- Such galaxy observations are relatively insensitive to the amount of dark energy in the Universe.

Remarkably, these approaches having senitivities to different parts of the cosmic fluid:

- Type Ia supernovae
- CMB fluctuations
- inventories of galaxy clusters

have reached a *concensus* for the values of cosmological parameters called the *concordance model*.

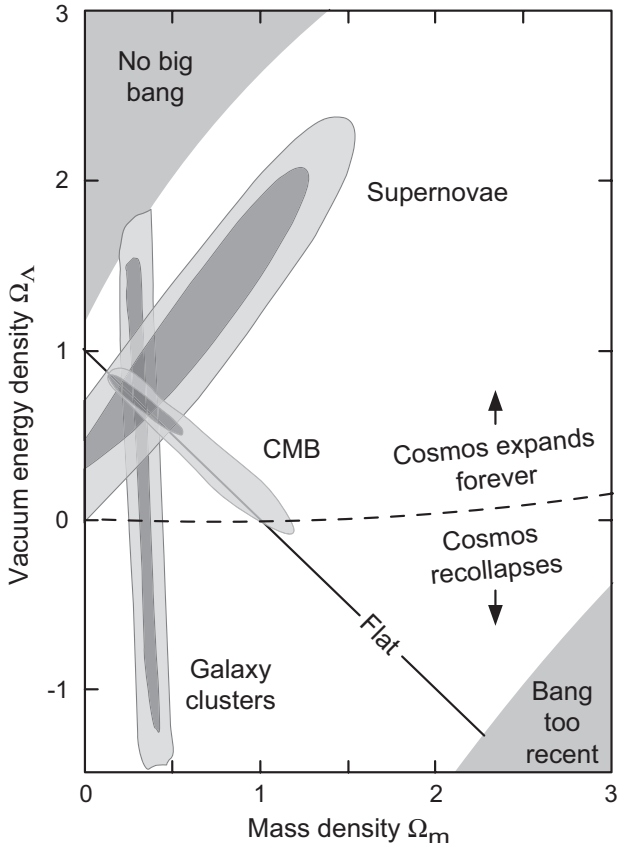
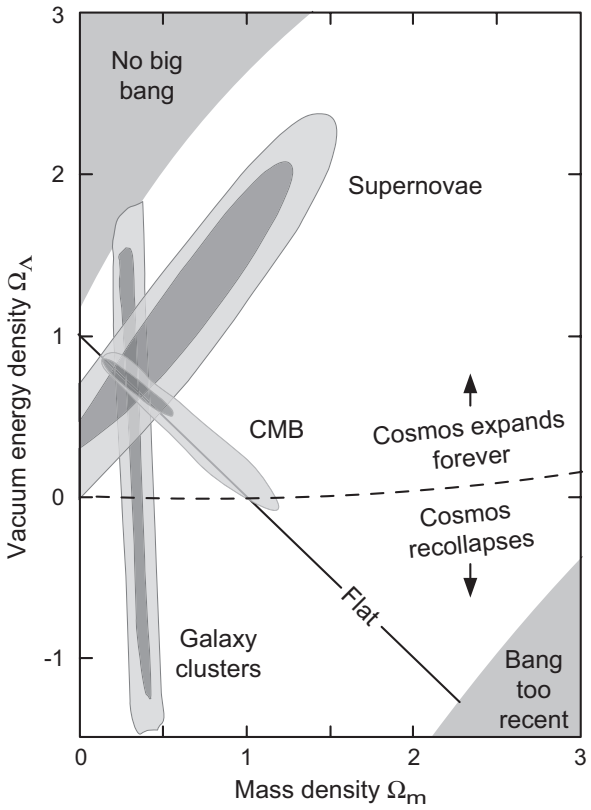


Figure 19.9: Cosmological parameter space. High redshift supernova data, CMB data, and galaxy cluster inventories converge on $\Omega_m \simeq 0.3$, $\Omega_\Lambda = 0.7$, and $\Omega_r \sim 0$, implying a cosmos that is flat but will expand forever.

In Fig. 19.9 supernova, CMB, and galaxy data are summarized in terms of confidence level contours.

- Notice the different dependence of the supernova data on the parameters than for the CMB and galaxy clusters data.
- The supernova and CMB data taken together imply that $\Omega_m \simeq 0.3$ and $\Omega_\Lambda \simeq 0.7$, with negligible contribution from radiation.



- In the figure above, data from galaxy cluster surveys are also displayed. These have *little sensitivity to vacuum energy density* but have *high sensitivity to matter density*.
- Remarkably, data from galaxy clusters also are consistent with the choice $\Omega_\Lambda \simeq 0.7$ and $\Omega_m \simeq 0.3$.
- The diagonal line indicates the flat space prediction of inflationary theory ($\Omega_m + \Omega_\Lambda = 1$), to be discussed later.
- The dashed line separates an eternally expanding Universe from one that eventually recollapses.
- Gray regions in the corners indicate parameters that would allow no big bang, or would cause it to occur too recently.

Baryon Acoustic Oscillations (BAO): Baryon acoustic oscillations are periodic fluctuations in the density of visible matter (galaxy clustering on large scales).

- They originated in flucutations present in the early Universe at the time of hydrogen recombination.
- This will be discussed in more detail in the next chapter.

Baryon acoustic oscillations are introduced here because BAO measurements are sensitive to cosmological density parameters, particularly the matter density.

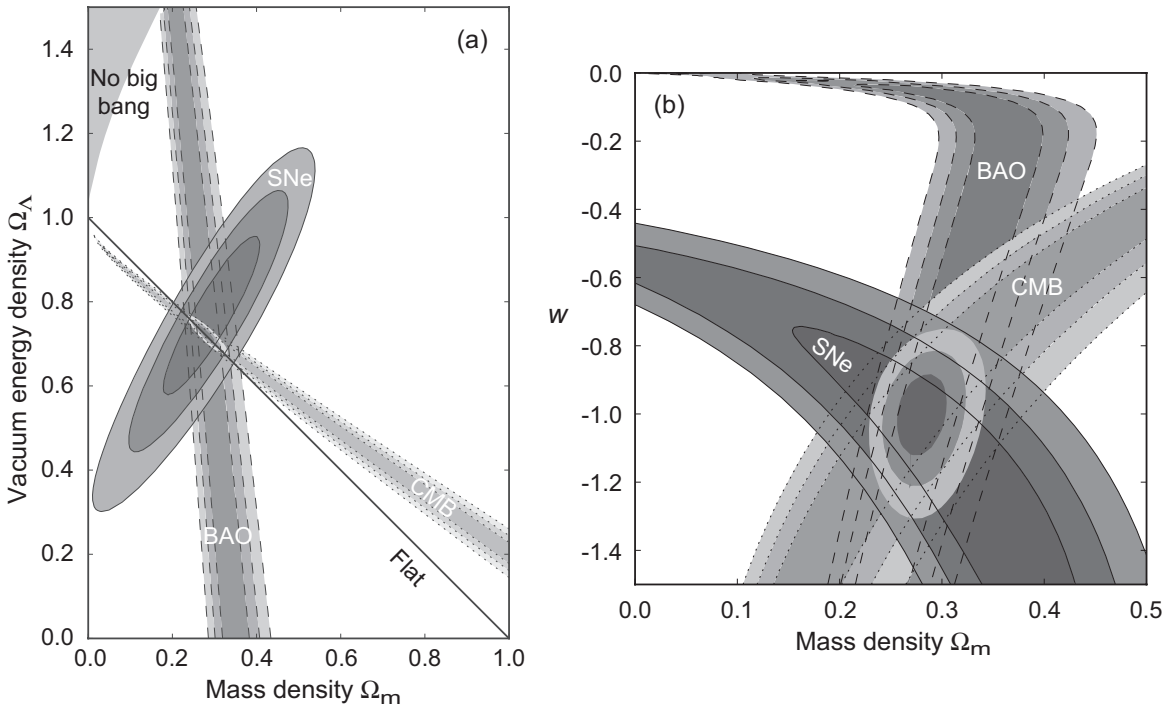


Figure 19.10: (a) Cosmological parameter space comparing data from six Type Ia supernovae lying between redshifts $z = 0.51$ and $z = 1.12$ (SNe), cosmic microwave background (CMB), and baryon acoustic oscillation (BAO) data. (b) Corresponding equation of state parameter w from $P = w\varepsilon$.

Fig. 19.10(a) illustrates parameter concordance for *supernova data*, *CMB fluctuations*, and *baryon acoustic oscillations*.

- Best-fit Ω_m and Ω_Λ are comparable to those found earlier.
- Constraint of the *equation of state parameter* w from the same comparison is illustrated in Fig. 19.10(b).

It was concluded from these data that if the Universe is flat,

$$w = -0.997^{+0.077}_{-0.082}$$

for the cosmic equation of state $P = w\varepsilon$.

Summary: We are led by recent cosmological data to adopt as a *benchmark model* the parameters

$$H_0 = 67 - 73\,\mathrm{km\,s^{-1}\,Mpc^{-1}} \qquad \Omega_\mathrm{r} = 8 \times 10^{-5} \simeq 0$$
$$\Omega_\mathrm{m} \sim 0.3 \qquad \Omega_\mathrm{v} \sim 0.7.$$

The reliable determination of such parameters changed cosmology from a notiously qualitative discipline to a precision science in little more than a decade.

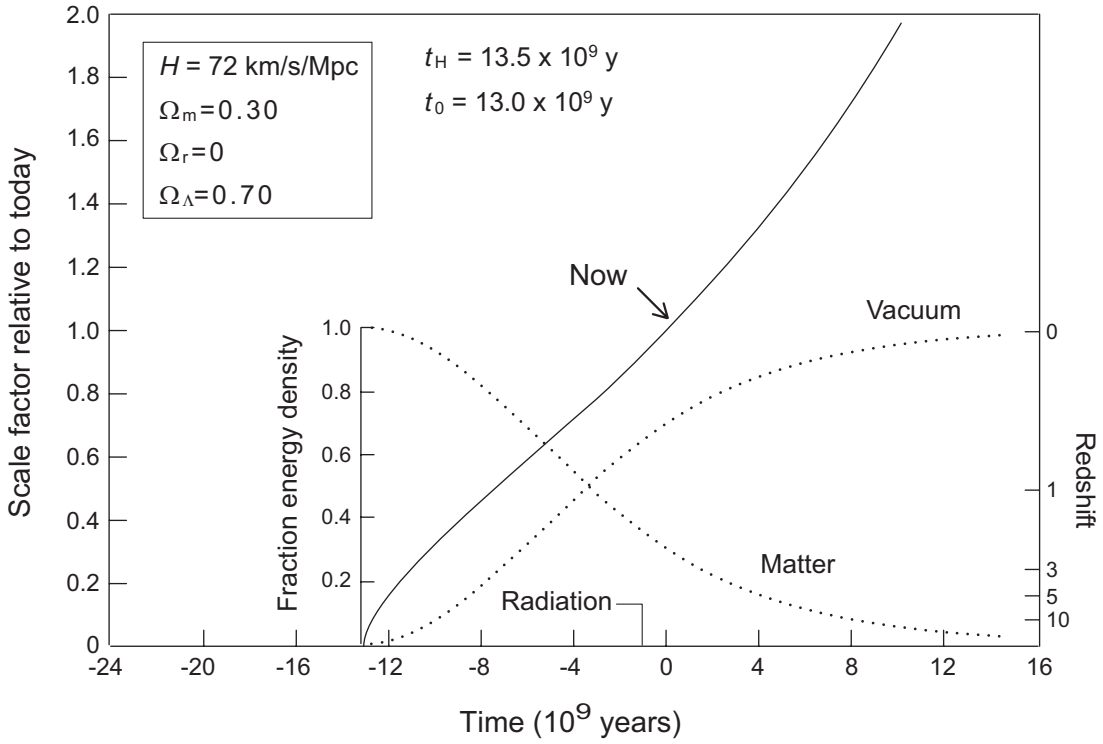
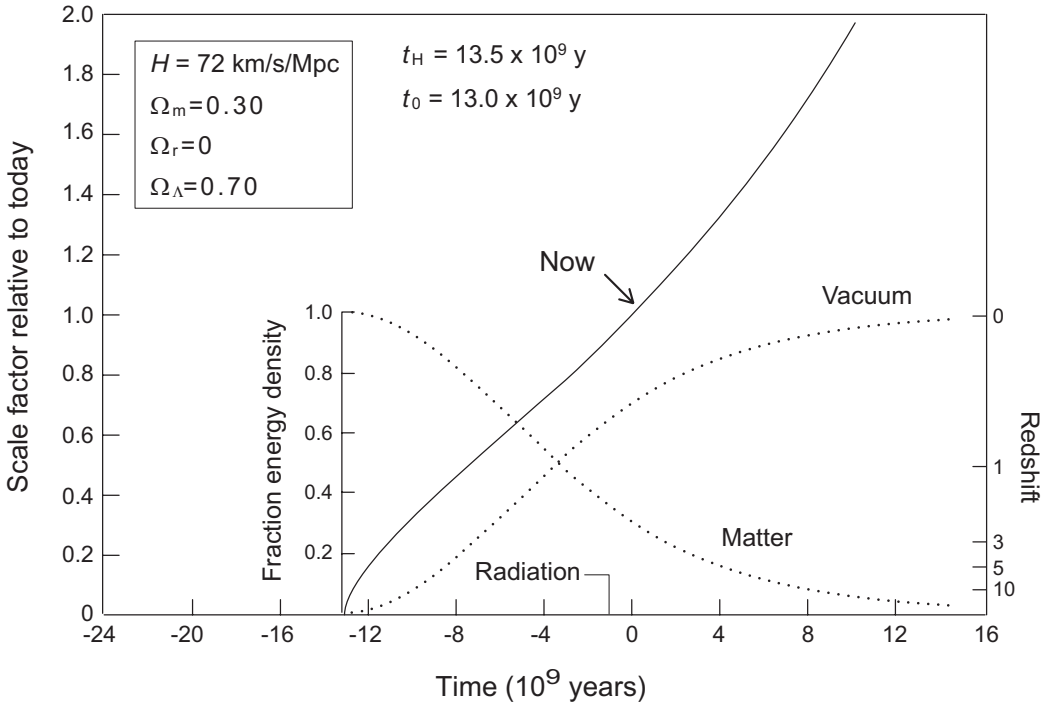


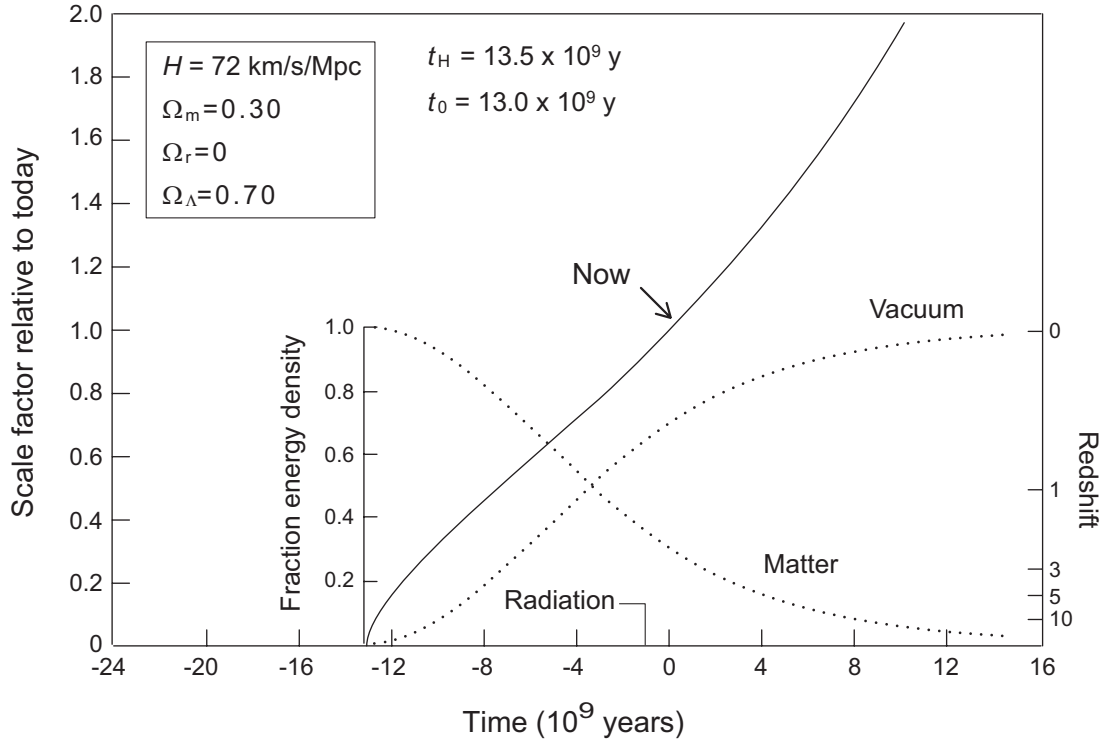
Figure 19.11: Scale factor versus time for a flat universe with 30% matter, 70% vacuum energy density, and negligible radiation. The energy density of matter and vacuum energy as a function of time are indicated in the inset by dotted lines. The left axis is marked in terms of the ratio of the scale factor to its value today and the right axis is marked in terms of redshift. In this model, which is consistent with current best values of the cosmological parameters, the expansion is presently dominated by vacuum energy and hence is accelerating.

19.4.6 Calculations with the Benchmark Parameters

In Fig. 19.11 a *solution of the Friedmann equations* using the *benchmark parameters* assuming $H_0 = 72\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ is displayed.



- The Universe was dominated by matter for most of its history since the very early radiation-dominated era (too short to see on this scale).
- About 4 billion years ago the vacuum energy gained ascendency relative to the non-relativistic matter and
- the Universe is now in an *accelerating phase* corresponding to steadily increasing *dominance of vacuum energy*.
- The Universe first decelerated (under the dominant influence of first radiation and then matter) and then began accelerating as the vacuum energy became dominant.
- This solution indicates that the Universe is *flat*, but it will *expand forever* because of vacuum energy.



- The calculation displayed above represents a *realistic evolution of the actual Universe*.
- However, more recent parameters constrained by cosmic microwave background data indicate
  - a somewhat smaller value of the Hubble constant and
  - a somewhat greater age for the Universe.

The calculation shown above, repeated for parameters determined by CMB analysis (next chapter) gives a larger age of 13.8 billion years, mostly because of choosing $H_0 = 67.8\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$

Chapter 20

The Big Bang

The Universe began life in a *very hot, very dense state* called the *big bang.* In this chapter we

- apply the Friedmann equations to the early Universe in an attempt to
- understand the most important features of the *big bang model,*

which is the *cosmologist's "standard model"* for the origin of the present Universe.

20.1 Radiation and Matter Dominated Universes

Because

- the influence of vacuum energy grows with expansion of the Universe, and
- vacuum energy is only today beginning to dominate,
- we may safely assume that it was *negligible in the early Universe* (once the inflationary epoch was over).

In that case, two extremes for the equation of state give us considerable insight into the early history of the Universe:

1. If the energy density resides primarily in light, relativistic particles *(radiation dominated)*, the equation of state is

$$P = \tfrac{1}{3}\varepsilon \qquad \text{(radiation dominated)}.$$

2. If on the other hand the energy density resides in massive, slow-moving particles *(matter dominated)*, the equation of state is

$$P \simeq 0 \qquad \text{(matter dominated)}.$$

In either extreme, the evolution of the Universe is then easily calculated using the Friedmann equations.

20.1.1 Evolution of the Scale Factor

The density of radiation and the density of matter scale differently in an expanding universe.

- If the Universe is *radiation dominated*, $P = \frac{1}{3}\varepsilon$ and

$$\dot{\varepsilon} + 3(\varepsilon + P)\frac{\dot{a}}{a} = 0 \quad \rightarrow \quad \frac{\dot{\varepsilon}}{\varepsilon} + \frac{4\dot{a}}{a} = 0,$$

  which has a solution

$$\varepsilon(t) \simeq \frac{1}{a^4(t)} \qquad \text{(radiation dominated)}.$$

- If on the other hand the Universe is *matter dominated*, we have $P \simeq 0$ and

$$\dot{\varepsilon} + 3(\varepsilon + P)\frac{\dot{a}}{a} = 0 \quad \rightarrow \quad \frac{\dot{\varepsilon}}{\varepsilon} + \frac{3\dot{a}}{a} = 0,$$

  which has a solution

$$\varepsilon(t) \simeq \frac{1}{a^3(t)} \qquad \text{(matter dominated)}.$$

As shown in the previous chapter, the corresponding scale factors behave as

$$a(t) \simeq \begin{cases} t^{1/2} & \text{(radiation dominated)} \\ t^{2/3} & \text{(matter dominated)} \end{cases}$$

20.1.2 Matter and Radiation Density

In the present Universe the *ratio of baryons to photons* is

$$\frac{n_{\rm b}}{n_\gamma} \simeq 10^{-9}.$$

However,

- the rest mass of a typical baryon is $\sim 10^9$ eV, while most
- photons are in the ~ 2.7 K *CMB*, with average energy

$$E_\gamma \simeq kT = (2.7\,{\rm K})\left(\frac{1\,{\rm GeV}}{1.2\times 10^{13}\,{\rm K}}\right) \simeq 2.3\times 10^{-4}\,{\rm eV}.$$

Thus the energy density of baryons relative to photons in the present Universe is

$$\frac{\varepsilon_{\rm b}}{\varepsilon_\gamma} \simeq 10^3 - 10^4$$

and the present Universe is *dominated by matter (and vacuum energy)*, with only a *small contribution from radiation*. But,

$$\varepsilon(t) \simeq \frac{1}{a^4(t)} \qquad \text{(radiation)}.$$

$$\varepsilon(t) \simeq \frac{1}{a^3(t)} \qquad \text{(matter)}.$$

Thus, as time is extrapolated backwards $a \to 0$ and *relativistic matter becomes dominant*.

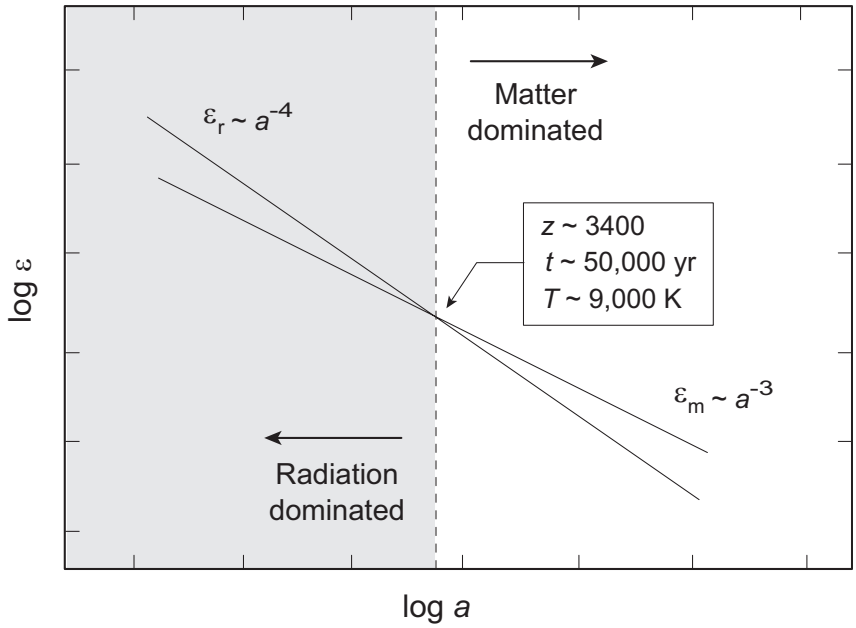
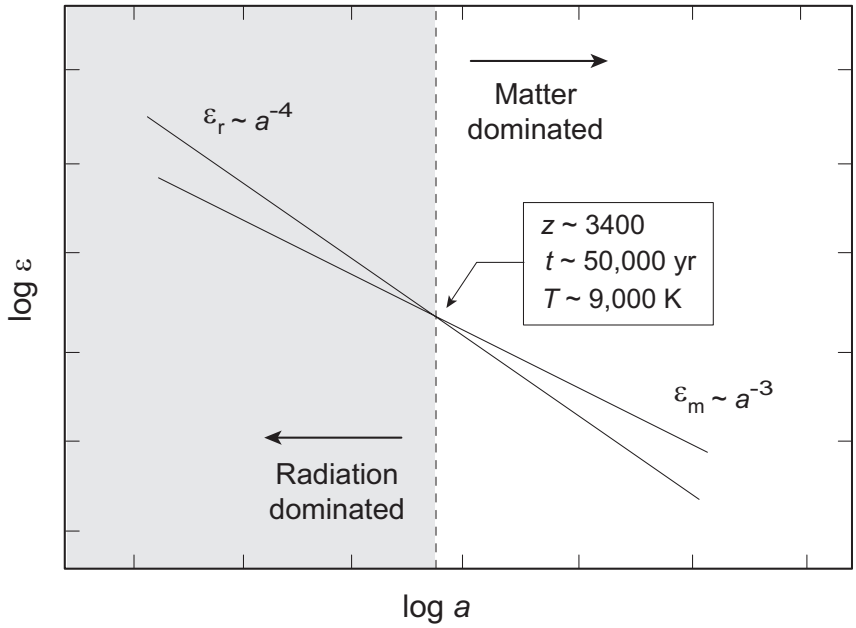


Figure 20.1: Schematic dependence of matter and radiation energy densities on the scale factor. The time of equivalence corresponds to a redshift of $z_{\rm eq} \sim 3400$, based on analysis of CMB fluctuations by the Planck satellite.

Fig. 20.1 illustrates. Thus, the early Universe should have been *radiation dominated*.



In this early, radiation-dominated Universe,

$$\varepsilon \simeq a^{-4} \qquad a \simeq t^{1/2} \qquad P = \tfrac{1}{3}\varepsilon,$$

- so the density and pressure tend to infinity as $t \to 0$.
- Furthermore, for a radiation dominated Universe,

$$T \simeq a^{-1} \simeq t^{-1/2}.$$

- Thus, as time is extrapolated backwards the temperature also tends to infinity.

These considerations suggest that the Universe started from a *very hot, very dense initial state* with $a(t \to 0) \to 0$.

- This initial state is called the *big bang*.
- If we take $t = 0$ when $a = 0$,
  - the transition between the earlier radiation dominated universe and one dominated by matter
  - happened approximately 50,000 years after the big bang,
  - corresponding to a redshift $z \sim 3400$,
  - by which time the temperature had dropped to about 9000 K.
- This *matter dominance* then continued until about 4-5 billion years ago, when the vacuum energy density began to overtake the matter density.

In the following sections we shall discuss in more detail the big bang and the *early radiation-dominated era* of the Universe.

> The name "big bang" was a term coined by opponents of this cosmology who favored the now discredited steady state theory. The name stuck, as did the theory.

20.2 Evolution of the Early Universe

The early Universe

- was *very hot* and *very dense*, and
- the major part of the energy density resided in photons and other *massless or nearly massless particles* like neutrinos.

We now give a description of the most important events in the big bang, and the transition from

- a universe *dominated by radiation* to
- a universe *dominated by matter*.

20.3 Thermodynamics of the Big Bang

We have established that in the initial *radiation dominated era* of the big bang, *curvature was negligible* and

$$H^2 \equiv \left(\frac{\dot{a}}{a}\right)^2 \simeq \frac{8\pi G}{3}\varepsilon_{\rm r} = a^{-4} \qquad a \simeq t^{1/2} \qquad H = \frac{\dot{a}}{a} = \frac{1}{2t}.$$

- The average evolution was that of an *ideal gas in thermal equilibrium*, with a number density for a particular species given by

$$dn = \frac{g}{2\pi^2\hbar^3}\frac{p^2 dp}{e^{E/kT} + \Theta},$$

where

  - p is the 3-momentum,
  - g is the number of degrees of freedom (helicity states: 2 for each photon, massive quark, and lepton),
  - and Θ and the energy E are given by

$$E = \sqrt{p^2c^2 + m^2c^4} \qquad \Theta = \begin{cases} +1 & \text{Fermions} \\ -1 & \text{Bosons} \\ 0 & \text{Maxwell–Boltzmann} \end{cases}$$

where Maxwell–Boltzmann statistics apply only if we make no distinction between fermions and bosons.

Let's assume that *the gas is ultrarelativistic* $(kT >> mc^2)$.

- Then the *particle energy* is $E = pc$, and
- the particle *number density* is obtained by integrating the previous expression for dn.

$$n = \int_0^\infty \frac{dn}{dp} dp = \frac{g}{2\pi^2\hbar^3} \int_0^\infty \frac{p^2 dp}{e^{E/kT} + \Theta} = \frac{g}{2\pi^2\hbar^3} \int_0^\infty \frac{p^2 dp}{e^{pc/kT} + \Theta}.$$

Integrals of this form may be *evaluated using*

$$\int_0^\infty \frac{t^{z-1}}{e^t - 1} dt = (z-1)!\,\zeta(z),$$

$$\int_0^\infty \frac{t^{z-1}}{e^t + 1} dt = (1 - 2^{1-z})(z-1)!\,\zeta(z),$$

where $\zeta(z)$ is the *Riemann zeta function*, with tabulated values

$$\zeta(2) = \frac{\pi^2}{6} = 1.645 \quad \zeta(3) = 1.202 \quad \zeta(4) = \frac{\pi^4}{90} = 1.082.$$

The results for the *number density of species* i are

$$n_i = g_i \frac{\zeta(3)}{\pi^2} T^3 \times \begin{cases} 1 & \text{Bose–Einstein} \\ 3/4 & \text{Fermi–Dirac} \\ \zeta(3)^{-1} & \text{Maxwell–Boltzmann} \end{cases}$$

where now we *introduce* $\hbar = c = k = 1$ *units*.

Likewise, the *energy density is*

$$\varepsilon = \rho c^2 = \int_0^\infty E\frac{dn}{dp}dp = \frac{g}{2\pi^2\hbar^3}\int_0^\infty \frac{Ep^2dp}{e^{E/kT}+\Theta} = \frac{g}{2\pi^2\hbar^3}\int_0^\infty \frac{Ep^2dp}{e^{pc/kT}+\Theta},$$

which gives for a species i,

$$\varepsilon_i = g_i\frac{\pi^2}{30}T^4 \times \begin{cases} 1 & \text{Bose–Einstein} \\ 7/8 & \text{Fermi–Dirac} \\ 90/\pi^4 & \text{Maxwell–Boltzmann} \end{cases}.$$

The *energy density for all relativistic particles* is then the sum,

$$\varepsilon = g_*\frac{\pi^2}{30}T^4 \qquad g_* \equiv \sum_{\text{bosons}} g_\text{b} + \tfrac{7}{8}\sum_{\text{fermions}} g_\text{f}.$$

If all species are in equilibrium, the *entropy density* s is

$$s = \frac{\varepsilon + P}{T} = \frac{4\varepsilon}{3T} = \frac{2\pi^2}{45}g_*T^3,$$

where we note from comparing this and

$$n_i = g_i\frac{\zeta(3)}{\pi^2}T^3 \times \begin{cases} 1 & \text{Bose–Einstein} \\ 3/4 & \text{Fermi–Dirac} \\ \zeta(3)^{-1} & \text{Maxwell–Boltzmann} \end{cases}$$

that $s \simeq \sum n_i$. The *entropy per comoving volume is constant*,

$$S \simeq sa^3 \simeq \text{ constant} \qquad \longrightarrow \qquad \frac{d(sa^3)}{dt} = 0$$

(adiabatic expansion), provided that g_* *does not change*.

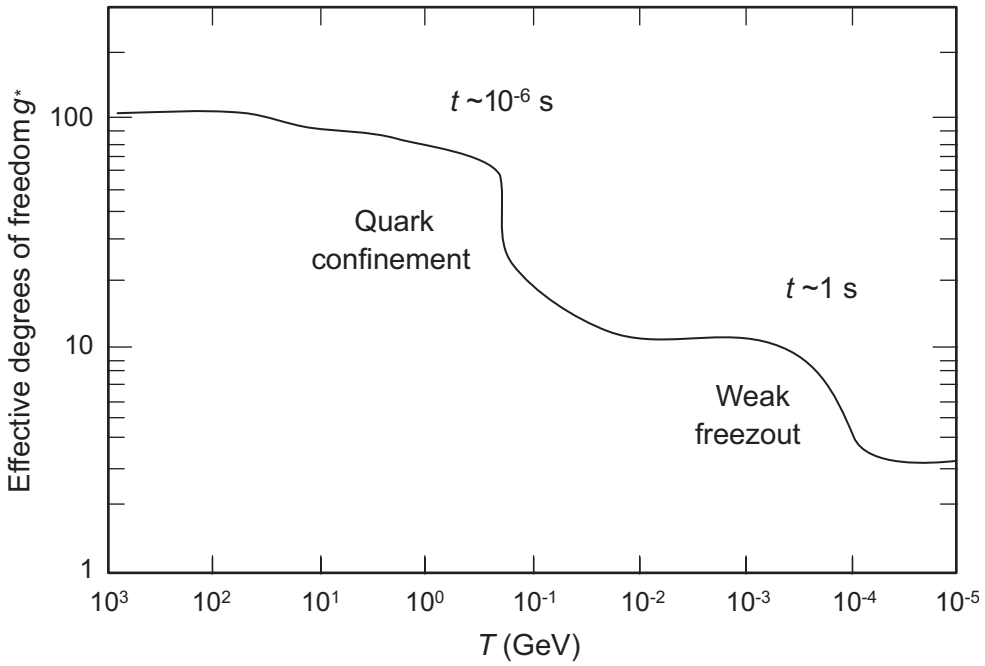


Figure 20.2: Variation of the effective number of degrees of freedom in the early Universe as a function of temperature.

In fact, we expect that *the effective number of degrees of freedom* g_* *is not constant*. Instead,

- g_* should be *approximately constant* for broad ranges of temperature but
- it will *change suddenly* at critical temperatures where kT *becomes comparable to the rest mass for a species*.

Fig. 20.2 illustrates the expected variation of g_* with temperature in the early Universe.

Table 20.1: Particles of the Standard Model

Particle	Symbol	Spin ($\hbar$)	Charge (e)	Mass (GeV/c^2)†
		– Leptons –		
Electron	e^-	$1/2$	-1	5.11×10^{-4}
Electron neutrino	ν_e	$1/2$	0	$< 2.2 \times 10^{-9}$
Muon	μ^-	$1/2$	-1	0.1057
Muon neutrino	ν_μ	$1/2$	0	$< 1.7 \times 10^{-4}$
Tau	τ^-	$1/2$	-1	1.777
Tau neutrino	ν_τ	$1/2$	0	$< 1.6 \times 10^{-2}$
		– Quarks –		
Down	d	$1/2$	$-1/3$	0.005
Up	u	$1/2$	$2/3$	0.002
Strange	s	$1/2$	$-1/3$	0.096
Charm	c	$1/2$	$2/3$	1.28
Bottom	b	$1/2$	$-1/3$	4.2
Top	t	$1/2$	$2/3$	173
		– Gauge and Higgs bosons –		
Photon	γ	1	0	0
Charged weak bosons	$W^\pm$	1	± 1	80.4
Neutral weak boson	Z^0	1	0	91.2
Gluons	$G_1, G_2, \ldots, G_8$	1	0	0
Graviton (?)	g	2	0	0
Higgs	H	0	0	125.1

†Quark masses are so-called current or Lagrangian masses.

The rest masses of particles from the Standard Model that are relevant for the early Universe are summarized in Table 20.1.

From the earlier expressions for the entropy density,

$$s = \frac{\varepsilon + P}{T} = \frac{4\varepsilon}{3T} = \frac{2\pi^2}{45} g_* T^3,$$

$$S \simeq sa^3 \simeq \text{constant} \qquad \longrightarrow \qquad \frac{d(sa^3)}{dt} = 0,$$

- we have that $sa^3 \simeq T^3 a^3$ *is constant;* hence the temperature varies as

$$T \simeq \frac{1}{a} \simeq t^{-1/2},$$

- where we have used the result that in a *radiation dominated universe* of negligible curvature,

$$a \simeq t^{1/2}.$$

To be precise, *evolution of the ultrarelativistic, hot plasma* characterizing the early big bang is described by

$$\frac{\dot{a}}{a} = -\frac{\dot{T}}{T} = \alpha T^2 \qquad t = \frac{1}{2\alpha T^2} = \frac{2.4 \times 10^{-6}}{g_*^{1/2} T^2} \,\text{GeV}^2\,\text{s}$$

$$\alpha = \left(\frac{4\pi^3 g_*}{45 M_\text{P}^2}\right)^{1/2} \qquad M_\text{P} \equiv \left(\frac{\hbar c}{G}\right)^{1/2} = 1.2 \times 10^{19}\,\text{GeV},$$

where M_P is the *Planck mass*.

20.3.1 Equilibrium in an Expanding Universe

Strictly, *equilibrium does not hold in an expanding universe*.

- However, a *practical equilibrium* can exist as the Universe passes through a *series of nearly equilibrated states*.
- We may expect both *thermal equilibrium* and *chemical equilibrium* to play a role in the expansion of the Universe.

A system is in *thermal equilibrium* if its *number density is given by*

$$dn = \frac{g}{2\pi^2\hbar^3}\frac{p^2dp}{e^{E/kT}+\Theta},$$

$$E = \sqrt{p^2c^2+m^2c^4} \qquad \Theta = \begin{cases} +1 & \text{Fermions} \\ -1 & \text{Bosons} \\ 0 & \text{Maxwell–Boltzmann} \end{cases}$$

A system is in *chemical equilibrium* if for $a+b \leftrightarrow c+d$ (say) the chemical potentials satisfy

$$\mu_a + \mu_b = \mu_c + \mu_d.$$

We shall illustrate the discussion primarily by *thermal equilibrium*, and will consider the equilibrium to be maintained by *two-body reactions* (the most common situation).

The *reaction rate* for a two-body reaction may be expressed as

$$\Gamma \simeq \langle nv\sigma \rangle,$$

- where n is the *number density*,
- v is the *relative speed*,
- σ is the *reaction cross section*, and
- the brackets $\langle\ \rangle$ indicate a *thermal average*.

A species will *remain in thermal equilibrium* in the radiation dominated Universe as long as

$$\Gamma >> \frac{\dot{a}}{a} \equiv H \simeq \frac{d(t^{1/2})/dt}{t^{1/2}} = \frac{1}{2t}.$$

- At early times, densities, velocities, and cross sections are large and it is *easy to fulfil this condition* for most species.
- However, as T and the density drop the number density and velocity factors will decrease steadily and
- at *certain reaction thresholds* the reaction rates maintaining equilibrium will become *too small for a particular species*, and it can drop out of thermal equilibrium.

> *Physical reason:* if reaction rates are slow compared with the rate of expansion, it is unlikely that particles can find each other often enough to react and maintain equilibrium.

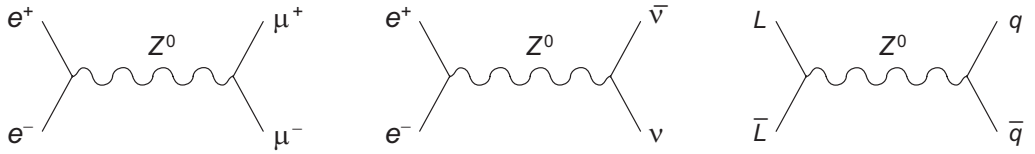


Figure 20.3: Some weak interactions important for equilibrium in the early Universe. Generic leptons are represented by L and generic quarks by q.

20.3.2 Example: Decoupling of the Weak Interactions

Consider *weak interactions in the early Universe*.

- The *strength of the weak interactions* varies as T^2.
- Thus, shortly after the big bang the weak interactions are not so weak and particles such as neutrinos are kept in equilibrium by reactions like $\nu\bar{\nu} \leftrightarrow e^+e^-$.
- Typical Feynman diagrams are shown in Fig. 20.3. From these, weak interaction cross sections vary as

$$\sigma_{\rm w} \propto G_{\rm F}^2 \qquad G_{\rm F} \simeq 1.17 \times 10^{-5}\,{\rm GeV}^{-2}.$$

- By *dimensional analysis*, this implies that the ratio of the weak reaction rate to the expansion rate is

$$\frac{\Gamma}{H} \simeq \frac{G_{\rm F}^2 T^5}{T^2/M_{\rm P}} \simeq \left(\frac{T}{1\,{\rm MeV}}\right)^3.$$

Weak interactions *decoupled from thermal equilibrium* ($\Gamma/H \sim 1$) at $T \sim 1\,{\rm MeV}$, which occurred about 1 second after the expansion began.

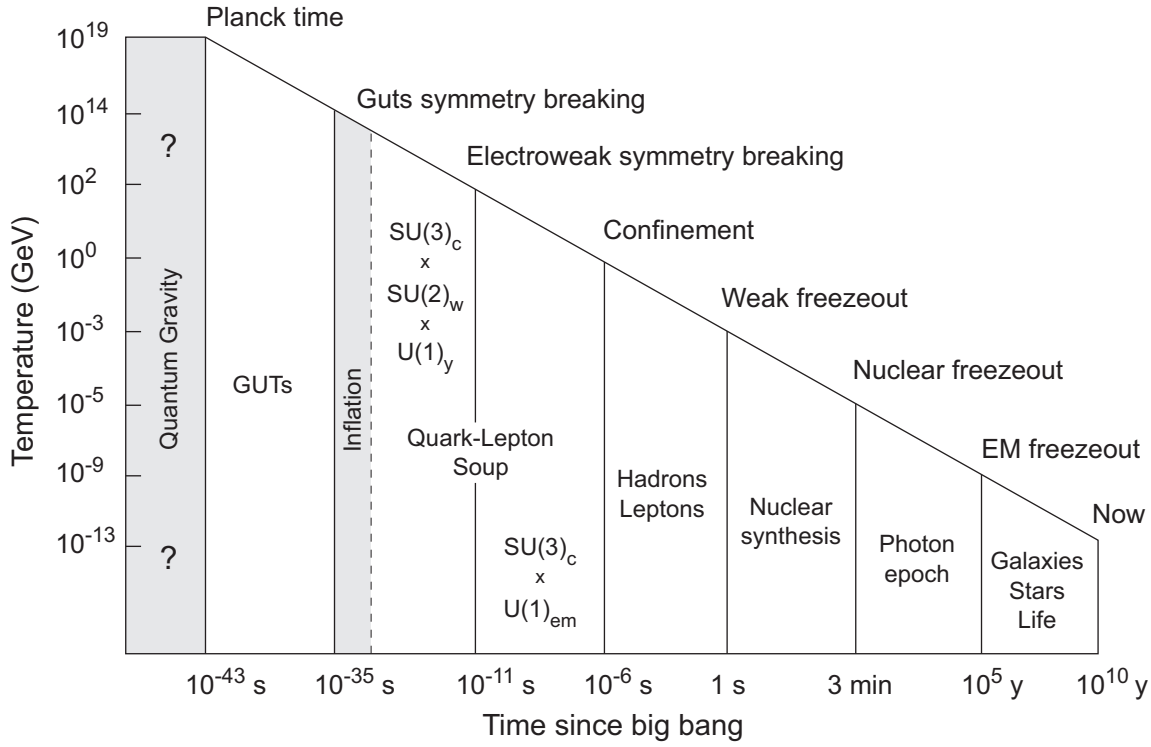
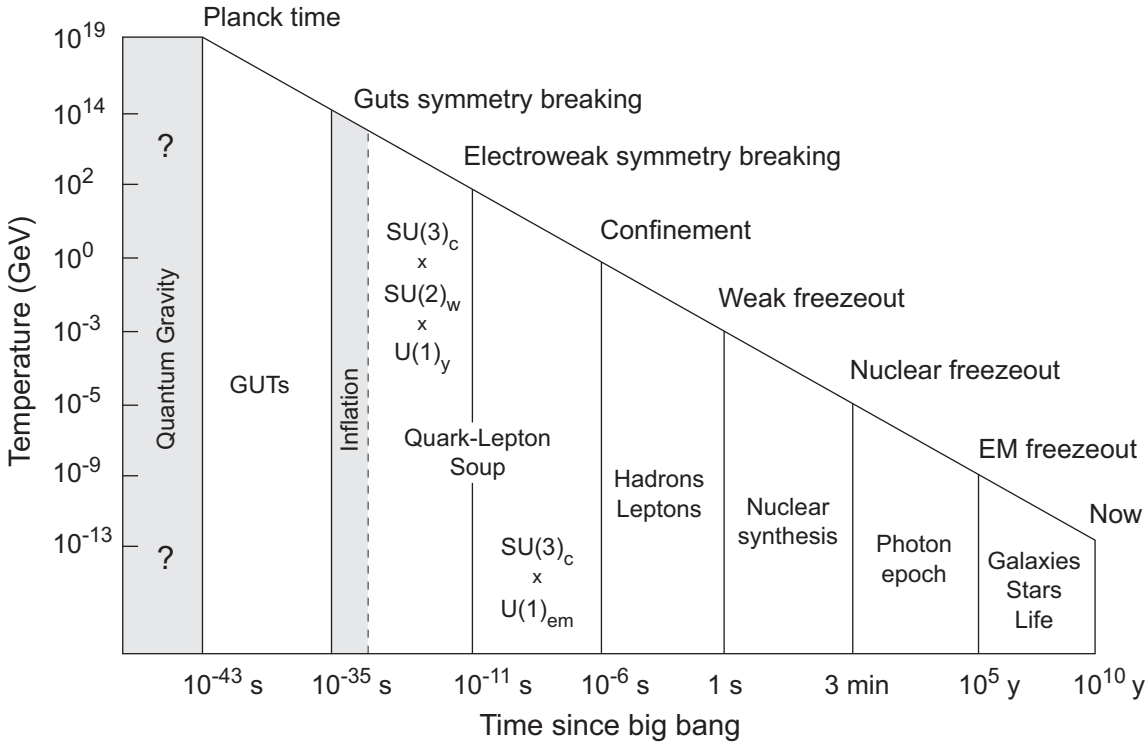


Figure 20.4: A history of the Universe. The time axis is highly nonlinear and 1 GeV $\simeq 1.2 \times 10^{13}$ K (after D. Schramm).

20.3.3 A Timeline for the Big Bang

The Friedmann equations allow us to *reconstruct the big bang*.

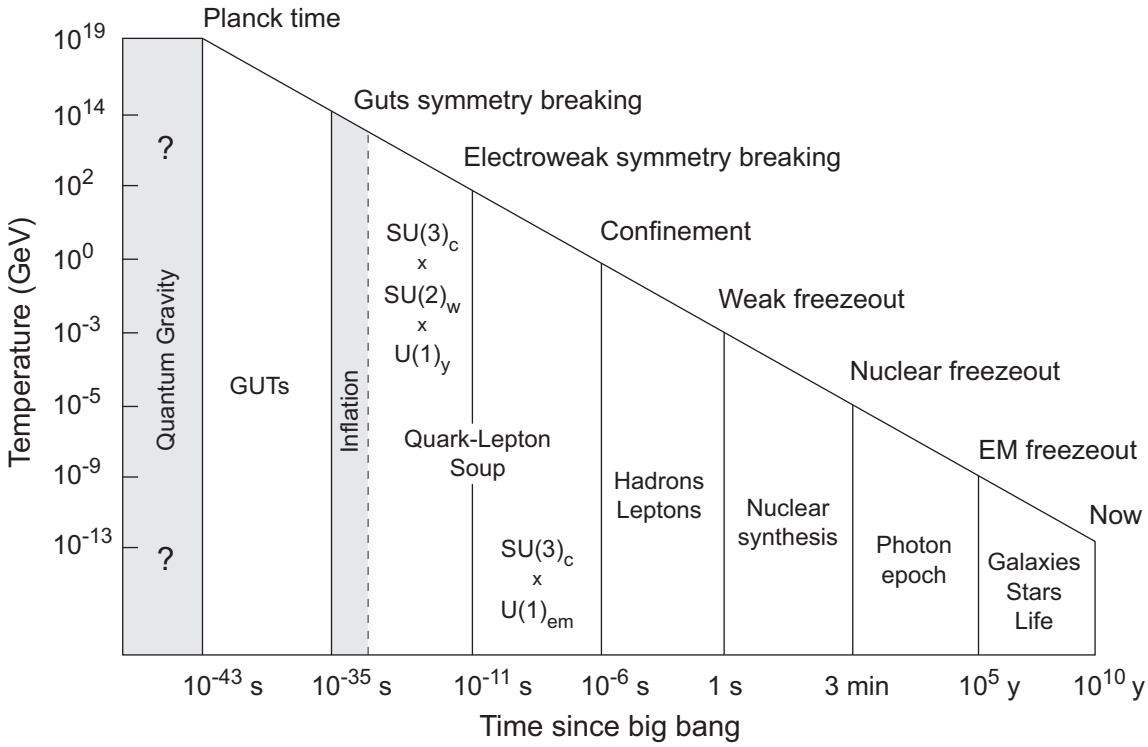
- Let's follow the sequence of events in terms of the time since the expansion began (see Fig. 20.4).
- The *primary cast of initial characters* includes:
  1. Photons
  2. Protons and neutrons
  3. Electrons and positrons
  4. Neutrinos and antineutrinos



Because of the equivalence of mass and energy, in a radiation dominated era

- the particles and their antiparticles are continuously undergoing reactions in which they annihilate each other, and
- photons can collide and create particle and antiparticle pairs.

Thus, under these conditions the radiation and the matter are in *thermal equilibrium* because they can freely interconvert.

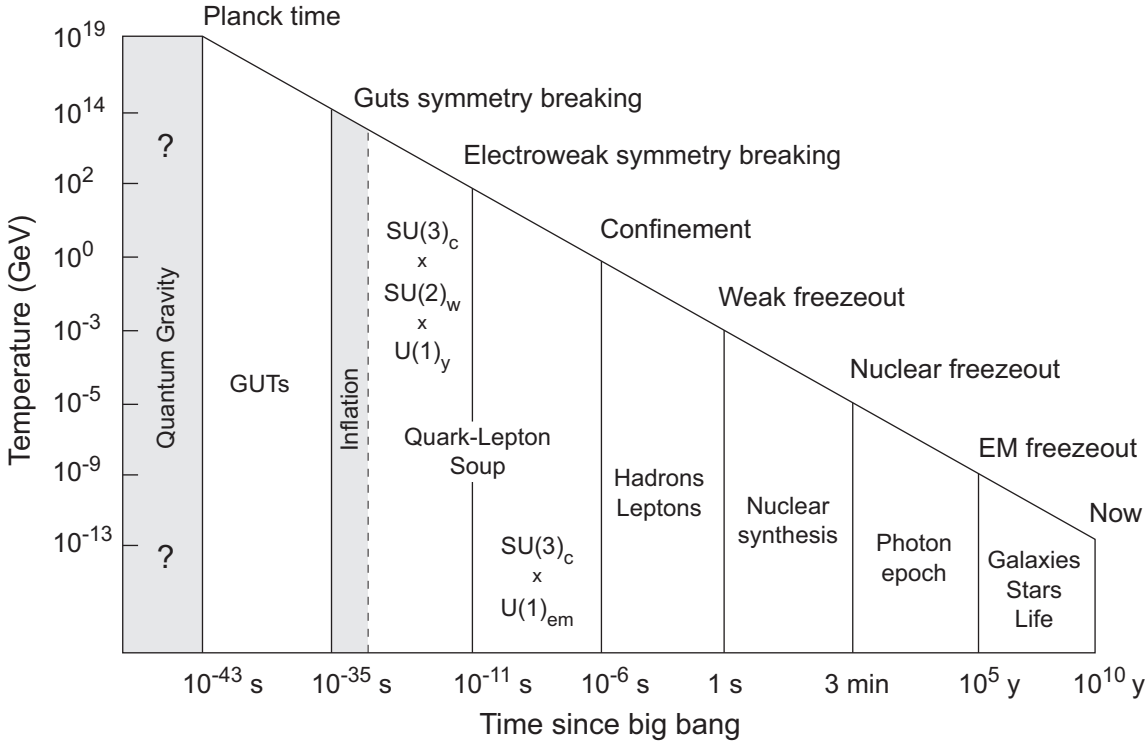


Time ~ 1 microsecond

The temperature is about 10^{13} K

- Quarks and gluons undergo a *confining transition* to form the hadrons.
- The Universe becomes *filled with protons and neutrons*, in about equal numbers.

This *quark–gluon to hadron phase transition* is labeled *"Confinement"* in the figure above.

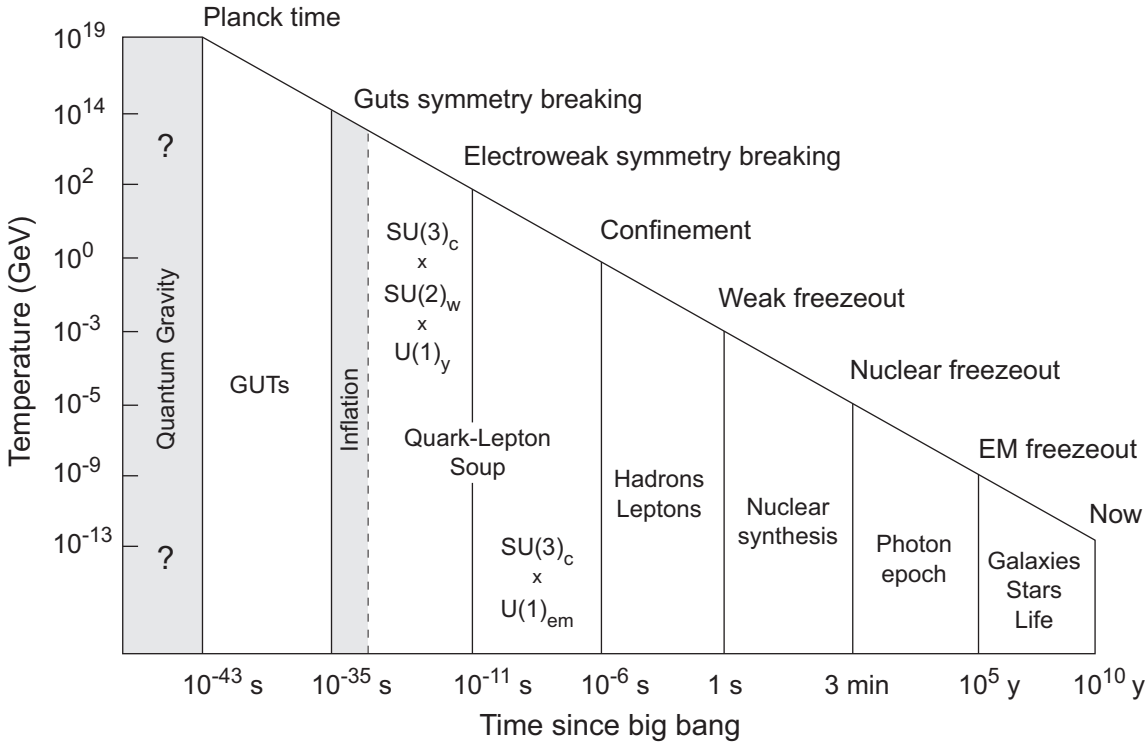


Time ∼ 0.01 seconds

- $T \simeq 10^{11}\,\mathrm{K}$.
- The Universe is expanding rapidly and consists of
  - a hot undifferentiated *soup of matter and radiation* in thermal equilibrium
  - with an *average particle energy* of $kT \simeq 8.6\,\mathrm{MeV}$.
- Equilibria:

$$e^- + e^+ \leftrightarrow \text{photons} \qquad \nu + \bar{\nu} \leftrightarrow \text{photons}$$

$$\bar{\nu} + p^+ \rightarrow e^+ + n \qquad \nu + n \rightarrow e^- + p^+ .$$



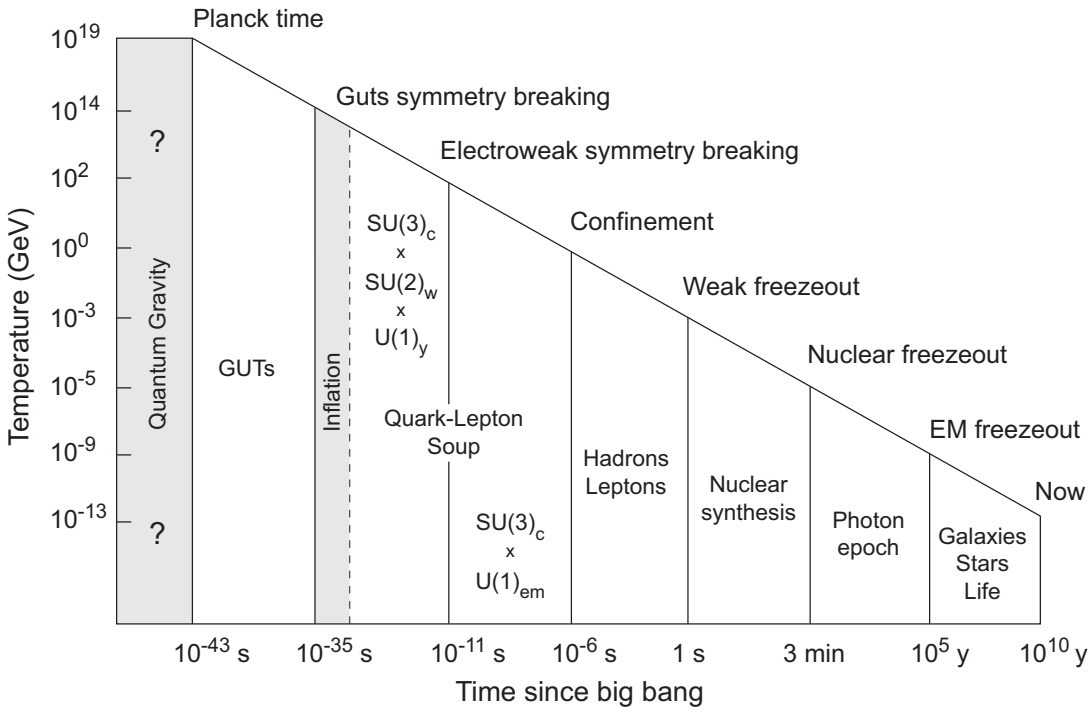
- A neutron is more massive than a proton by

$$Q \equiv (m_n - m_p)c^2 = 1.3\,\text{MeV}.$$

- Therefore, the equilibrium neutron–proton number density ratio is

$$\frac{n_n}{n_p} \sim \exp(-Q/kT).$$

- Since $kT \gg Q$ for $T = 10^{11}$ K, at this temperature $n_n \sim n_p$.

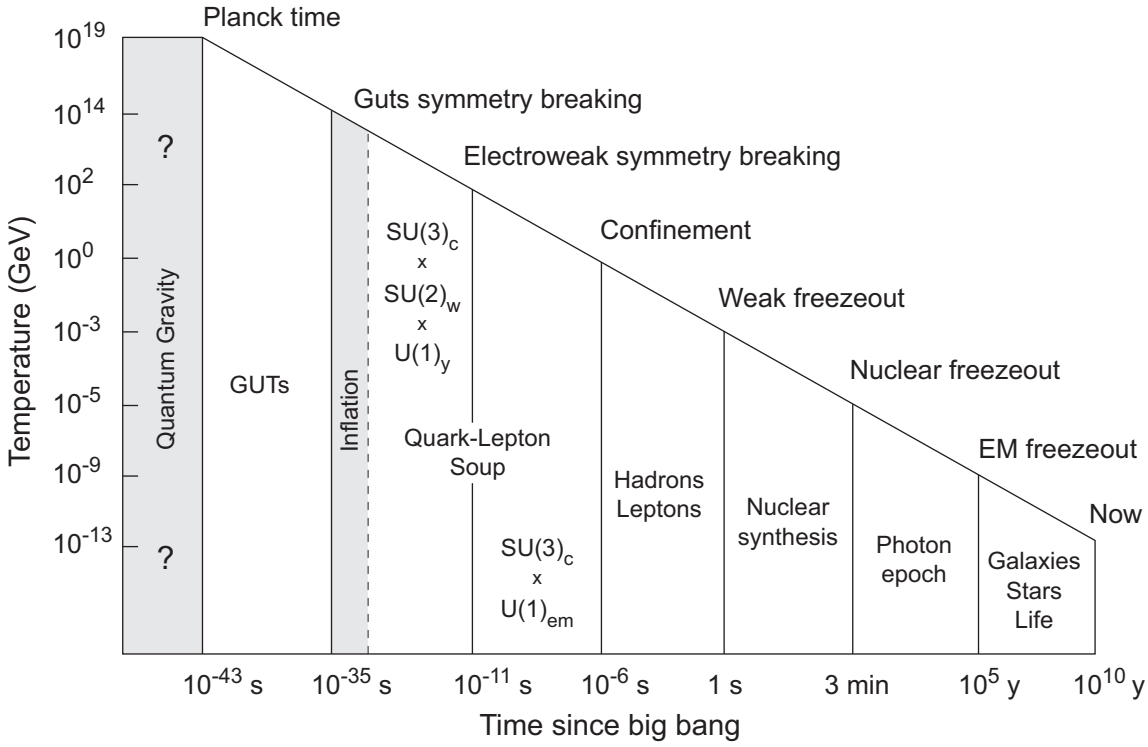


Time ∼ 0.1 seconds

- $T \simeq 3 \times 10^{10}\,\mathrm{K}$, corresponding to $kT \sim 2.6\,\mathrm{MeV}$.
- Thus the initial balance between neutrons and protons begins to be *tipped in favor of protons*, with now

$$\frac{n_n}{n_p} \sim 0.6.$$

- No composite nuclei can form yet because *deuterium has a binding energy* of only 2.2 MeV.
- Free neutrons will be converted steadily to protons until composite nuclei can form.
- Because there are *so many photons*, this requires kT an order of magnitude below the deuterium binding energy.

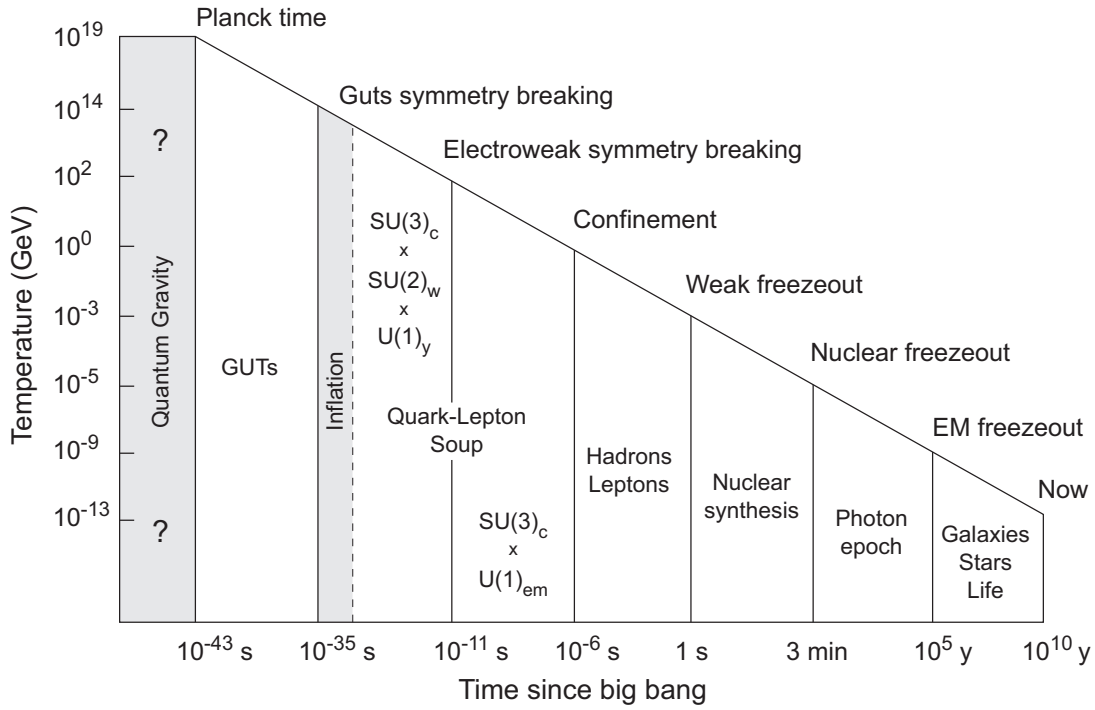


Time ∼ 1 second

- $T \simeq 10^{10}\,\mathrm{K}$, corresponding to $kT \simeq 0.8\,\mathrm{MeV}$.
- The *neutrinos drop out of equilibrium* and cease to play a role in the continuing evolution *("Weak freezeout")*.
- The temperature is still *too high for composite nuclei*.
- The neutron to proton ratio is now

$$\frac{n_n}{n_p} \sim 0.2,$$

and *continues to fall*.

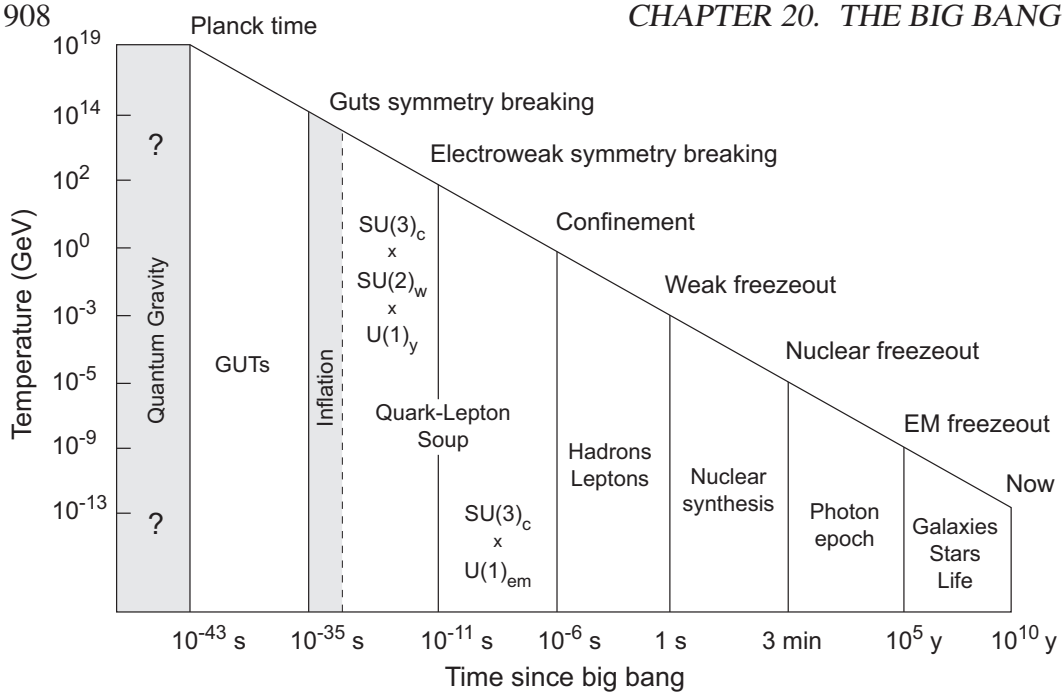


Time ∼ 10 seconds

- $T \sim$ several $\times 10^9$ K, corresponding to $kT \sim 0.25$ MeV.
- This is too low for photons to produce e^+e^- pairs, so they fall out of thermal equilibrium and
- the electrons annihilate the positrons to form photons:

$$e^- + e^+ \rightarrow \text{photons.}$$

- This *reheats all particles in thermal equilibrium with the photons* (not neutrinos, which dropped out of equilibrium at $t \sim 1$ s.)
- It is still *too hot to form composite nuclei* and n/p *continues to decrease* as free neutrons decay to protons.



Time ∼ 100–1000 seconds

- Finally the temperature drops sufficiently low (about 10^9 K) that *light composite nuclei can hold together*.
- The *neutron/proton ratio* has now fallen to $n_n/n_p \sim 0.15$.
- A rapid *sequence of nuclear reactions* ensues

$$n + p^+ \to {}^2_1\mathrm{H} \qquad {}^2_1\mathrm{H} + p^+ \to {}^3_2\mathrm{He} + n \to {}^4_2\mathrm{He}$$

$$ {}^2_1\mathrm{H} + n \to {}^3_1\mathrm{H} + p^+ \to {}^4_2\mathrm{He}$$

- and all remaining free neutrons are cooked rapidly into helium and trace amounts of a few other light elements.

The end of primordial nuclesynthesis is labeled *"Nuclear freezeout"* in the figure above.

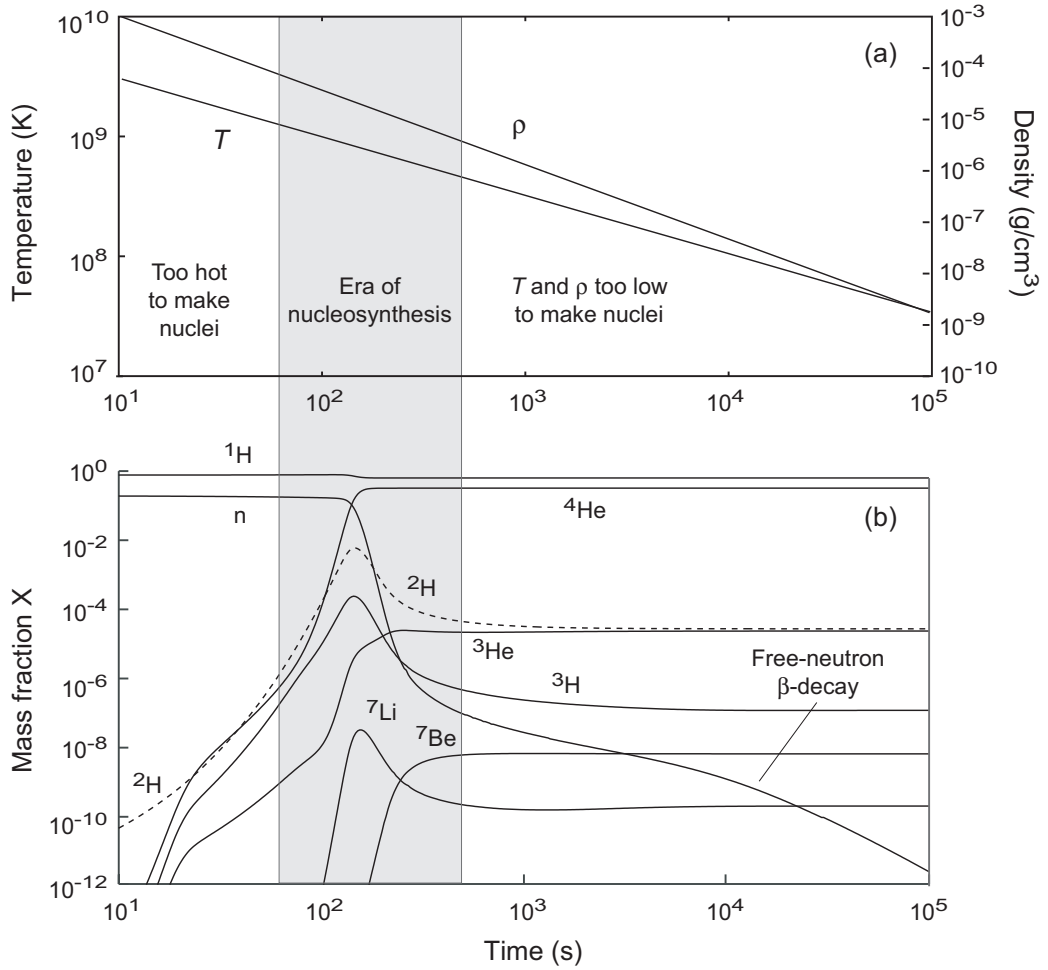
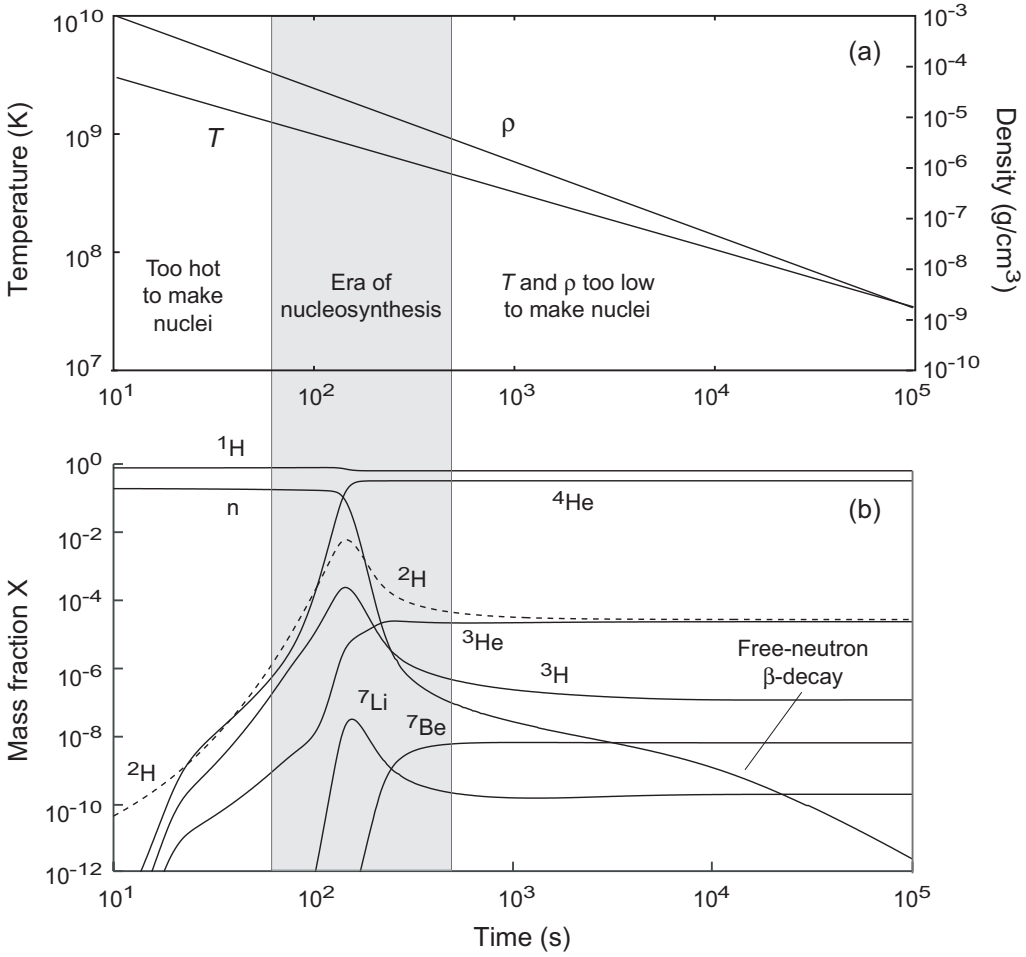


Figure 20.5: Big bang nucleosynthesis. (a) Variation of T and ρ assuming $\eta = 6 \times 10^{-10}$. (b) Mass fractions assuming mass fractions of 0.15 for neutrons and 0.85 for protons at the beginning of nucleosynthesis.

Production of the light elements occurs only in the narrow window (*era of nucleosynthesis*) illustrated in Fig. 20.5.

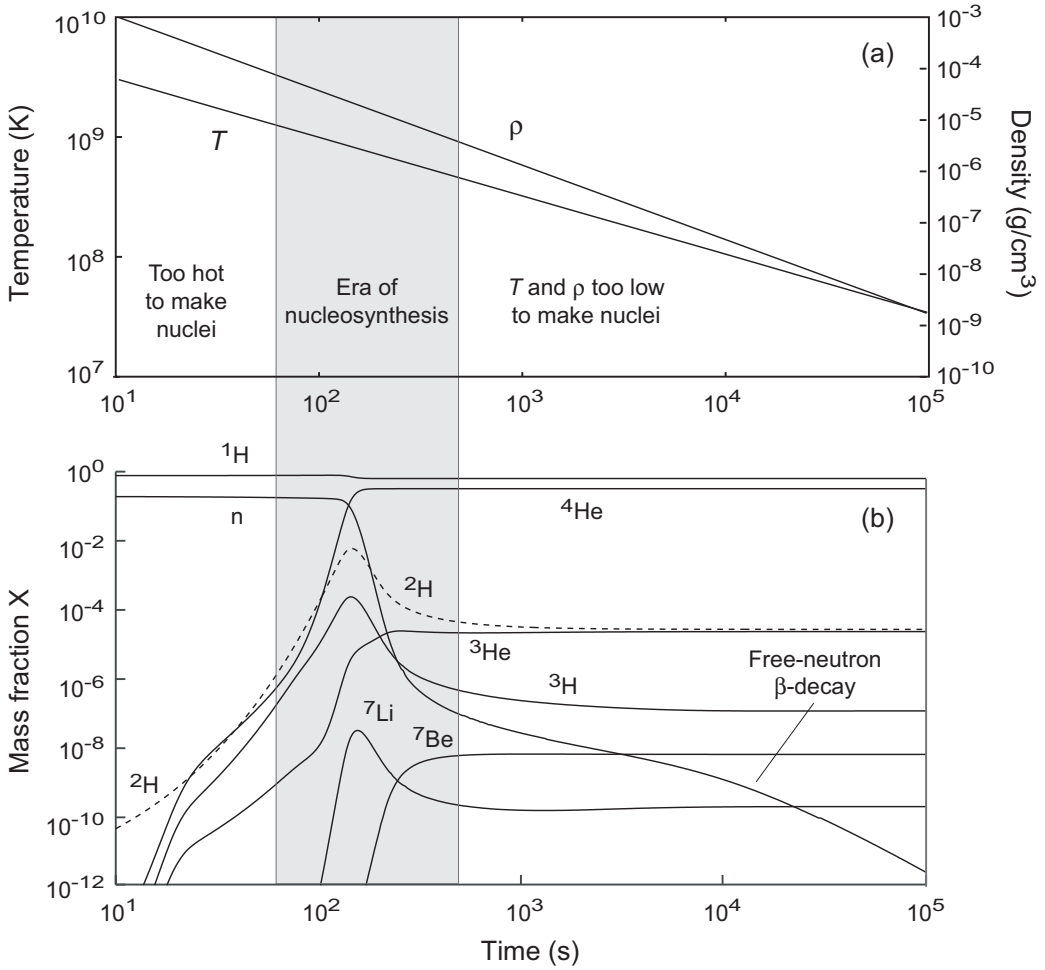
- In the era of nucleosynthesis *the temperature is*
  - *high enough* for nuclear reactions to occur, but
  - *low enough* to suppress photodisintegration of nuclei.



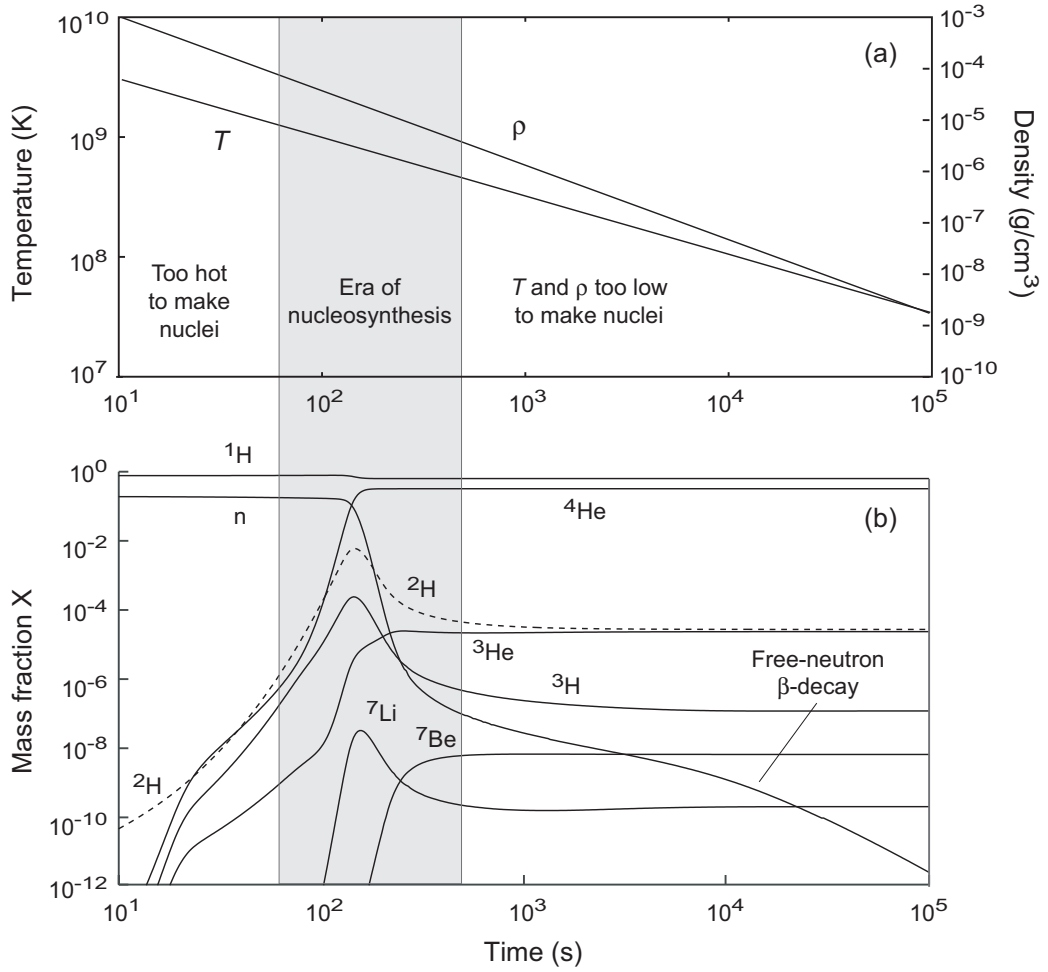
- Elements beyond ^{4}He aren't formed because
  - there are *no stable mass-5 or mass-8 isotopes*, and
  - the *density is too low* to make carbon by $3\alpha \rightarrow {}^{12}\mathrm{C}$.

In the era of nucleosynthesis $\rho \sim 10^{-5}\,\mathrm{g\,cm^{-3}}$.

- 100 times lower than the density of air.
- 10^{10} times lower than for 3α in *red giants*.

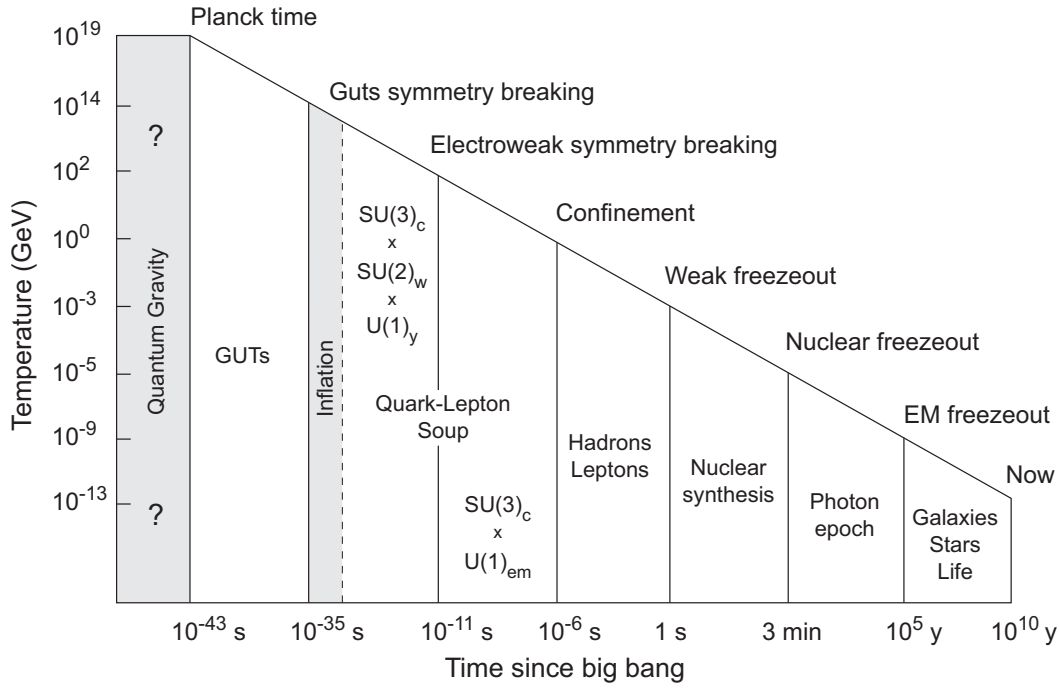


- A free neutron is unstable against β-decay into a proton with a halflife $t_{1/2} \sim 10.3$ minutes.
- Thus, once nucleosynthesis is complete any neutrons not bound up into composite nuclei *quickly β-decay into protons*.
- This is illustrated by the curve labeled *"Free-neutron β-decay"* in the figure above.



- The net effect of this short epoch of nucleosynthesis was to leave the Universe a few hundred seconds after its birth
  - consisting primarily of ^{1}H and ^{4}He,
  - with trace amounts of ^{2}H, ^{3}H, ^{3}He, ^{7}Be, and ^{7}Li.

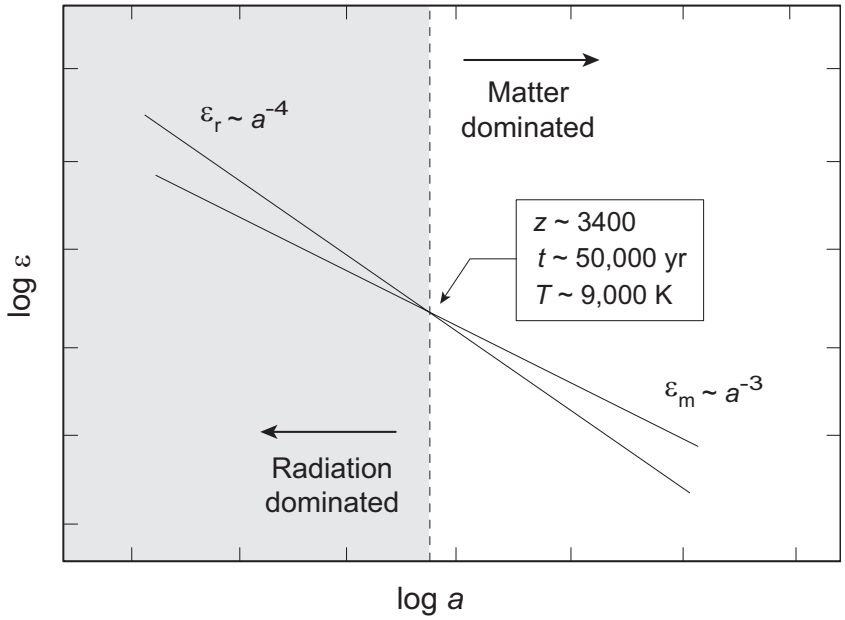
^{3}H is unstable against β^- decays to ^{3}He with a halflife of 12.3 yrs, so it doesn't stick round.



Time ∼ 30 minutes

- The temperature is now $T \sim 3 \times 10^8$ K.
- the Universe consists primarily of
  - protons,
  - excess electrons that didn't annihilate with positrons,
  - ^{4}He ($\sim 26\%$ abundance by mass), and
  - photons and neutrinos.
- Leftover neutrons continue to β-decay to protons and the *neutron abundance* has decreased now to $\sim 10^{-8}$.

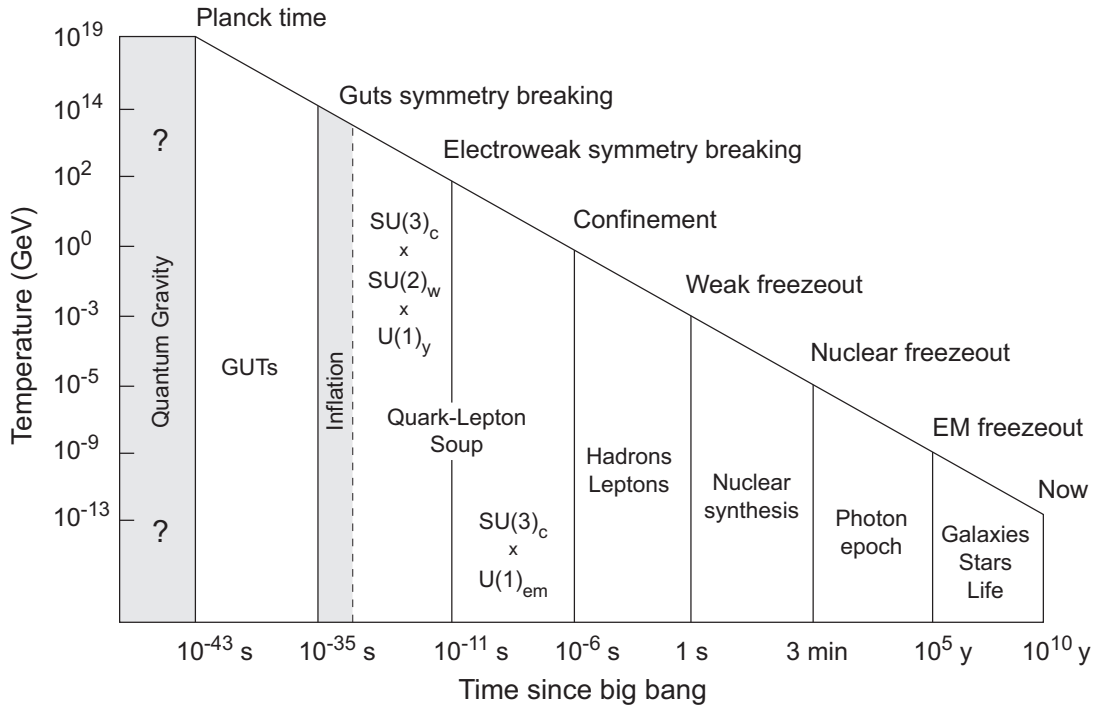
There are *no atoms yet* because the temperature is too high for protons and electrons to bind.



Time ∼ 50,000 years

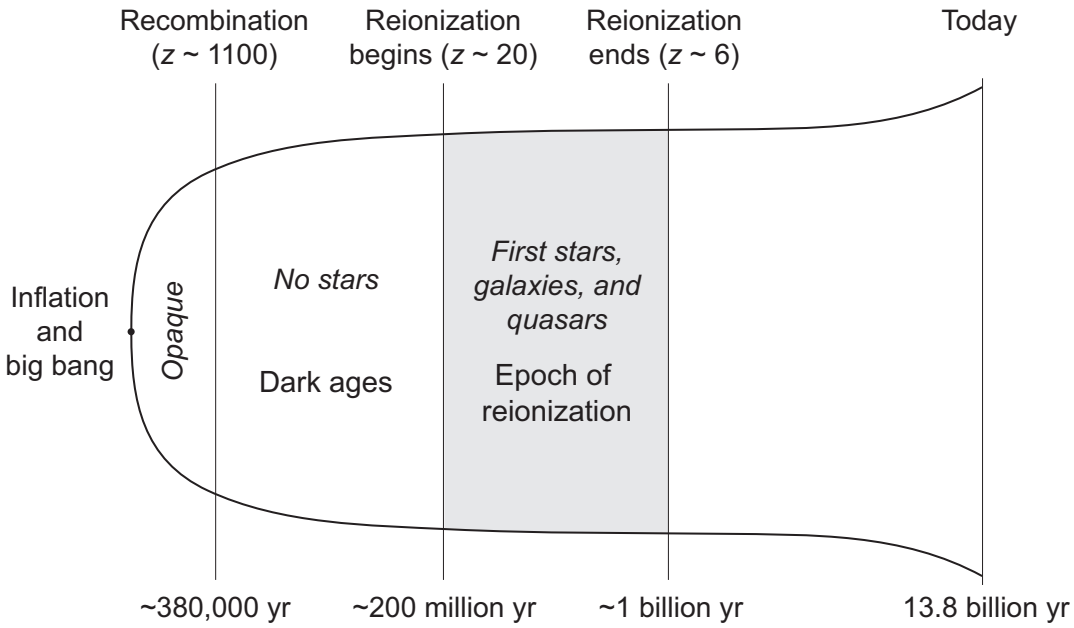
- Radiation dominated the early Universe (figure above).
- But the density of radiation ($\varepsilon_r \propto a^{-4}$) falls faster than the density of matter ($\varepsilon_m \propto a^{-3}$) as the Universe expands.
- Hence at some point ε_m *began to exceed* ε_m.
- Observations indicate that this occurred at $z \sim 3400$, corresponding to $T \sim 9,000\,\mathrm{K}$, at $t \sim 50,000$ years after the big bang.
- At this point the Universe began to be *matter dominated*.

Vacuum energy will become dominant much later, but it is *relatively insignificant at this stage*.



Time ∼ 400,000 years

- The temperature is now several thousand K, and *electrons and protons begin binding to form hydrogen atoms*.
- Until this point, matter and electromagnetic radiation have been in thermal equilibrium but *now they decouple* (labeled *"EM freezeout"* above).
- As free electrons are bound up in atoms in this *decoupling transition*, the primary cross section for photon scattering (Thomson scattering) is suppressed.
- The Universe, which had been opaque, *becomes transparent:* visible light can now freely travel large distances.



Time ∼ 400 million years

Observations like *spectra of high-z quasars* indicate that at some time after recombination *much of the hydrogen reionized*.

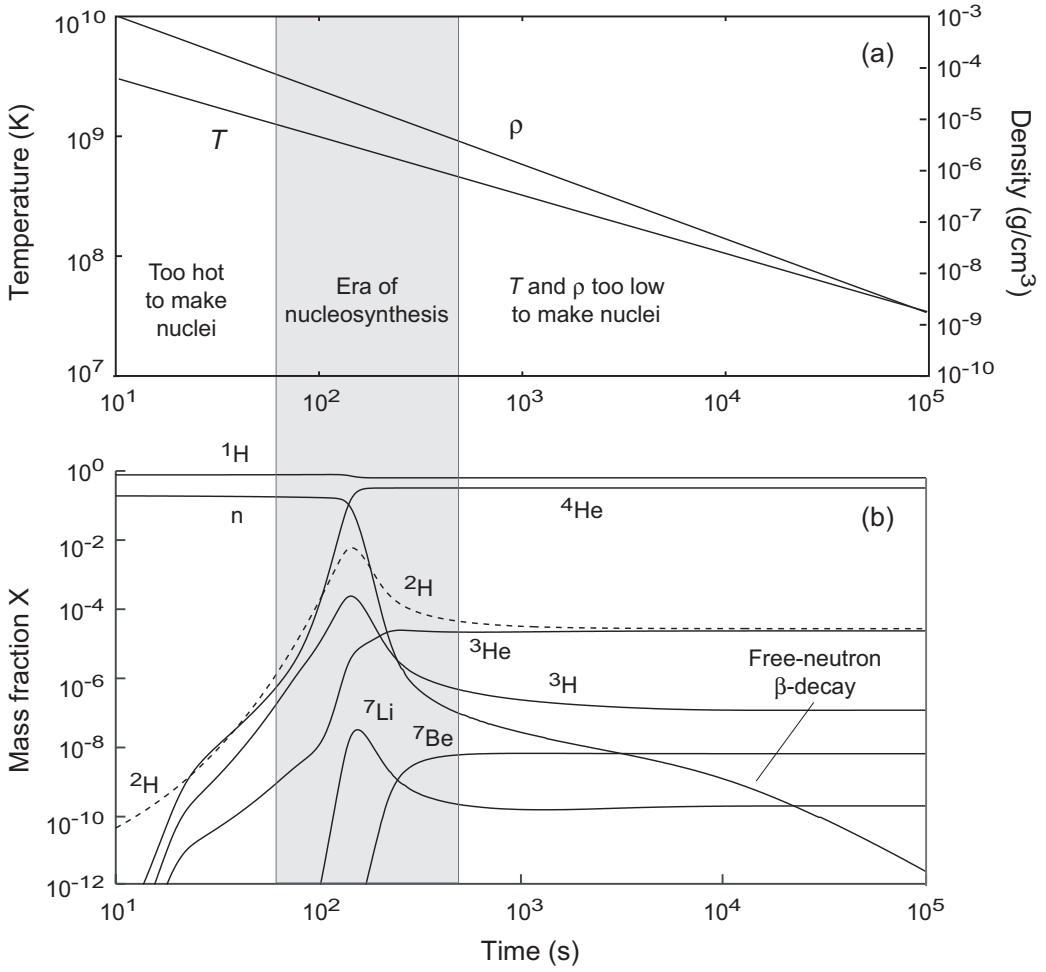
- This *reionization transition* occurred over about 800 million years between redshifts $z \sim 20$ and $z \sim 6$, and coincided with the *appearance of the first stars* (figure above).
- The period between the recombination transition and the appearance of the first stars is called the *dark ages*.
- Likely candidates for the reionizing agent were the first generation of stars (*Pop III*), and early *AGN and quasars*.
- The reionized Universe *remained relatively transparent* because expansion diluted the gas sufficiently to suppress collisions between photons and free electrons.

20.4 Nucleosynthesis and Cosmology

In *big bang nucleosynthesis*, ^{2}H can't form until T falls below a critical value, and then it is quickly converted into ^{4}He.

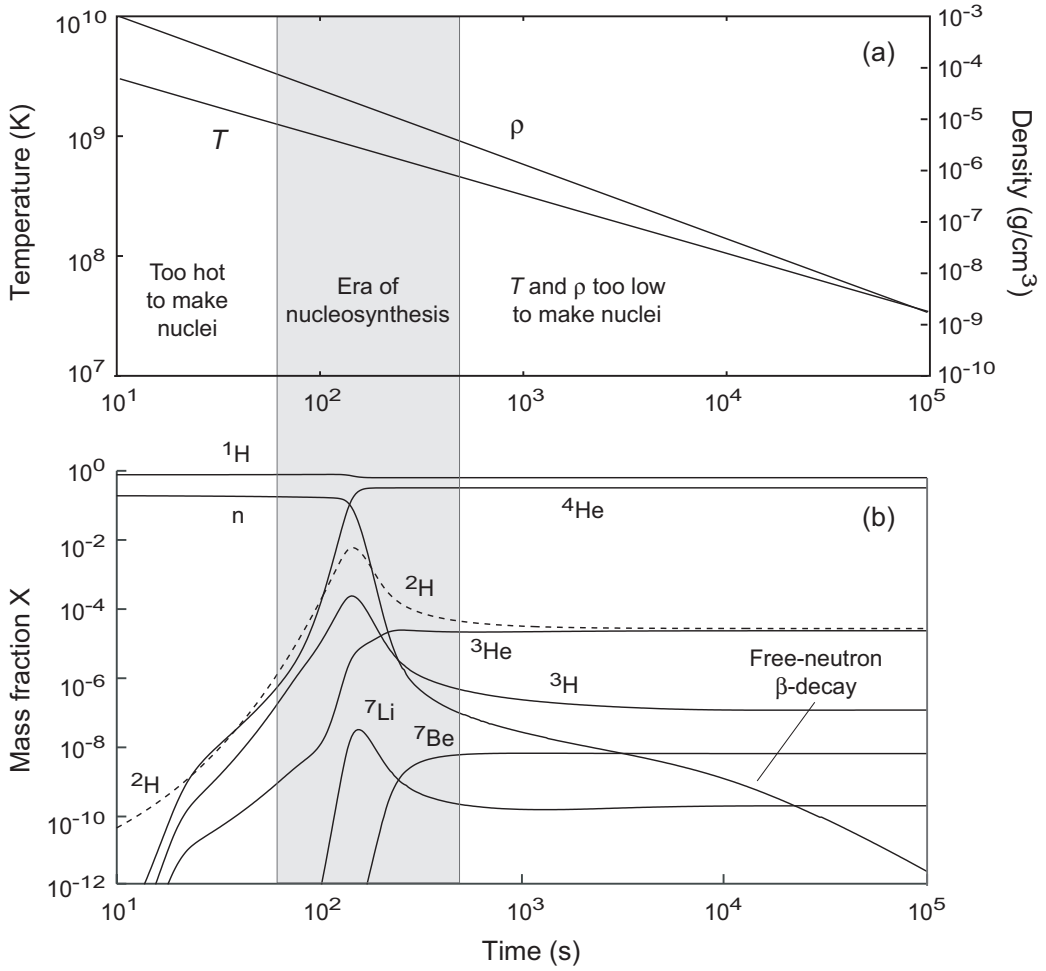
- Therefore, present abundances of ^{4}He and ^{2}H (and *other light elements* produced by the big bang in trace amounts)
- are a *sensitive probe of conditions in the early Universe*.
- The *oldest stars* contain material that is
- the *least altered* from that produced originally in the big bang.
- *Analysis of their composition* indicates abundances in very good agreement with predictions of the hot big bang.

This is *one of the strongest pieces of evidence* in support of the big bang theory.

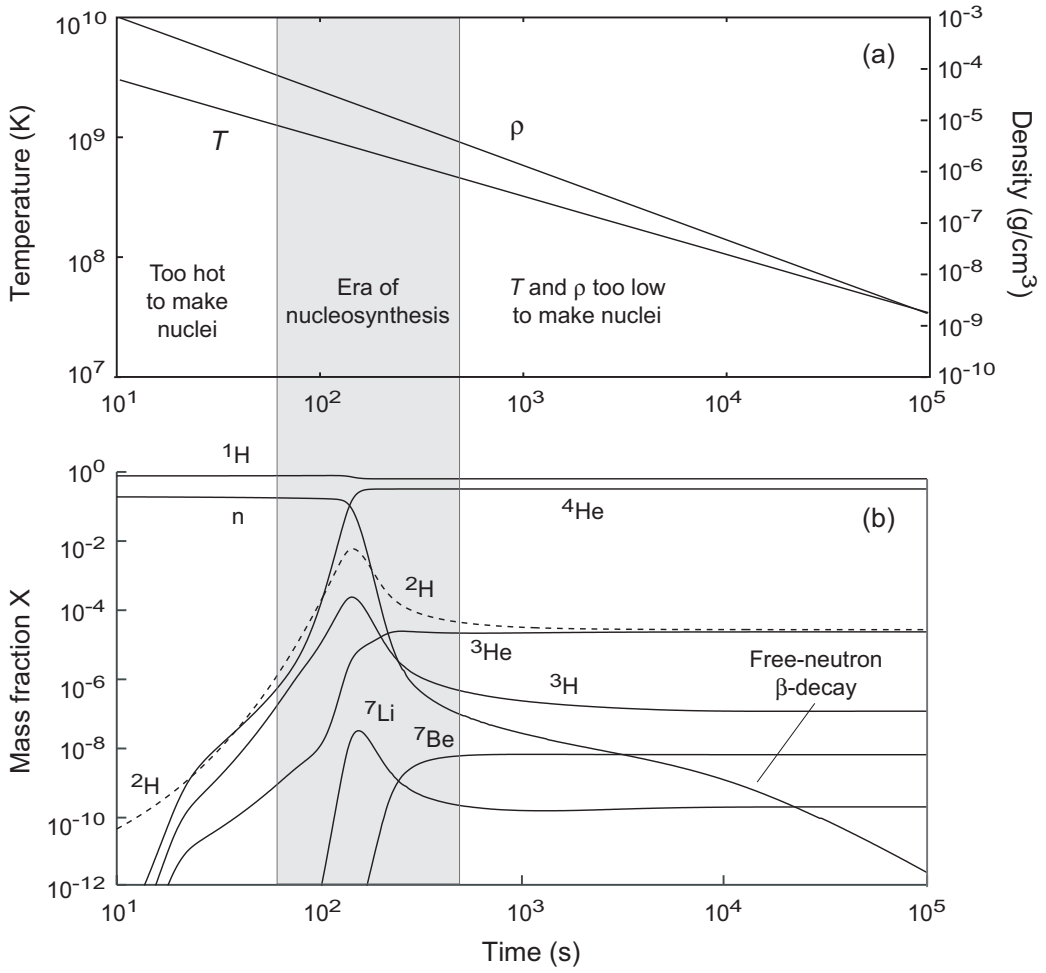


20.4.1 Elements Synthesized in the Big Bang

The outcome of big bang nucleosynthesis is conversion of neutrons and protons primarily to ^{4}He and free protons, with small concentrations of isotopes like ^{2}H, ^{3}He, and ^{7}Li.



- The specific abundance of each isotope is *sensitive to conditions in the short era of nucleosynthesis*.
- For the simulation above the *mass fractions are 0.253 for* ^{4}*He and 0.745 for* ^{1}*H*, once nucleosynthesis stops.
- The simulation is subject to some parameter uncertainty.
- However, it clearly suggests that the big bang should have left the Universe with about *a quarter helium by mass*, with almost *all the rest hydrogen*.



The simulation displayed above indicates that

- the big bang left the Universe with about *25% He by mass*, with almost *all the rest hydrogen*.
- That is rather *close to the current measured mass fractions* for helium and hydrogen.
- This is an *important general test*, since processes occurring later in stars cannot change these mass fractions by too much.

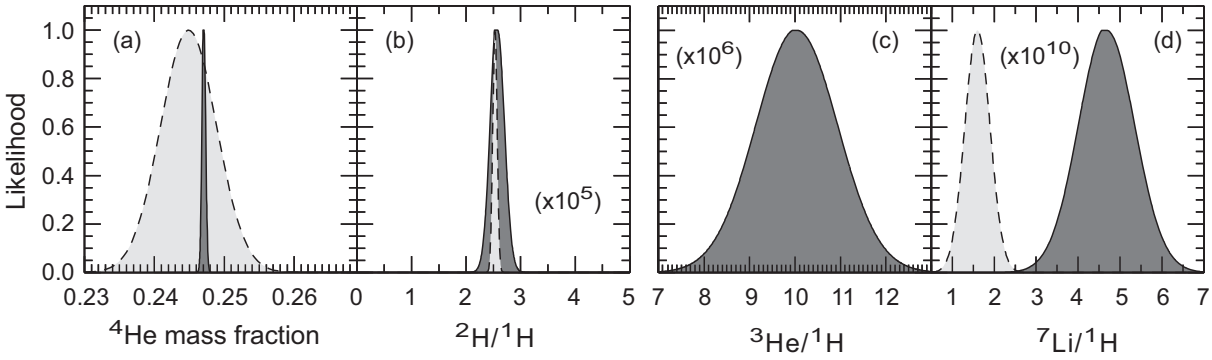


Figure 20.6: Likelihoods normalized to unity for abundances of light elements. (a) Mass fraction for ^{4}He; (b) Ratio of ^{2}H (deuterium) to ^{1}H abundance; (c) Ratio of ^{3}He to ^{1}H abundance; (d) Ratio of ^{7}Li to ^{1}H abundance. Dark gray with solid lines are predictions based on big bang nucleosynthesis and CMB analysis. Light gray with dashed lines indicates primordial abundances from astronomical observations, except for ^{3}He where no reliable measurements exist. The agreement between big bang theory (dark gray) and observations (light gray) is very good except for ^{7}Li, where there is a factor of three discrepancy.

Fig. 20.6 illustrates a *comparison of calculated and observed abundances* for the light elements produced in the big bang.

- Agreement between observation and theory is excellent, except for ^{7}Li, where there is a factor of three discrepancy.
- The ^{7}Li discrepancy is not understood at present.
- It occurs for the isotope with (by far) the tiniest abundance.
- Thus it is a small effect that might be explained by observational issues, or by missing physics having a small impact on overall nucleosynthesis.

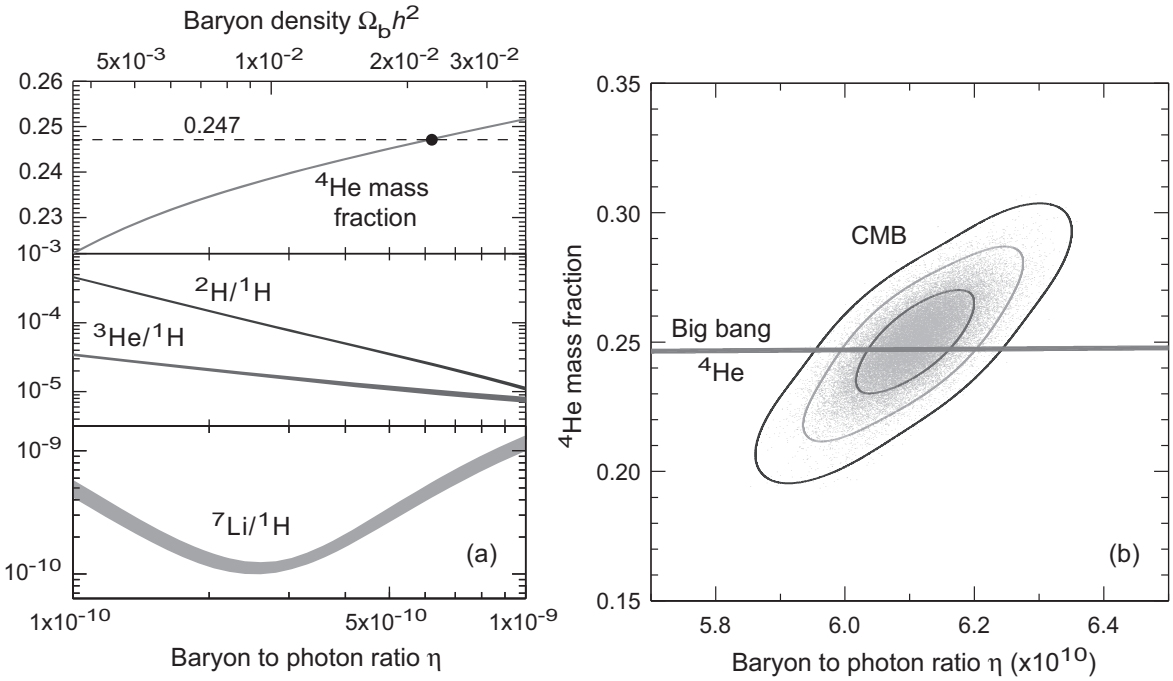
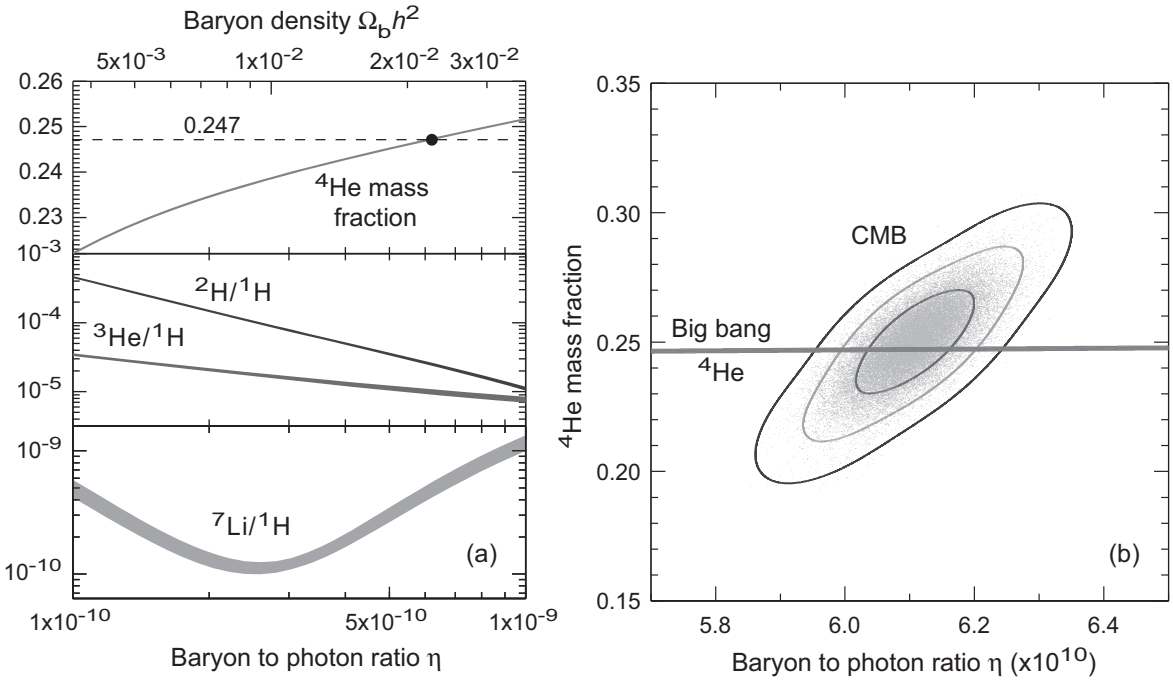


Figure 20.7: (a) Big bang abundances as a function of the baryon to photon ratio η or baryon density parameter Ω_b. (b) Correlation of helium abundance and η from Planck satellite CMB data. The ^{4}He mass fraction predicted from big bang nucleosynthesis is indicated by the near-horizontal curve. Both plots suggest $\eta \sim 6.1 \times 10^{-10}$, or equivalently $\Omega_b \sim 0.04$.

Figure 20.7(a) illustrates calculated abundances for big bang light elements as a function of

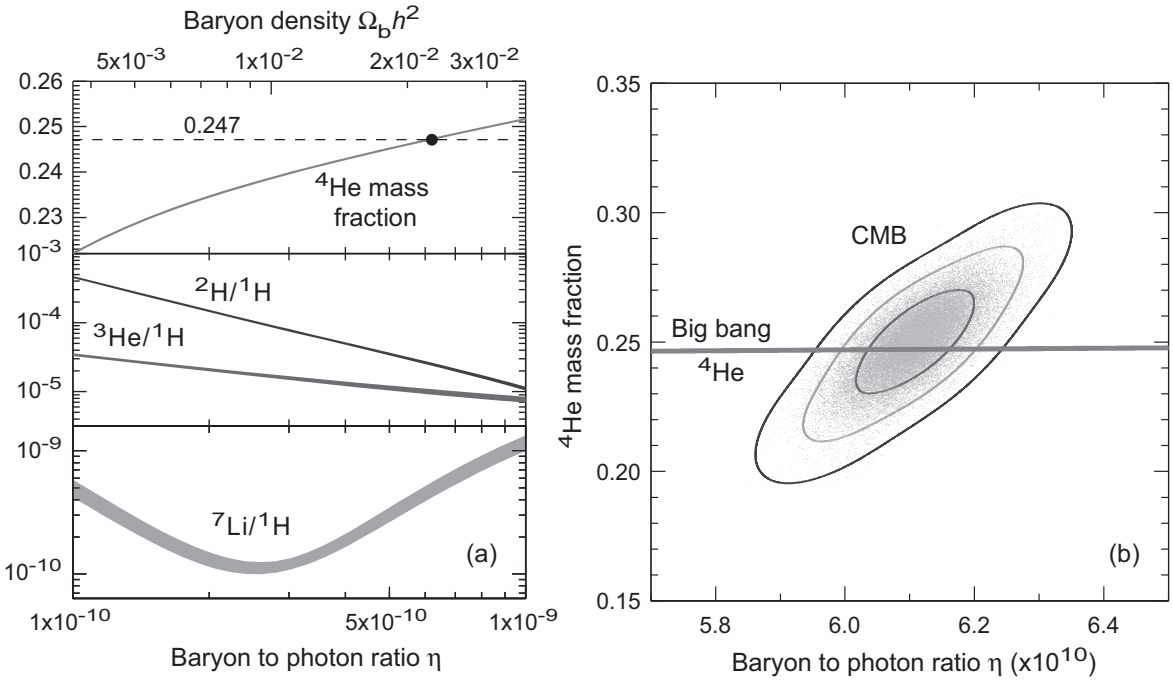
- the current *baryon to photon ratio* η, or equivalently
- the current *baryon density parameter* Ω_b.
- Widths of curves indicate theoretical uncertainties.
- By comparing calculated and observed abundances, we may (1) *test big bang nucleosynthesis* and (2) determine the *baryon to photon ratio*.



- The analysis indicates that the *mass fraction of* ^{4}*He is* 0.247 [dashed horizontal line in Fig. (a) above].
- Intersection of this curve with the ^{4}He curve implies $\eta \sim 6 \times 10^{-10}$.
- As illustrated in Fig. (b) above, independent analysis of the cosmic microwave background by the Planck satellite finds a *similar value of* η.
- From these results it may be concluded that

$$\eta \equiv \frac{n_{\rm b}}{n_\gamma} \simeq 6.1 \times 10^{-10},$$

implying *several billion photons for every baryon*.



- Most of these baryons are neutrons and protons, while
- most of the photons are in the cosmic microwave background radiation.
- The total number of each kind of particle is not expected to change much in the absence of interactions.

Thus the current ratio η is approximately the value of η at the time when matter and radiation decoupled.

20.4.2 Constraints on Baryon Density

Agreement between theory and observation for light-element abundances also *constrains the baryonic mass of the Universe*.

- That constraint is the basis for our earlier assertion that *most dark matter cannot be baryonic*.
  - If enough baryons were present in the Universe to make the dominant matter baryonic, and
  - our understanding of the big bang is anywhere near correct,

  the *distribution of light element abundances* would have to *differ substantially from that observed*.

The implication is that the matter we are made of *(baryonic matter)* is but a small impurity compared to the dominant matter of the Universe *(non-baryonic matter)*.

20.5 The Cosmic Microwave Background

There are *three important observables* in the present Universe that presumably date back to its early history:

- Abundance of the light elements
- The cosmic microwave background (CMB) radiation
- Dark matter

> We have just discussed formation of the light elements. The remainder of this chapter will emphasize the nature of the CMB and dark matter.

The *cosmic microwave background (CMB)* is the faint glow left over from the big bang itself.

- It was discovered accidentally by Penzias and Wilson in 1964 while testing a new microwave antenna.
- They initially believed the signal that they detected coming from all directions to be electronic noise.
- Once careful experiments had ruled that possibility out, they were initially unaware of the significance of their discovery.
- Then it was pointed out that the big bang theory *predicted* that the Universe should be permeated by radiation left over from the big bang itself.
- But now the radiation would have been *redshifted by the expansion* over some 14 billion years to the microwave spectrum.

Dark matter appears to represent the major part of the mass in the Universe, but we don't yet know what it is.

Both the CMB and the nature of dark matter provide crucial diagnostics for a fundamental issue in cosmology, the *formation of large-scale structure*.

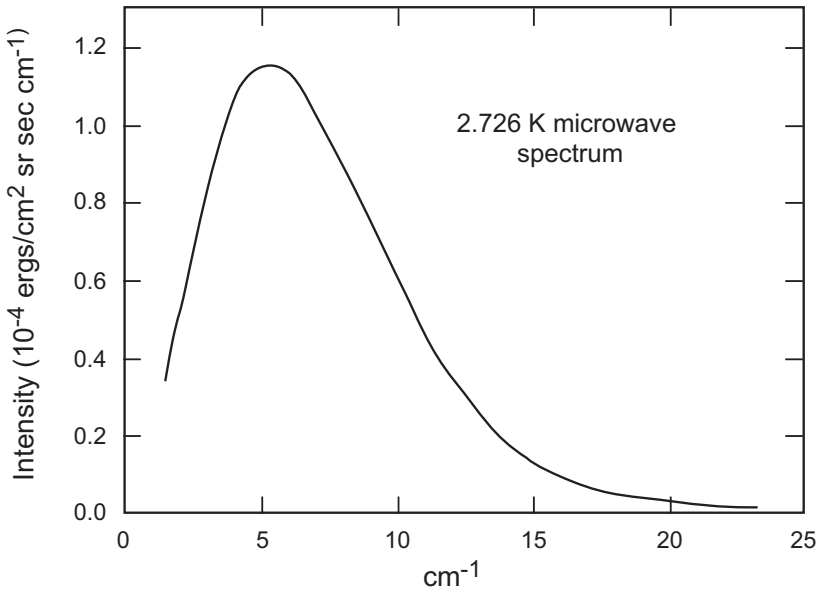


Figure 20.8: The 2.726 K microwave background spectrum (COBE).

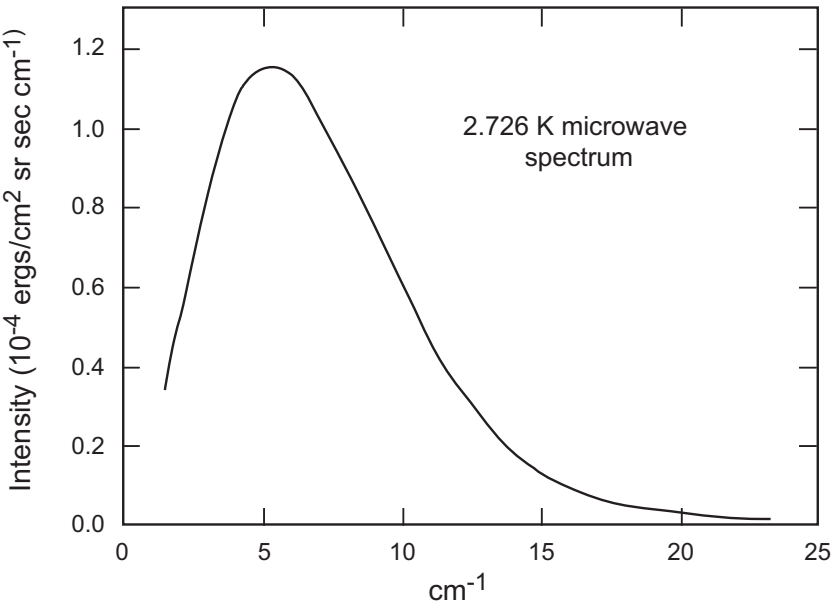
20.6 The Microwave Background Spectrum

Measurements by Penzias and Wilson that are relatively crude by modern standards established that

- The radiation was *coming from all directions in the sky*, and had a blackbody temperature of $T \sim 3\,\mathrm{K}$.
- Later, precise measurements from the *Cosmic Background Explorer (COBE)* satellite found an almost *perfect blackbody spectrum* having average temperature

$$\langle T \rangle \equiv \frac{1}{4\pi} \int T(\theta,\varphi) \sin\theta d\theta d\varphi = 2.726\,\mathrm{K},$$

with $T(\theta,\varphi)$ the temperature measured at coordinates (θ,φ) on the sky; Fig. 20.8 illustrates.



- Data points and the theoretical curve for a 2.726 K spectrum are *indistinguishable* (figure above).
- This is the *best blackbody spectrum ever measured*.
- The temperature was revised to 2.275 ± 0.001 by the *Wilkensen Microwave Anisotropy Probe (WMAP)*.
- By applying basic statistical mechanics to the observed spectrum, we may deduce a *CMB photon density* of

$$N_\gamma \simeq 410 \,\text{photons}\,\text{cm}^{-3}$$

Theory predicts also a *cosmic neutrino background* remnant of the big bang, but these low-energy neutrinos are *not presently detectable*.

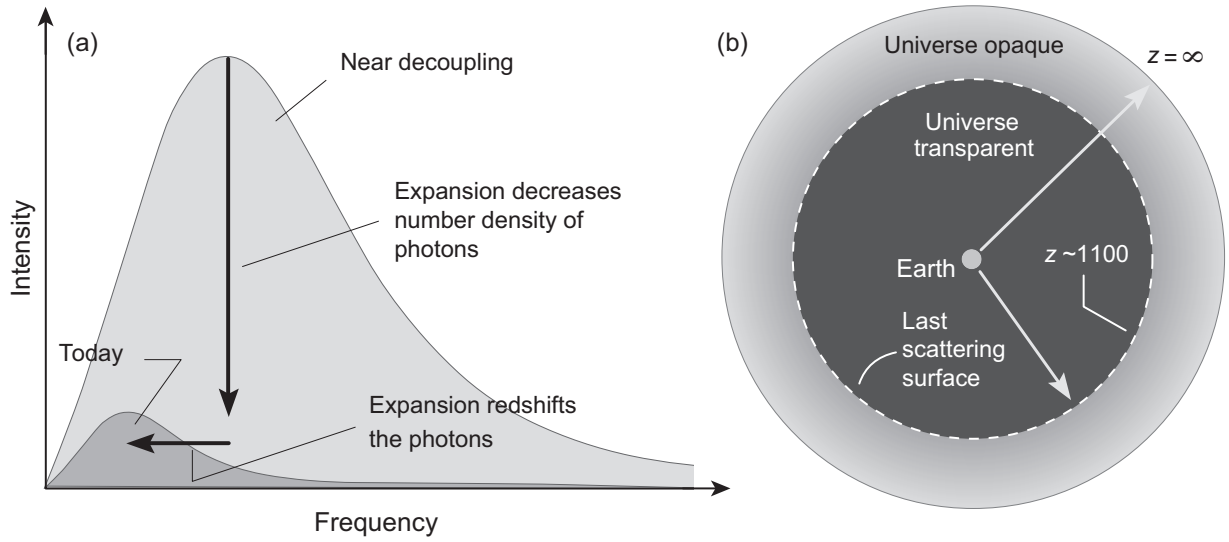


Figure 20.9: (a) Schematic evolution of the cosmic microwave background. As the Universe expands the spectrum remains blackbody but the photon frequencies are redshifted and the number density of photons is lowered. The 2.7 K CMB is the faint, redshifted remnant of the cosmic fireball in which the Universe was created. Decoupling occurred at $z \sim 1100$. The photon temperature then of about 3000 K is lowered by the redshift factor to the present value of a little less than 3 K. (b) Last scattering surface for the CMB. The Universe is opaque to photons at larger redshift (earlier times).

The CMB is the photon remnants of the big bang, redshifted into the microwave spectrum by expansion [Fig. 20.9(a)].

- The photons detected in the CMB were emitted from the *last scattering surface (LSS)* illustrated in Fig. 20.9(b).
- The LSS lies at $z \sim 1100$.
- This represents the time when the present CMB photons decoupled from matter ($\sim 400,000$ yr after the big bang).
- At greater z the Universe was opaque to photons because then *matter and radiation were strongly coupled*.

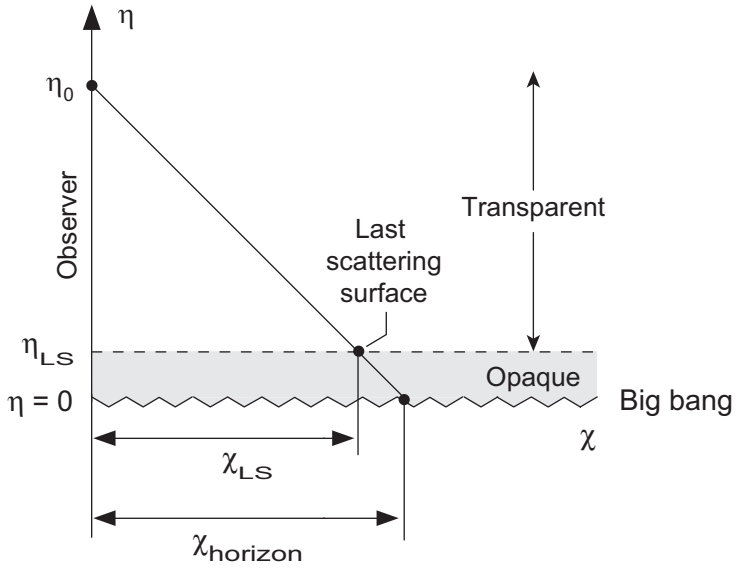


Figure 20.10: The CMB last scattering surface illustrated in a $\eta - \chi$ space-time diagram, where χ is a comoving coordinate and η is a conformal time coordinate. The present time is η_0, the time of the big bang is $\eta = 0$, and the time of last scattering is $\eta_{\rm LS}$.

The *last scattering surface* is illustrated further in Fig. 20.10 using *conformal time* η.

- The comoving distance to the last scattering surface $\chi_{\rm LS}$ is the *greatest actual distance* from which photons can be detected.
- It is *less than the horizon distance* $\chi_{\rm horizon}$.
- The Universe is
  - opaque to photons before $\eta_{\rm LS}$ and
  - transparent afterwards.

20.7 Anisotropies in the Microwave Background

COBE, *WMAP*, and *Planck* satellites (as well as *high-altitude balloons*) have measured the *angular distribution of the CMB*.

- It is isotropic down to a dipole anisotropy $\sim 10^{-3}$.
- This corresponds to a *Doppler shift* associated with motion of the Earth relative to the microwave background.
- Once the peculiar motion of the Earth with respect to the CMB is subtracted, the *background is isotropic* down to the 10^{-5} level (tens of μK).
- The *temperature fluctuation* at a given point on the sky is defined by

$$\frac{\delta T}{T}(\theta,\varphi) \equiv \frac{T(\theta,\varphi) - \langle T\rangle}{\langle T\rangle} \qquad \langle T\rangle \equiv \frac{1}{4\pi}\int T(\theta,\varphi)\sin\theta d\theta d\varphi$$

- *COBE measured an anisotropy* that corresponds to

$$\sqrt{\left\langle \left(\frac{\delta T}{T}\right)^2 \right\rangle} = 1.1\times 10^{-5}.$$

Even more precise CMB anisotropies have been measured by WMAP and Planck, as illustrated on the following page.

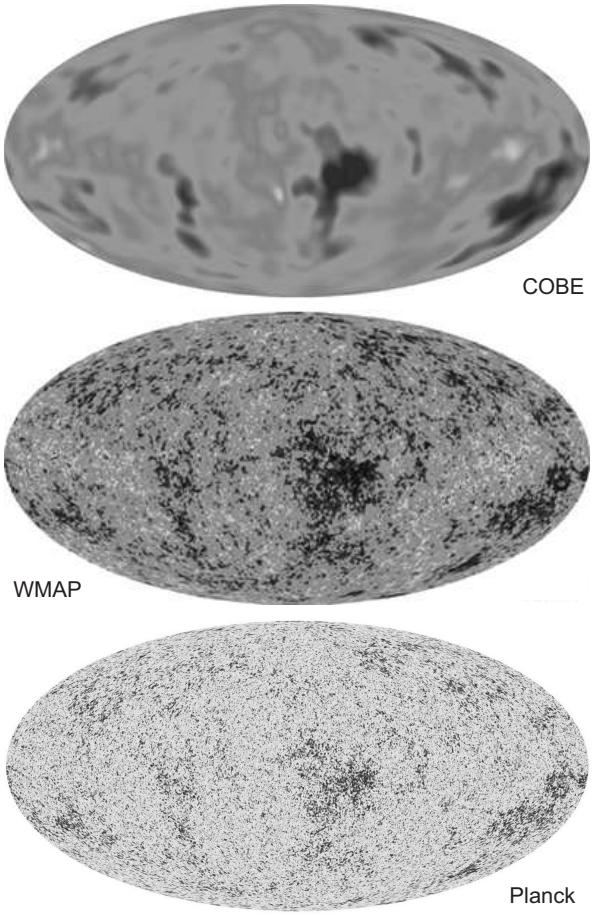


Figure 20.11: COBE, WMAP, and Planck fluctuation maps of the full sky.

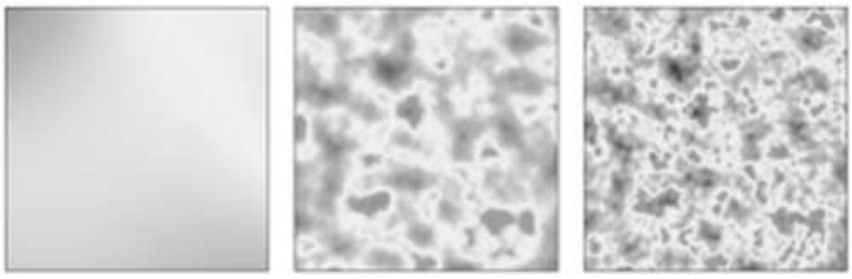


Figure 20.12: COBE, WMAP, and Planck maps of a 10 degree square. Note the dramatic difference in resolution.

The primary interest in cosmology is in *statistical correlations of the CMB temperature fluctuations*.

- It is convenient to expand the temperature fluctuations as a function of angular position (θ, φ) on the sky using *spherical harmonics* $Y_{\ell m}(\theta, \varphi)$,

$$\frac{\delta T}{T}(\theta, \varphi) = \sum_{\ell m} a_{\ell m} Y_{\ell m}(\theta, \varphi).$$

- The *2-point correlation function* $C(\theta)$ is then defined by,

$$C(\theta) \equiv \left\langle \frac{\delta T}{T}(\hat{\boldsymbol{n}}_1) \frac{\delta T}{T}(\hat{\boldsymbol{n}}_2) \right\rangle_{\hat{\boldsymbol{n}}_1 \cdot \hat{\boldsymbol{n}}_2 = \cos\theta},$$

  - where $\hat{\boldsymbol{n}}_1$ and $\hat{\boldsymbol{n}}_2$ are *unit vectors specifying the directions to two points on the celestial sphere* separated by an angle θ, and
  - the angular brackets indicate an *average over all pairs of points* separated by θ.

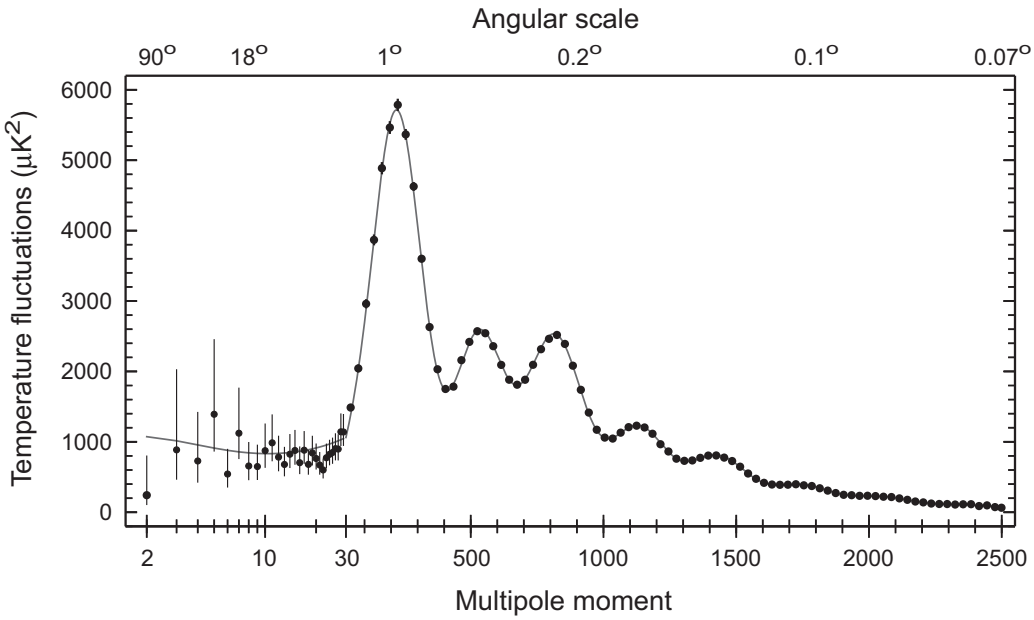


Figure 20.13: Planck CMB power spectrum. Multipole order ℓ on the bottom axis (with a log scale below 30 and linear above) and a corresponding angular scale θ on the top axis. They are related by $\theta \sim 180^\circ/\ell$.

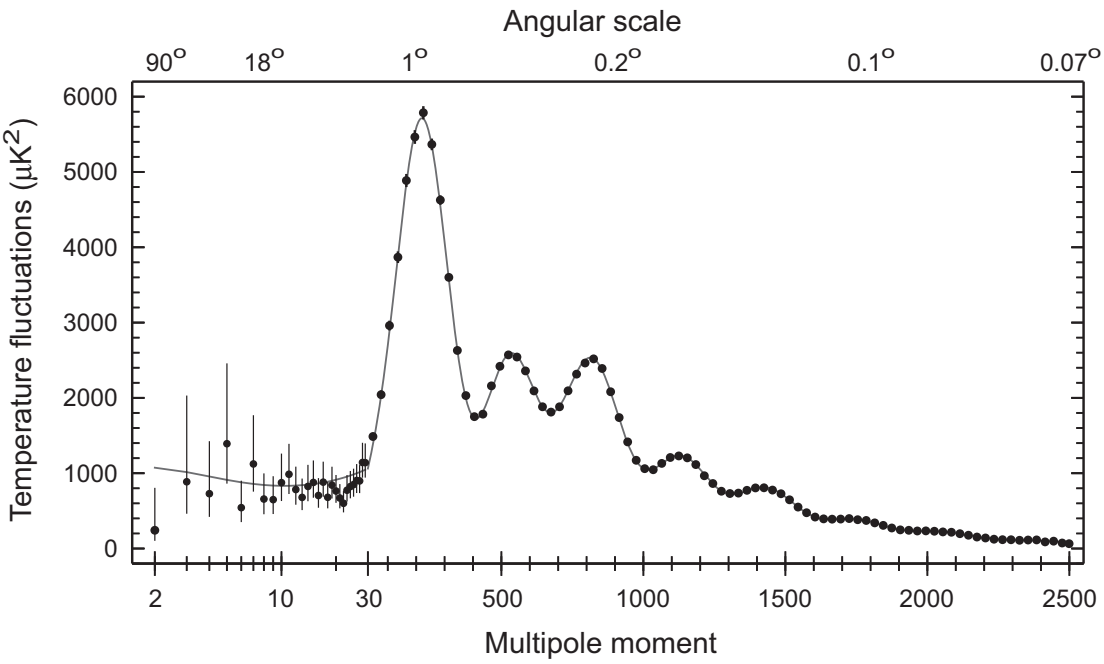
- *Inflation* predicts fluctuations that are *maximally random (*gaussian*)*, meaning that *multipoles are uncorrelated*.
- Assuming *gaussian fluctuations* and the *addition theorem*

$$P_\ell(\cos\theta) = \frac{4\pi}{2\ell+1}\sum_{m=-\ell}^{m=+\ell} Y^*_{\ell m}(\hat{\boldsymbol{n}}_1)Y_{\ell m}(\hat{\boldsymbol{n}}_2),$$

with θ the angle between $\hat{\boldsymbol{n}}_1$ and $\hat{\boldsymbol{n}}_2$, the correlation function may be expanded as *Legendre polynomials* $P_\ell(\cos\theta)$,

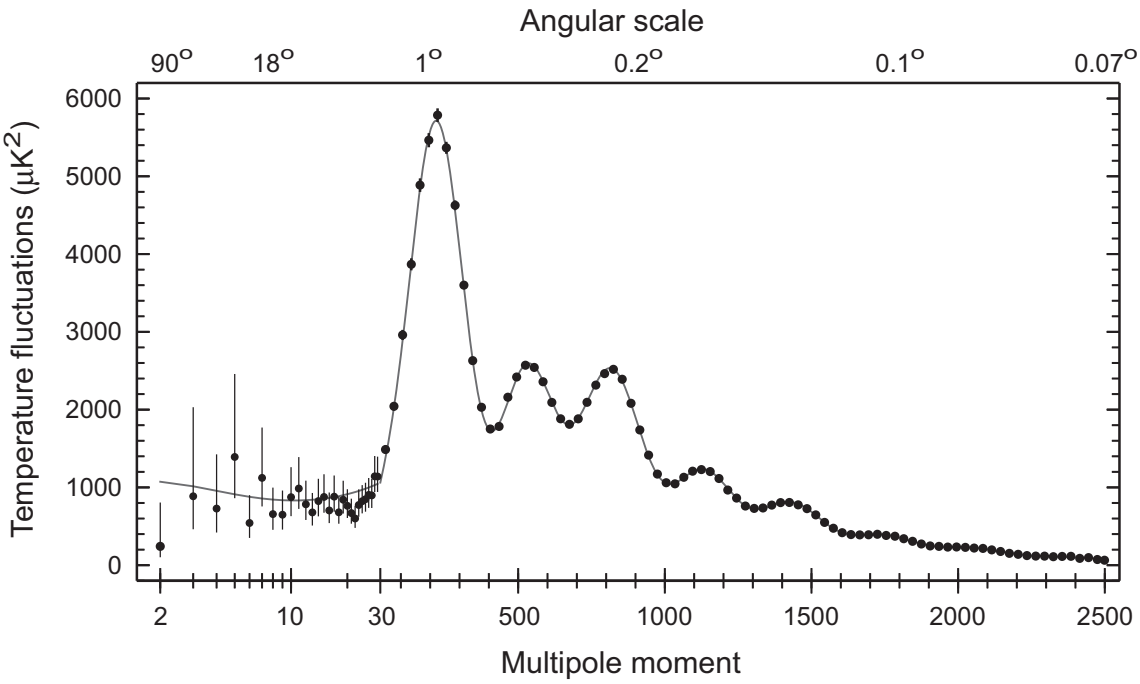
$$C(\theta) = \frac{1}{4\pi}\sum_{\ell=0}^{\infty} c_\ell(2\ell+1)P_\ell(\cos\theta).$$

A plot of *some function of* $C(\theta)$ *vs. multipole order* is called a *power spectrum*. An example is shown in Fig. 20.13.



- Physically, the $\ell = 0$ component is normally zero and the
- the $\ell = 1$ component is motion relative to the CMB.
- Thus, *cosmological interest centers on* $\ell \geq 2$.
- *Multipole moments* on the bottom axis of the figure above are related to an *angular scale* (top axis in the figure).
- Roughly, a multipole moment of order ℓ is sensitive to an angular region equal to $\sim 180^\circ/\ell$. Thus,

  - *Lower multipoles* carry information about the CMB on *large angular scales*.
  - *High multipoles* probe *small angular scales*.



Examples:

- The $\ell = 10$ multipole above is sensitive to structure on an angular scale of

$$\frac{180^\circ}{10} \simeq 18^\circ,$$

- while the $\ell = 1800$ multipole reflects structure on an angular scale of

$$\frac{180^\circ}{1800} \simeq 0.1^\circ,$$

Thus in CMB temperature fluctuation maps, *statistically*,

- lower multipoles $\longleftrightarrow$ larger features
- higher multipoles $\longleftrightarrow$ smaller features.

20.8 The Origin of CMB Fluctuations

The CMB represents photons liberated near the *decoupling transition*.

- Decoupling occurred about *380,000 years after the big bang*.
  - At decoupling the visible Universe was a *thousand times smaller* and
  - a *billion times more dense* than it is today.
- The pattern of CMB temperature fluctuations displayed in the sky maps reflects the *perturbation of big bang photons at decoupling*.
- These were caused by the *density fluctuations* present at the time of *last scattering* (when statistically photons were unlikely to undergo further scattering before reaching us).
- Thus, this pattern contains *detailed information about the state of the Universe near the decoupling transition* and it is important to understand the cause of the fluctuations.

The origin of the CMB fluctuations is *different for the low multipoles and high multipoles* in the power spectrum.

20.8.1 Angular Size Distance

Cosmological distances are a matter of *definition:* physical distances in general relativity are *proper distances,* but

> *What we can measure* on cosmological scales is *not directly a proper distance*.

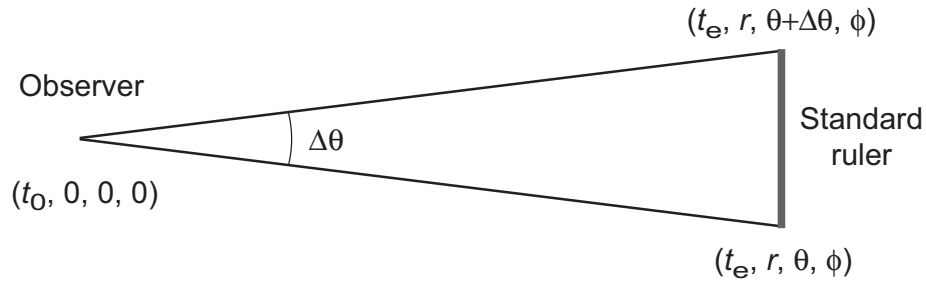
- One such defined distance was introduced earlier, the *luminosity distance* d_{L}.
- The luminosity distance is based on the idea of a *standard candle* (an astronomical object of *known intrinsic luminosity*).
- In studying the CMB we are often concerned with *angular measure* and it is useful to introduce another defined distance, the *angular size distance* d_{A}.
- Angular size distance is based on the idea of a *standard ruler* (an astronomical object of *known transverse size*).

The *angular size distance* is defined to be

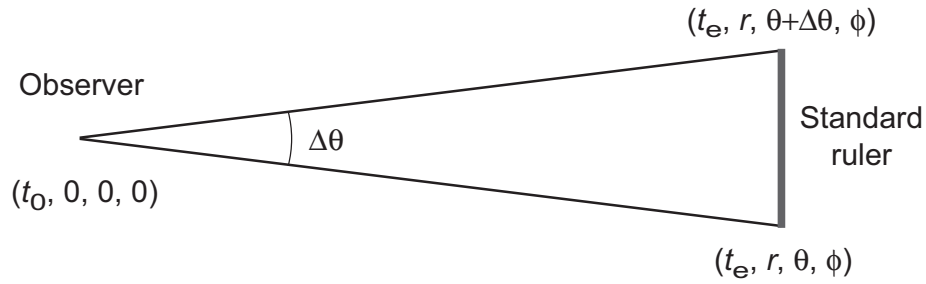
$$d_{\mathrm{A}} = \frac{D}{\Delta\theta}$$

where $\Delta\theta$ is the observed angular size and D is the physical length of the standard ruler.

- In *static euclidean space* d_{A} is the proper distance but in GR we must account for *curvature* and *expansion*.
- Assume the RW metric with
  - an *observer* at comoving coordinates $(t_0, r, \theta, \varphi) = (0,0,0,0)$, and
  - a *standard ruler at redshift* z stretching from comoving coordinates $(t_{\mathrm{e}}, r, \theta, \varphi)$ to $(t_{\mathrm{e}}, r, \theta + \Delta\theta, \varphi)$:



- Light is *emitted from the standard ruler* at time t_{e} and *travels to the observer on a geodesic* with $\theta = \varphi =$ constant, arriving at time t_0.



- The proper length of the standard ruler at t_e is known to be D, and it also can be calculated from the line element.
- Using the RW metric in the form

$$ds^2 = -dt^2 + a(t)^2[dr^2 + S_k(r)^2\, d\Omega^2]$$

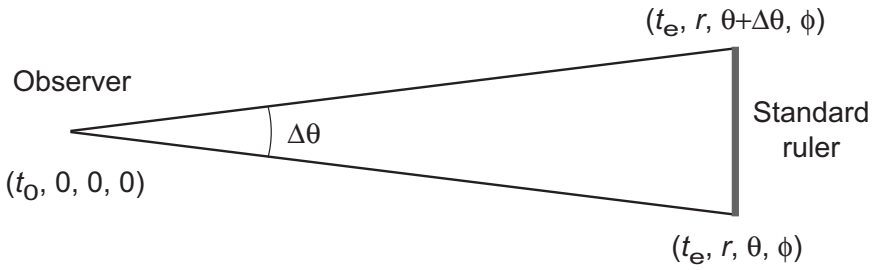
$$S_k(r) = \begin{cases} \sin r & (k=+1) \\ r & (k=0) \\ \sinh r & (k=-1) \end{cases}$$

with $dr = d\varphi = dt = 0$, and $d\Omega = \Delta\theta$ since φ is fixed, gives for the *length of the standard ruler*

$$D = ds = a(t_e) S_k(r) \Delta\theta .$$

- Inserting this in the defining equation $d_A = D/\Delta\theta$, the *angular size distance* is

$$d_A = \frac{D}{\Delta\theta} = \frac{a(t_e) S_k(r) \Delta\theta}{\Delta\theta} = a(t_e) S_k(r) = \frac{S_k(r)}{1+z}.$$



- Comparing with the luminosity distance $d_{\rm L}$ defined earlier,

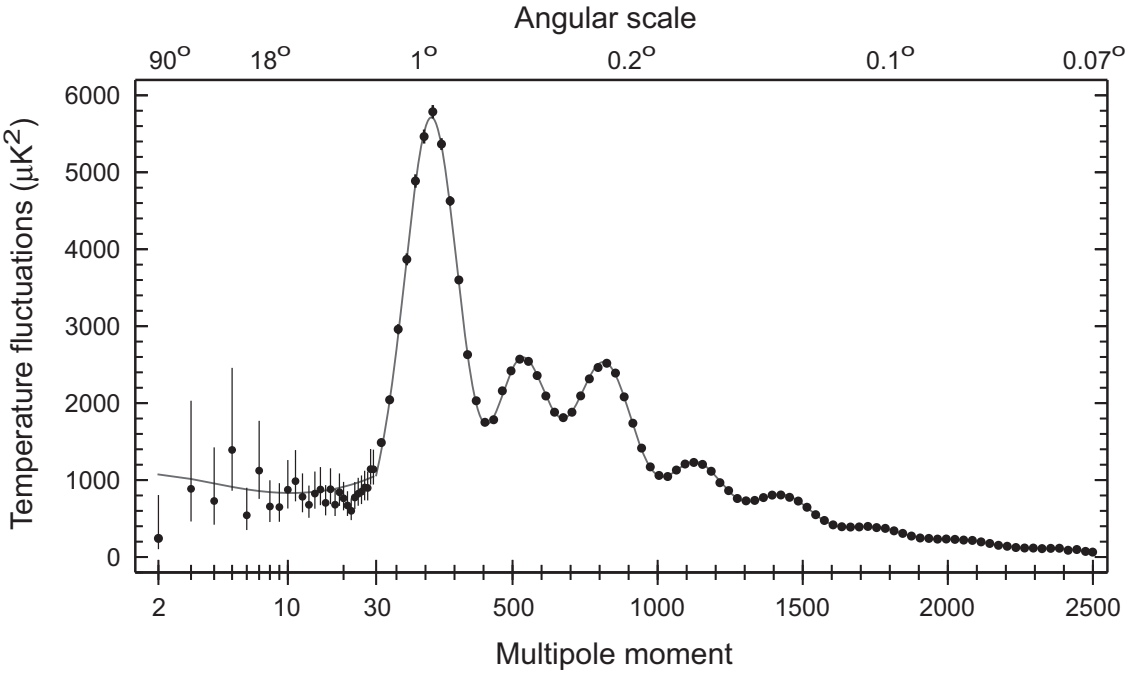
$$d_{\rm L} = S_k(r)(1+z) \simeq (1+z)r = (1+z)\ell(t_0),$$

we find that

$$d_{\rm A} = \frac{d_{\rm L}}{(1+z)^2}.$$

- Thus $d_{\rm A}$ and $d_{\rm L}$ are *very different at large redshift*.

For a flat Universe $d_{\rm A}$ equals the proper distance *at the time the light was emitted* (Problem).



The origin of the CMB fluctuations is *different for low and high multipoles* in the above power spectrum.

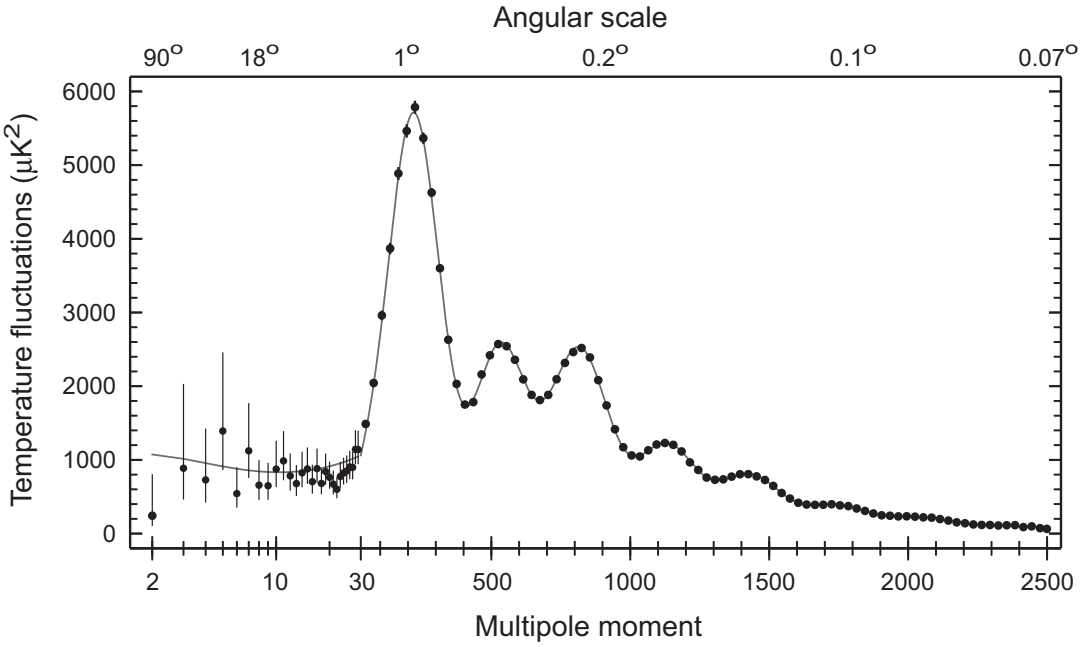
20.8.2 The Sachs–Wolf Effect

Consider multipole orders in the above spectrum below $\ell \sim 30$.

- This corresponds to *large angular scales on the sky*.
- At decoupling ($z \sim 1080$) energy densities for dark matter, radiation, and baryonic matter were in the ratio

$$\varepsilon_{\rm dm} : \varepsilon_\gamma : \varepsilon_{\rm b} \sim 5.5 : 1.8 : 1.$$

- Thus *gravity was dominated by the dark matter*.



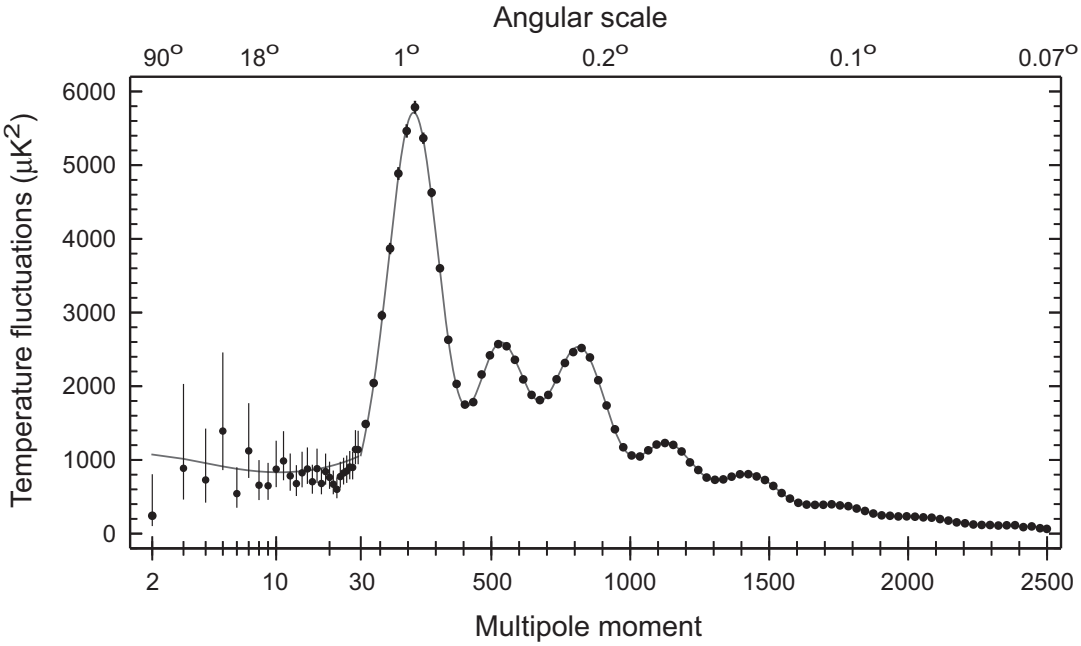
- Dark matter was not coupled to photons so it could *contract under its self-gravity before baryonic matter could*.
- Small dark matter overdensities *grew prior to decoupling*.
- These *fluctuations in the dark matter energy density*

$$\varepsilon(\boldsymbol{r}) = \varepsilon_0 + \delta\varepsilon(\boldsymbol{r})$$

caused *fluctuations in the gravitational potential*. In Newtonian approximation, the *Poisson equation* implies

$$\boldsymbol{\nabla}^2(\delta\varphi) = 4\pi G\left(\frac{\delta\varepsilon}{c^2}\right).$$

- A CMB photon climbing out of a local potential minimum at decoupling is *redshifted* and
- one falling off a local potential maximum is *blueshifted*.



- This causes *fluctuations in the CMB temperature*

$$\frac{\Delta T}{T} = \frac{\delta\varphi}{3c^2}$$

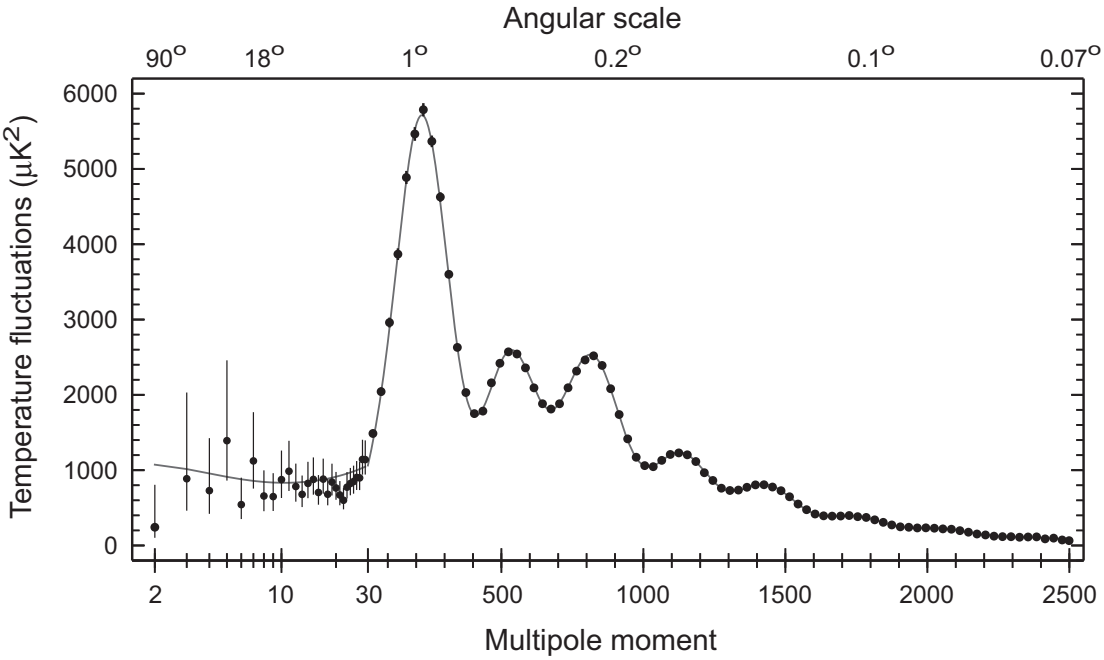
  called the *Sachs–Wolfe effect*.

- The Sachs–Wolfe effect is the main source of CMB fluctuations for angular scales larger than $\sim 10^\circ$ [multipole orders below $\ell \sim 20$].

- From the above figure, CMB power is *nearly constant for* $\ell < 20$.

- Thus, Sachs–Wolfe fluctuations are *nearly constant over a broad range of angular scales* and thus distance scales.

- This *scale invariance of density fluctuations* is a central prediction of the *theory of inflation* (next chapter).



CMB fluctuations on *small angular scales (large multipole order)* have a different origin that is associated with *sound waves in the early Universe*.

20.8.3 Sound Waves

- Consider a small region of the young Universe that is *slightly overdense* relative to the surrounding region.
- Because it is overdense, it will *compress gravitationally*.
- What happens next depends on whether the fluid is *neutral or ionized*.

If the *fluid is uncharged,* little radiation pressure opposes gravity and the region is *unstable to collapse*.

- Overdense regions become *more dense* and
- underdense regions become *less dense*.
- Over cosmological time this leads to the formation of *superclusters of galaxies surrounding large voids*.

The situation is *quite different if the cosmic fluid is ionized*, as it is earlier than the decoupling transition.

- Strong Thomson *scattering of photons from free electrons* couples the plasma into an effective *baryon–photon fluid*.
- This fluid has only about half of the density of the dark matter at decoupling.
- Hence the baryon–photon fluid moves *mostly under the gravitational influence of the dark matter* and not its own self-gravity.
- Gravitational compression of a region *increases density, temperature, and pressure*, producing *radiation pressure* that opposes the gravitational contraction.
- This drives an *expansion of the region* that causes the pressure to drop, eventually leading to *another gravitational contraction*, and so on.

- Thus the baryon–photon plasma develops *sound waves (pressure fluctuations)* called *acoustic oscillations,* for which
  - the *driving force is gravity* and
  - the *restoring force is radiation pressure* of the photons.
- These sound waves *travel at very high speed* in the ionized fluid.
- Near decoupling $T \sim 3000\,\mathrm{K}$ and *photons are* $\sim 10^9$ *more abundant than baryons* in the fluid.
- At constant entropy the *speed of sound* v_s is

$$v_s = \left(\frac{dP}{d\rho}\right)^{1/2} \simeq \frac{c}{\sqrt{3}},$$

since photons dominate and for a photon gas

$$P = \tfrac{1}{3}\varepsilon = \tfrac{1}{3}\rho c^2.$$

Thus radiation pressure

- keeps density fluctuations from collapsing
- and creates *acoustic oscillations*

in the plasma before the decoupling transition.

As we will now discuss, there is a *preferred length scale* associated with these acoustic oscillations.

- This preferred distance is called the *acoustic scale* or the *sound horizon*.
- It is associated with the *distance sound could have traveled between the big bang and decoupling*.
- As we shall now see, this length scale is expected to leave its imprint on both
  - *the radiation* (through *CMB fluctuations*) and
  - on *the matter* (through *baryon acoustic oscillations* in the presently observed distribution of galaxies).

20.8.4 The Acoustic Scale

The *sound horizon scale* is set by the proper distance $\ell_s(t_{ls})$ sound could travel from the big bang to the decoupling time t_{ls}.

- This may be evaluated from

$$\ell_s(t_{ls}) = a(t_{ls}) \int_0^{t_{ls}} \frac{v_s dt}{a(t)},$$

  where $v_s \sim c/\sqrt{3}$ is the speed of sound.

- From cosmological data the acoustic scale corresponds to a proper distance $\ell(t_{ls}) \sim 0.144\,\text{Mpc}$ on the last scattering surface (Problem).

- *Due to expansion*, this implies a *proper distance today* of

$$\ell(t_0) = (1 + z_{ls})\ell(t_{ls}) \sim 156\,\text{Mpc}.$$

- Hence a *preferred proper distance scale*

$$\underbrace{\ell(t_{ls}) \sim 0.144\,\text{Mpc}}_{z\sim 1080} \quad \longrightarrow \quad \underbrace{\ell(t_0) \sim 156\,\text{Mpc}}_{z=0}$$

  is established by the *sound horizon at last scattering*.

- If, *at last scattering*, the fluid in an oscillation region is
  - at *maximum compression,* its density will be *higher* and photons emitted from it will be *slightly hotter;*
  - if it is at *maximum expansion,* the density will be *lower* and the emitted photons *slightly cooler,*

  relative to the average temperature of the CMB.

20.8.5 Acoustic Signature in the CMB

The *acoustic length scale* is based on well-known physical principles (speed of sound in a photon-dominated plasma).

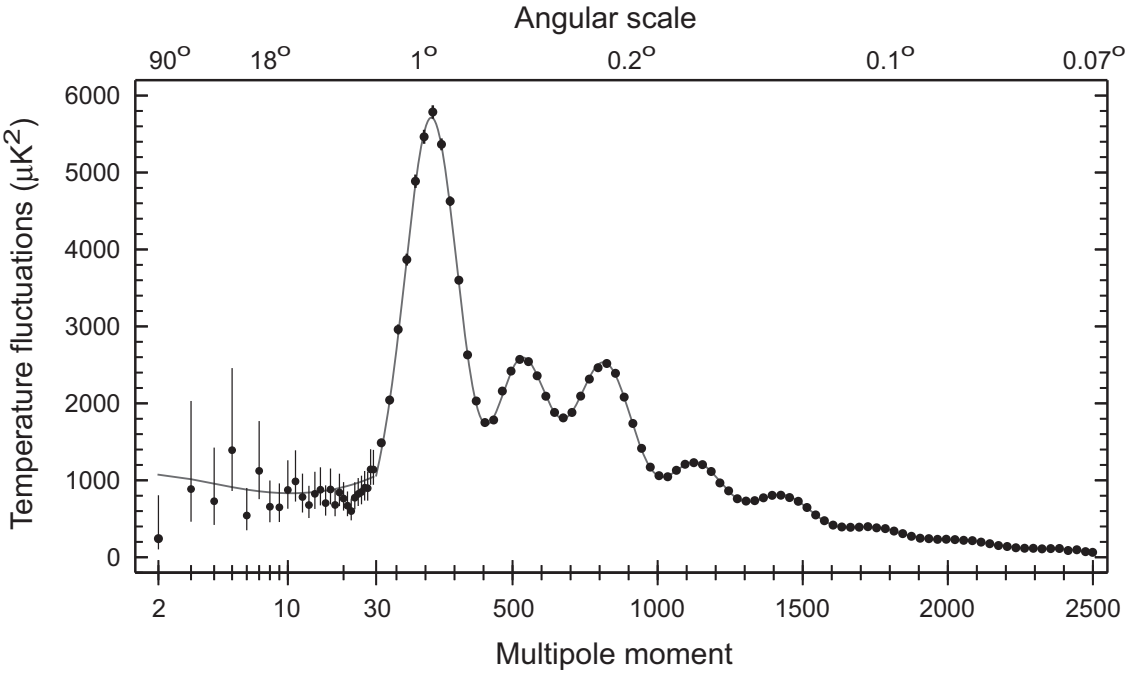
- Therefore it defines a *standard ruler*.
- Then an *observed angular size* can be associated with *a physical distance* on the last scattering surface (LSS).
- Specifically, fluctuations in the CMB of angular size $\Delta\theta$ are related to a physical transverse size D on the LSS by

$$D = d_{\mathrm{A}}\Delta\theta,$$

  where d_{A} is the angular size distance.
- A numerical solution assuming the parameters taken from Planck satellite data gives $d_{\mathrm{A}} = 12.9\,\mathrm{Mpc}$.
- Thus a *proper transverse distance* D on the LSS is related to an *angle* $\Delta\theta$ *observed today* on the celestial sphere by

$$\Delta\theta = 4.45 \times 10^{-3} \left(\frac{D}{\mathrm{kpc}}\right) \mathrm{deg}.$$

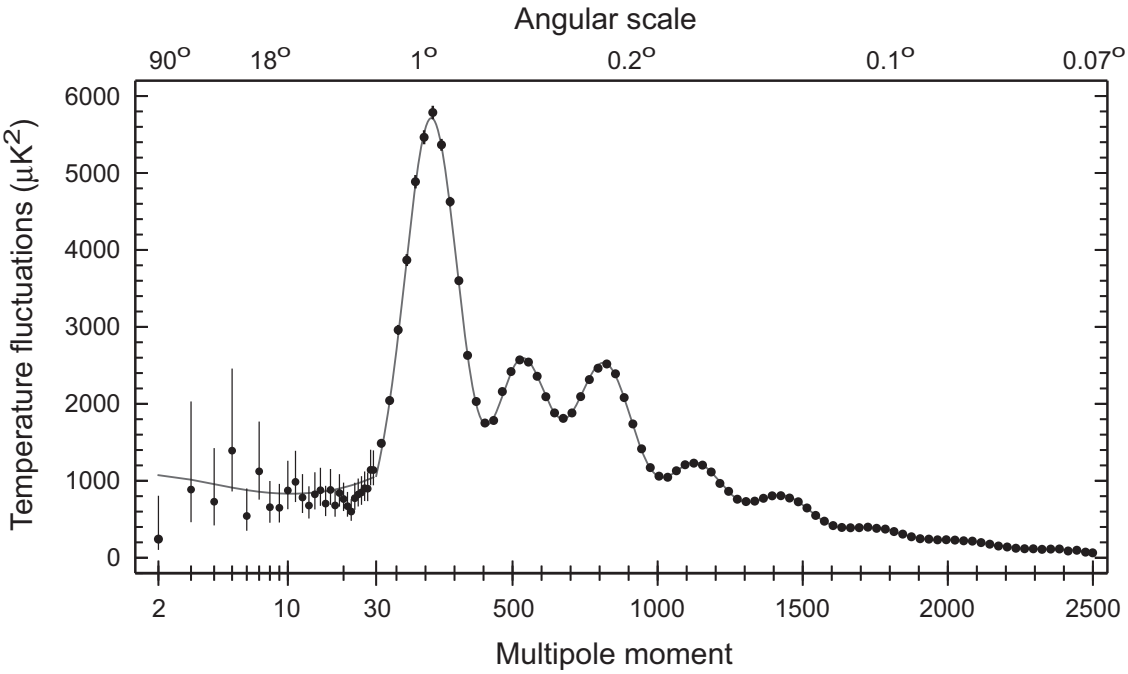
- The *preferred distance scale* established by acoustic oscillations at the time of last scattering translates into a *preferred angular scale* for the CMB observed today.
- Inserting the acoustic scale of 144 kpc in this equation implies that a *preferred angular size of* $\Delta\theta \sim 0.6^\circ$ should be observable in the temperature fluctuations of the CMB.



The patterns of hot and cold spots on small angular scales in the CMB are a snapshot of the *sound wave pattern* in the cosmic fluid at the time of decoupling.

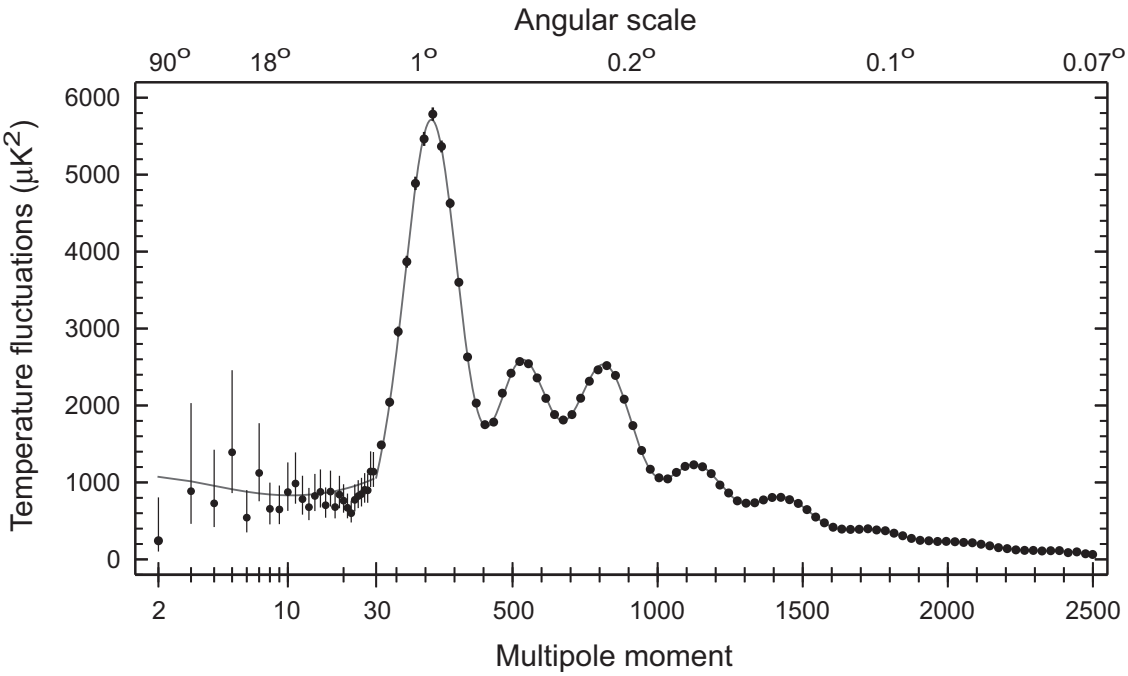
- The corresponding *CMB power spectrum* (figure above) then contains a *wealth of cosmological information*.
- The *first peak is strongest.*
  - It occurs at a multipole order associated with the preferred angular size of $0.6°$.
  - This represents acoustic oscillations that were at *maximum compression* at the time of last scattering.

The *harmonic series of peaks* represents overtones of the *preferred* $0.6°$ *angular scale*.



- Observed *angular sizes depend on curvature*.
  - *Positive curvature* shifts the first peak to *larger angle;*
  - *negative curvature* shifts it to *smaller angle*.

  At this point, *data are consistent with zero curvature*.
- The amplitude of the first peak is
  - *increased* by lower sound speed v_s and
  - *decreased* by higher v_s.
- Thus the *height of the first peak* is
  - a diagnostic of the *speed of sound* at decoupling.
  - This is in turn a diagnostic for the *baryon density*.



- Equal spacings of the peaks indicates that the initial fluctuations were *adiabatic*.
  - *Adiabatic:* all species varied together in density.
  - Adiabatic fluctuations as initial big bang conditions are *predicted by inflation* (next chapter).
- The relative heights of the peaks are sensitive to the *dark matter density* and the *baryon density*.
- The position of peaks measures the *acoustic length scale*.
  - This *scale is physical,* so it is a *standard ruler*.
  - This allows an *independent determination of distance* to the last scattering surface.

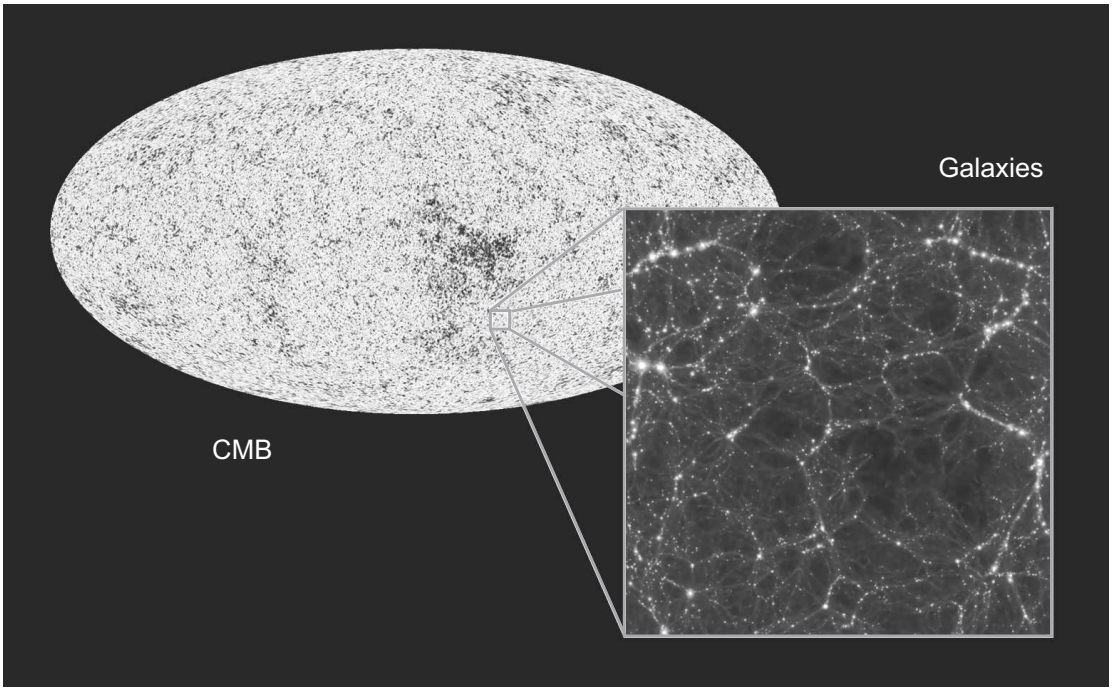
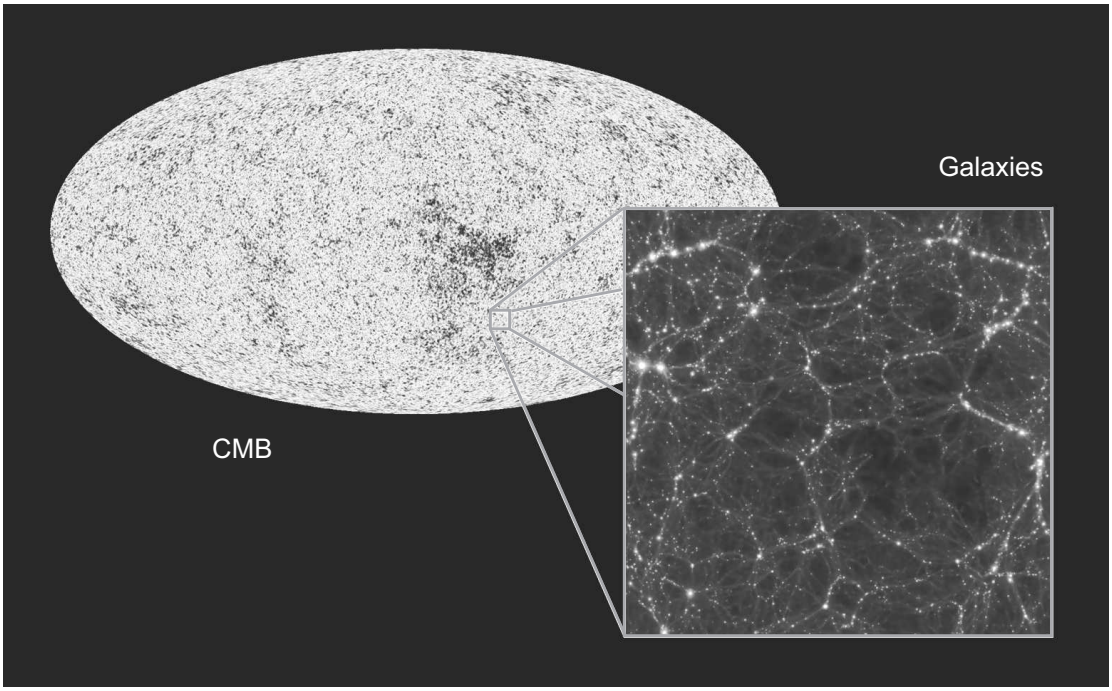


Figure 20.14: Acoustic oscillations at the decoupling transition leave an imprint on the CMB at last scattering, and on the large-scale distribution of galaxies at later times.

20.8.6 Acoustic Signature in Galaxy Distributions

A *characteristic length scale* was imprinted on the CMB by *acoustic oscillations at decoupling*.

- This length scale also should be *visible in the baryonic matter*.
- This is illustrated schematically in Fig. 20.14.
- The possibility of such *baryon acoustic oscillations (BAO)* was introduced earlier; now be discuss it in more detail.



A *remnant signature of the BAO* should be observable in the *clustering of galaxies* today on large scales.

- The acoustic length on the last scattering surface has been stretched by a factor of $1+z_{\rm ls}$ to $\sim 150\,{\rm Mpc}$ today.
- Using the average matter density parameter $\Omega_{\rm m} \sim 0.3$, a volume of this size contains more than $10^{17}\,M_\odot$.
- This greatly exceeds the mass of a galaxy supercluster.
- Furthermore, the original BAO signal has been
  - *diluted* by mixing with the dominant dark matter and
  - *smeared* by peculiar local velocities and fluid flows

  in the evolution of the Universe.

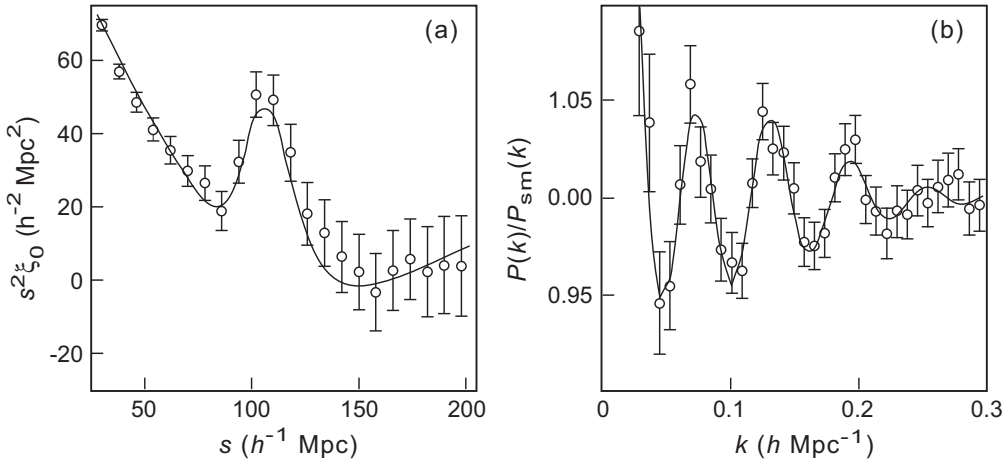


Figure 20.15: Baryon acoustic oscillation for approximately 10^6 galaxies. (a) Correlation function versus comoving radius. (b) Power spectrum versus wavenumber. The data have been reconstructed to remove effects of matter flows and peculiar velocities on intermediate scales and give sharper peaks. The curves are calculations for a best-fit BAO model. The galaxy power spectrum (b) is similar in spirit to the CMB power spectrum except that the 3D data are expanded in a Fourier series rather than in spherical harmonics, and the index k is a wavenumber rather than a multipole moment.

Finding a BAO requires a *huge sample of 3D galaxy locations*.

- Surveys of angular positions that also determine redshifts are effectively 3D since redshift is a proxy for distance.
- Evidence for a BAO in a survey of $\sim 10^6$ galaxies from the *Sloan Digital Sky Survey (SDSS)* is shown in Fig. 20.15.
- Assuming $h \sim 0.7$, the peak in Fig. 20.15(a) indicates a preferred length scale of about 150 Mpc.
- This is roughly the estimated *present size of the acoustic length scale* established by how far sound waves could have traveled before the decoupling transition.

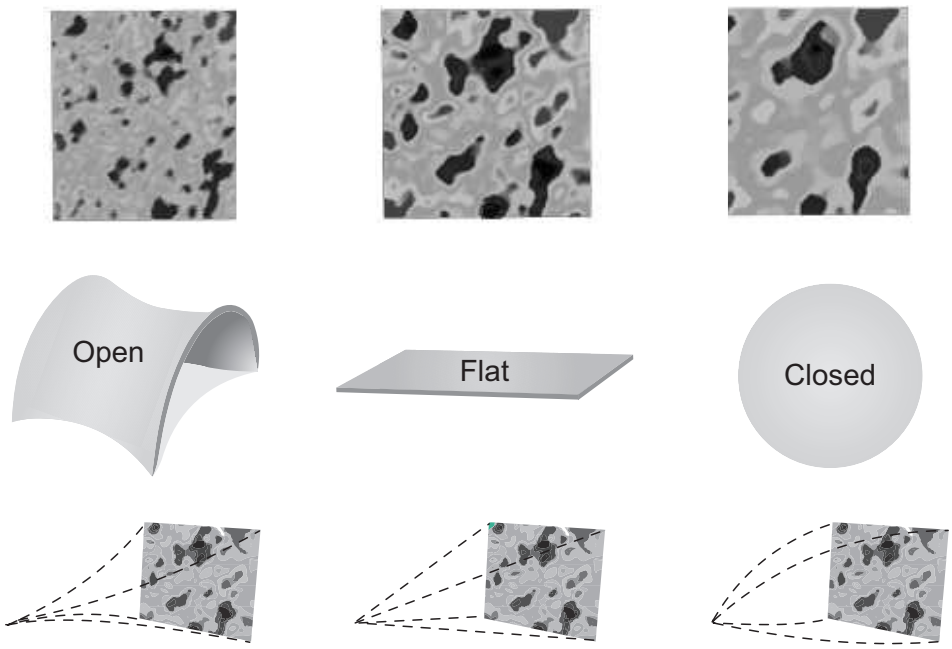


Figure 20.16: Schematic influence of spacetime curvature on observed CMB fluctuations.

20.9 Precision Measurement of Cosmology Parameters

WMAP and *Planck* observations in particular have yielded *precise constraints on important cosmological parameters*.

- This is because the detailed pattern of CMB fluctuations is extremely sensitive to many cosmological parameters.
- For example, Fig. 20.16 illustrates schematically that lensing effects on the CMB distort it in a way that depends on the overall curvature of the Universe.

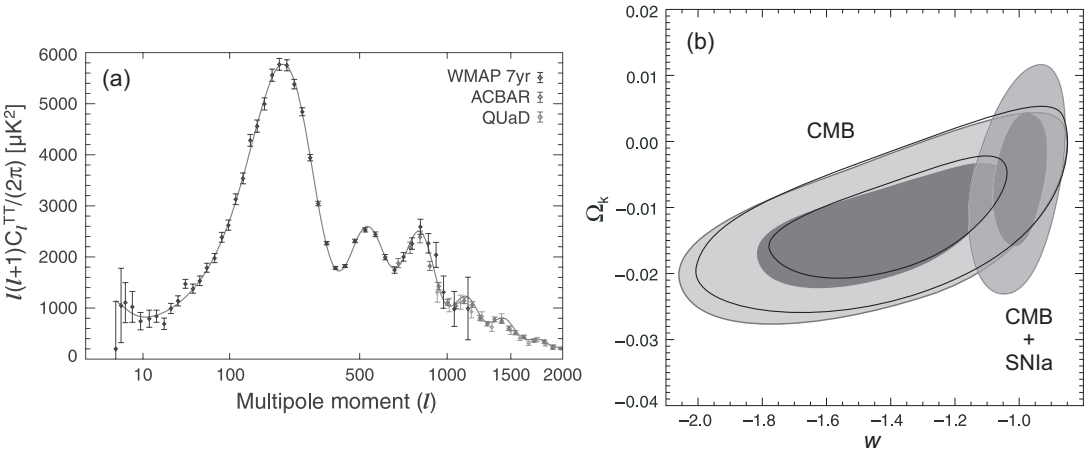


Figure 20.17: (a) CMB power spectrum. (b) Curvature density Ω_k versus w for an equation of state $P = w\varepsilon$. The diagonal locus indicates CMB analysis; the more vertical locus combines CMB and Type Ia supernova data.

As a second example,

- Fig. 20.17(a) illustrates a CMB power spectrum and
- Fig. 20.17(b) shows confidence intervals for
  - the *curvature density* Ω_k and
  - the *equation of state parameter* w in $P = w\varepsilon$

  implied by the CMB data.
- The inferred equation of state parameter is $w = -1 \pm 0.02$.
- This is consistent with $w = -1$ *(cosmological constant)*.

From Fig. 20.17, the *total density parameter* is

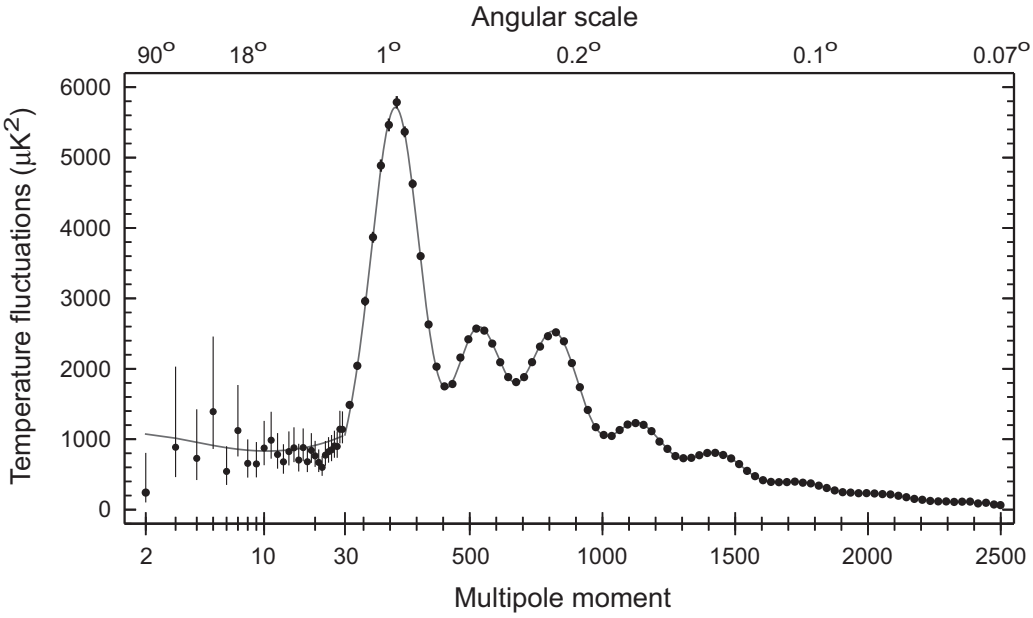
$$\Omega = 1 \pm 0.03,$$

which is consistent with a *flat Universe*.

Table 20.2: Cosmological parameters from Planck CMB data

Parameter	Symbol	Value
Hubble parameter	H_0	$67.8 \pm 0.9\,\text{km s}^{-1}\,\text{Mpc}^{-1}$
Age of the universe	t_0	$13.80 \pm 0.04\,\text{Gyr}$
Baryon density	Ω_b	0.0484 ± 0.0005
Matter density	Ω_m	0.308 ± 0.012
Dark energy density	Ω_Λ	0.692 ± 0.012
Dark energy equation of state†	w	-1.006 ± 0.045
Sum of neutrino masses	$\sum m_\nu$	$< 0.23\,\text{eV}$
Curvature density	$\lvert\Omega_k\rvert$	< 0.005
Redshift for rad/matter equality	z_eq	3365 ± 44

† The coefficient w in $P = w\varepsilon$.



Fits to power spectra determine cosmological parameters.

- The Planck CMB power spectrum is shown above and
- Planck cosmological parameters are shown in Table 20.2.

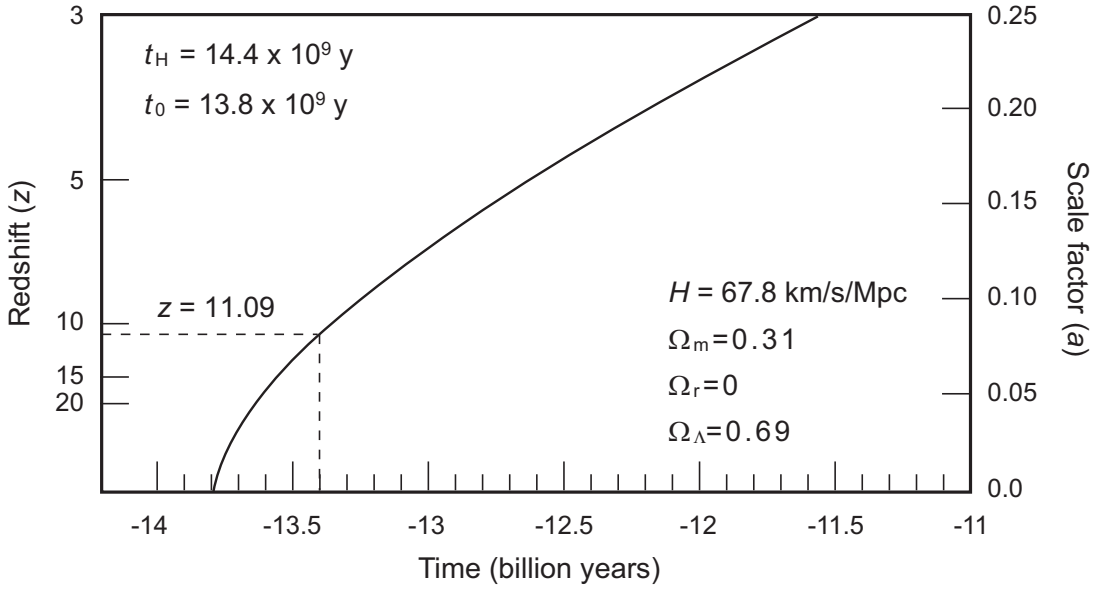


Figure 20.18: Redshift and scale factor in the early Universe using cosmological parameters consistent with Planck satellite data. Redshift $z = 11.09$ corresponds to the most distant galaxy that had been observed as of 2017.

Evolution of the early Universe for a Friedmann cosmology using parameters from Planck data is illustrated in Fig. 20.18.

- As of late 2017, the largest redshift observed was $z = 11.09$ for the irregular galaxy GN-z11.
- From Fig. 20.18, this redshift corresponds to light emitted about 400 million years after the big bang.

This indicates that *galaxy formation was well underway* during *reionization of the Universe*, which began $\sim$ 200 million years after the big bang.

The precision with which parameters are now being determined from

- CMB data
- high-redshift supernovae
- (supplemented by lensing, baryon acoustic oscillation, and other observational data from more traditional observational astronomy)

is unprecedented.

Over little more than a decade these new observations transformed cosmology into a *quantitative science constrained by precise data*.

20.10 Seeds for Structure Formation

CMB fluctuations reflect conditions when matter and radiation decoupled.

- If the CMB were perfectly smooth, it would be difficult to understand how structure could have formed.
- A period of *exponential growth* in the early Universe called *cosmic inflation* may have been crucial to producing the observed CMB density fluctuations (next chapter).
- However, the observed CMB fluctuations are *too small by themselves* to account for the large-scale structure observed today without help from *dark matter*.
- Dark matter was crucial in aiding structure formation.
    1. Because dark matter does not couple to photons, it could *begin to clump earlier* than the normal matter.
    2. Because there is so much more dark matter than normal matter, it could *clump more effectively*.
- Thus, dark matter provided the initial regions of higher density that *seeded the early formation of structure* in the Universe.
- This permitted structure formation to *proceed faster* than it would have otherwise.
- These ideas are illustrated schematically on the next page.

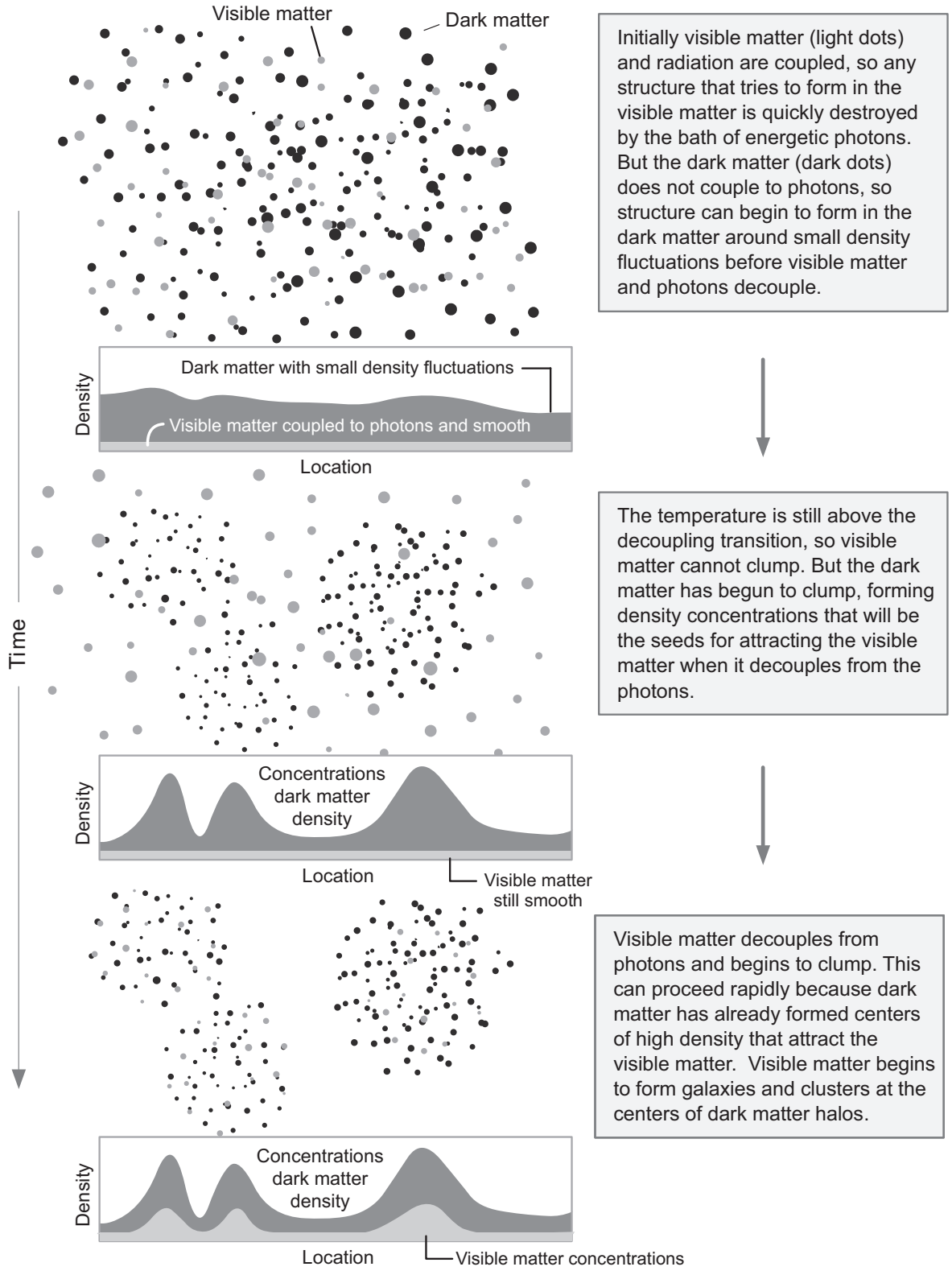


Figure 20.19: Formation of structure. Dark dots are dark matter; lighter dots are normal matter.

As already discussed, hot and cold dark matter suggest *different paradigms for formation of structure* in the early Universe.

- *Hot dark matter* implies high velocities for the dark particles.
- They will be stopped only after passing through a volume containing of order $10^{15} M_\odot$ *(mass of galaxy superclusters)*.
- This implies that they can *stream freely over 10–100 Mpc distances*.
- Hot dark matter streaming freely on these scales *disfavors structure formation on any smaller scales:*

In a Universe dominated by *hot dark matter*

- the formation of structure occurs *first on large scales* and
- this structure must then *break apart to form smaller structures*.

This is called the *top-down model* for structure formation.

On the other hand,

- *Cold dark matter* has relatively low velocity and
- tends to travel only a small distance before being stopped.
- This favors the *formation of structure on small scales* such as galaxies, and then
- *larger structures can form by aggregation* of these smaller structures.

This is termed the *bottom-up model* of structure formation.

The *top-down scenario* has some *serious problems:*

- Observations the furthest back in time see *quasars, gamma-ray bursts, and galaxies*.
- They *do not see superclusters of galaxies*.

The best current simulations indicate that formation of structure was

- *dominated by cold dark matter* in a *bottom-up scenario*, though
- quantitative details may require a small admixture of hot dark matter.

As we have already noted, the cold dark matter is

- *essential for accelerating the bottom-up scenario* because otherwise
- the Universe would *not be old enough* to have assembled the large-scale structure seen in the current Universe.

20.11 Summary: Dark Matter, Dark Energy, and Structure

Let's conclude this chapter by summarizing our understanding of *dark matter*, *dark energy*, and *formation of structure*.

- If *inflation is correct* (next chapter) and the *cosmological constant were zero*,
  - the matter density of the Universe would be *exactly the closure density*,
  - which would lead to *flat geometry*.
  - However, that is *not the nature of our Universe* according to observations.
- Current data indicate that the Universe *is indeed flat*, as predicted by inflation, but
  1. *it does not contain a closure density of matter* because there is a *non-zero cosmological constant*. Instead,
  2. about 30% of the closure density is *supplied by matter*
  3. and about 70% is supplied by *dark energy* (*vacuum energy* or a *cosmological constant*).
  4. Luminous matter contributes a small fraction of the closure density, implying that
  5. the vast majority of the mass density is *dark matter*.

Thus, the present Universe is *dominated by dark matter and dark energy*.

- Known neutrinos are *relativistic (hot dark matter)*.
    1. They could aid the formation of large structures like superclusters but not smaller structures like galaxies.
    2. Thus, (the known) neutrinos are *not likely to account for more than a small fraction of the dark matter*.
    3. WMAP and Planck data support this conclusion.

- On the scale of galaxies and clusters of galaxies, *90% of the total mass is not seen*.
    1. A significant fraction of the dark matter could be *normal (baryonic),* in the form of small, *very low luminosity objects* like white dwarfs, neutron stars, black holes, brown dwarfs, red dwarfs, Jupiters, ...
    2. However, *microlensing* and *searches for subluminous objects* have not found enough of these "normal" objects to account for the mass of galaxy halos.

- Further, *strong constraints from big bang nucleosynthesis* compared with observed abundances of light elements indicate that *most of the dark matter cannot be baryonic*.
    1. Thus, much of the dark matter is likely to be *nonbaryonic and not neutrinos*, and to be *cold (massive)*.
    2. Current speculation centers on *undiscovered elementary particles* as candidates for this cold dark matter.
    3. However, *no such particles have been found* so far.

- *Large-scale structure and its rapid formation* in the early Universe is hard to understand,
  - given the *smallness of the CMB fluctuations* implied by COBE, WMAP, and Planck measurements,
  - unless *cold dark matter seeds initial structure formation*.
- The models of structure formation most consistent with current data are the class of Λ*CDM models* that
  - combine a *cosmological constant* (denoted by Λ) with
  - *cold dark matter* (CDM)
  - to give an *accelerating but flat* universe,

  with the *cold dark matter seeding structure formation*.
- The origin of dark energy is an *almost complete mystery*.
  - The most economical explanation would be the energy of *quantum vacuum fluctuations*.
  - However, with our present understanding of how to calculate the vacuum energy that explanation is *highly inconsistent with observations*.

We will consider further the *energy of vacuum fluctuations* in the last chapter.

Chapter 21

Extending Classical Big Bang Theory

The big bang is our *standard model for the origin of the Universe*. This place is well earned.

- *At a broad conceptual level*, many cosmology observations make sense only if there was a big bang in our past.
- *At a more nuts and bolts level*, the standard big bang model accounts quantitatively for precise observational data not easily be explained by any competing theory.

However, some observations in modern cosmology suggest that the classical big bang is an

- essentially *correct*, but
- perhaps *incomplete*

picture of the origin and evolution of our Universe. In this chapter we address some of these issues.

21.1 Successes of the Big Bang

Let's begin by summarizing the role of the big bang theory in modern cosmology. The standard cosmology rests on a *relatively few observations and concepts:*

1. The *redshifts of distant galaxies* imply that
   - we live in an *expanding Universe* that is
   - described at the simplest level by the *Hubble law*.
2. On *large enough scales* (beyond superclusters of galaxies) the Universe appears to be both
   - *homogeneous* and
   - *isotropic*.

   This is called the *cosmological principle*.
3. *Properties of distant quasars* suggest that these powerful energy sources were once
   - *more energetic* and
   - *more closely spaced* than they are today,

   implying that the Universe has *evolved with time*.
4. The *cosmic microwave background (CMB)* is all-pervading and
   - *highly isotropic* but with *fluctuations at the* 10^{-5} *level*, and
   - has a *blackbody spectrum,* with $T = 2.725\,\mathrm{K}$.

5. The *elemental composition* of the Universe is (by mass) *three parts H to one part He*, with but a trace of heavier elements.

6. The Universe appears to be *charge neutral* on large scales. Conversely, observations indicate that

   - the Universe is *highly asymmetric with respect to matter and antimatter,*
   - with no evidence for significant equilibrium concentrations of antimatter.

7. There are *many fewer baryons than photons* in the Universe.

8. In contrast to the homogeneity on very large scales,

   - matter on scales comparable to superclusters of galaxies and smaller exhibits *complex and highly evolved structure*.
   - This contrasts sharply with the smoothness of the CMB.
   - Furthermore, observations indicate that this structure began to develop very *early in the history of the Universe*.

9. Detailed analysis of fluctuations in the CMB and of the brightness of distant Type Ia supernovae indicate that

   (a) The *expansion of the Universe* is presently *accelerating*.

   (b) The *geometry of the Universe* is remarkably *flat (euclidean)*.

10. The bulk of the matter in the Universe is

    - not luminous *(dark matter)* and
    - observable *only through its gravity*.

    Various observational constraints imply that *most dark matter is not baryonic*.

The *first five observations* fit well within classical big bang cosmologies:

- Observations (1)–(2) indicate that *we live in an expanding, isotropic universe*,
- Observations (3)–(5) indicate that
  - this Universe has *evolved over time* and
  - had a beginning in a *very hot, very dense initial state* (the big bang).

The remaining observations *need not be inconsistent with the classical big bang*, but they require either

- ad hoc *imposition of particular initial conditions* on the Universe, or
- *assumption of specific microscopic properties* for the matter and energy fields that the Universe contains.

Furthermore, observation (4): existence of a very smooth CMB but with fluctuations at the 10^{-5} level

- is *one of the great triumphs* of the big bang theory, but
- it *raises a potentially serious problem*.

21.2 Problems with the Big Bang

As indicated in the previous section, *observational properties (6)–(11)* constitute potential *problems for the classical big bang model*.

- They *do not invalidate* the big bang.
- However, they indicate that the big bang in its minimal form may be *incomplete*.
- Much of this incompleteness is likely to originate in
  - an inadequate understanding of how the particle and field content of the Universe influences its history.
  - Thus, the incompleteness might be addressed by a better understanding of how those particles and fields couple to the evolution.

Let's now discuss these isues in more depth.

21.2.1 The Horizon Problem

Observational property (4) *(isotopy of the CMB)* presents a potential *conflict with causality:*

- The CMB has *nearly the same temperature* in widely separated parts of the sky.
- This is understandable only if those regions were *in causal contact in the past*.
- But in the standard big bang model it is easy to show that
  - regions on the sky separated by more than a degree or two in angle *could not have exchanged light signals* since the big bang.
  - Thus, they could *never have been in past causal contact*.
- That is, they lie *outside each other's horizons*, as illustrated in Fig. 21.1 on the following page.

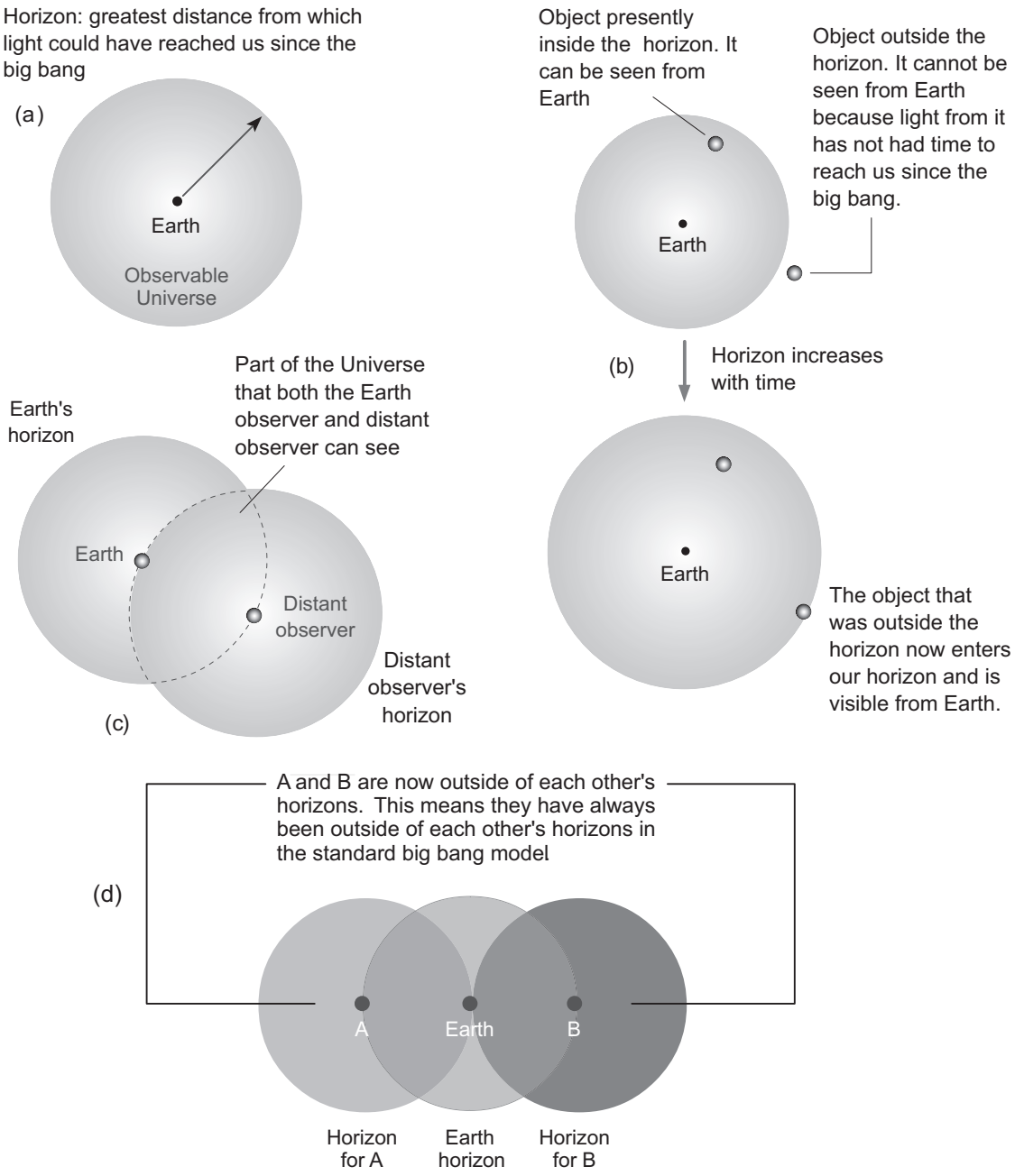


Figure 21.1: Horizons in an expanding universe. (a) A (cosmological) horizon is the greatest distance from which light could have reached us since the beginning of time. (b) Horizons expand with time, so objects currently outside our horizon may come within our horizon in the future. (c) Cosmological horizons are defined relative to each observer, so each has her own horizon. (d) Horizon problem produced by the CMB having identical temperatures on opposite sides of the sky for an observer: how do A and B know to have the same temperature if they could never have exchanged signals in the past?

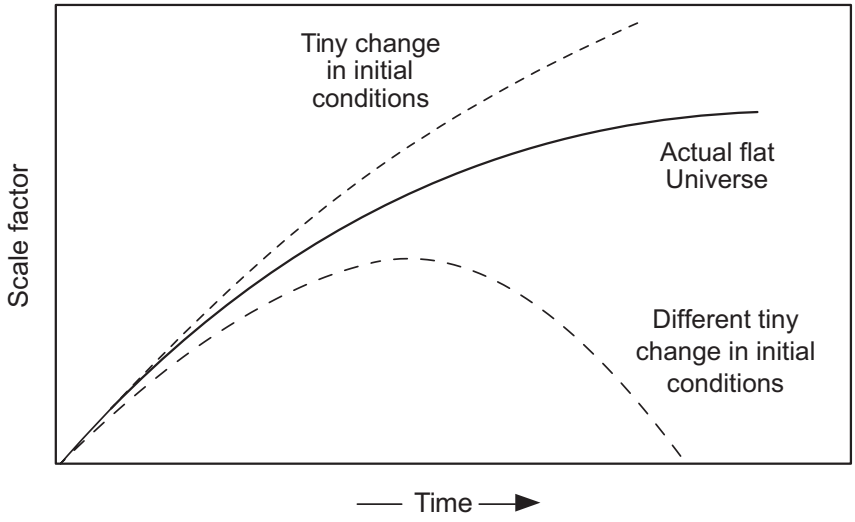


Figure 21.2: The flatness problem: producing a flat Universe today requires remarkable fine-tuning of the initial curvature for the Universe.

21.2.2 The Flatness Problem

Observational property 9(b) implies that *the Universe is flat*.

- Hence, the Universe must be *very near the closure density*.
- This condition is possible, but only if parameters are *very finely tuned* in the early universe (Fig. 21.2).
- The fractional deviation of density from critical density at any time in the evolution of the Universe is (Problem)

$$\frac{\Delta\rho}{\rho} = \frac{\rho - \rho_c}{\rho} = \frac{3kc^2}{8\pi G a^2 \rho},$$

- Unless flatness is *tuned to one part in* 10^{60} *at the Planck time*, we end up with a Universe that is *far from flat today*.

This is a *possible*, but not very *natural* initial condition.

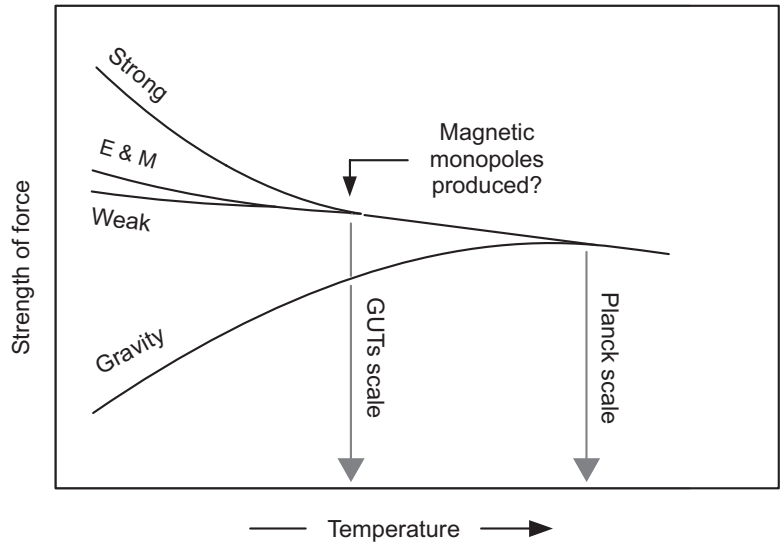


Figure 21.3: The magnetic monopole problem of the standard big bang: where are the massive relic particles that would be expected to be produced at phase transitions like grand unification (labeled GUTs)?

21.2.3 The Magnetic Monopole Problem

Massive *magnetic monopoles* might be *produced at phase transitions* in the early universe (Fig. 21.3). This causes problems:

- Such particles have *never been observed*.
- They would have caused a *more rapid change from radiation to matter dominated evolution* in the early Universe.
- This would *negate many correct big bang predictions*.

These problems are removed if we *assume* that such particles are *not produced in large numbers*. But *why* is this initial condition the correct one?

21.2.4 The Structure and Smoothness Dichotomy

Observational properties (4) and (8) (*smoothness of CMB* vs. *existence of large-scale structure*) present a compatibility issue:

- The remarkably high smoothness of the CMB implies that the early Universe was strikingly devoid of density perturbations.
- Where then did the density perturbations responsible for the growth of rich structure in the present Universe on the supercluster and smaller scale originate?

The form of the density perturbations required to give the observed structure is known so they could be *postulated as initial conditions*, but again we would like to know *why.*

21.2.5 The Vacuum Energy Problem

Observation 9(a) *(accelerated expansion)* has a simple explanation only if the Universe contains *dark energy*.

- This would be most naturally explained if dark energy is a *consequence of vacuum fluctuations*.
- However, estimates of the vacuum energy content of the Universe are *spectacularly wrong* in comparison with the corresponding observational constraints.
- The accelerated expansion is consistent with the big bang picture if we simply *postulate the existence of dark energy in the Universe* in the required amount.

But it is highly unsatisfying to have no understanding of where this fundamental influence on the evolution of the Universe comes from.

21.2.6 The Matter–Antimatter and Baryogenesis Problem

Observation (6) indicates that the physical universe contains almost entirely matter with *little corresponding antimatter*.

- We can *impose this as an initial condition.*
- But that is bothersome given that *matter and antimatter enter on an equal footing in modern elementary particle physics*.
- A closely-related problem (because annihilation of baryons with antibaryons produces photons) is how to account for the *large excess of photons over baryons* in the Universe (Observation 7).

21.2.7 Modifying the Classical Big Bang

We shall now discuss some possible resolutions of these problems.

- In attempting to resolve the problems we have to be careful to *preserve the successes of the big bang model*.
- The successful predictions of the big bang model rest primarily on the *evolution of the Universe at times later than about one second* after the initiation of the big bang.
- Therefore, any modification of our cosmological model that influences the Universe at *times earlier than about one second after the big bang*
  - will leave the successes of the big bang intact
  - if they leave appropriate initial conditions for the subsequent evolution.

We begin with a proposed modification of the evolution of the Universe that

- is operative only in the first tiny fraction of a second of the big bang and that
- has the potential to *resolve the first four problems in a single stroke*.

21.3 Cosmic Inflation

The theory of *cosmic inflation* is based on a simple but striking idea that has been discussed already in conjunction with the *de Sitter solution:*

- In the presence of an *energy density constant over all space*, the Einstein equation admits an *exponentially growing solution.*
- A *short burst of exponential growth* before settling down into more normal big bang evolution would have potentially large implications for the evolution of the Universe.
- We shall take the essential point of inflation to be this general idea that the *early Universe experienced a period of exponential growth*.
- There are many specific versions of inflationary theory that implement this in different ways.
- For the most part we shall leave those specifics for the interested reader to pursue in the specialist literature.
- Our reason is that
  - there is compelling evidence that the basic idea of inflation is necessary to explain evolution of the early Universe, but
  - no specific current version of inflation gives a completely satisfactory accounting of the cause and detailed effects of the inflationary period.

21.3.1 Inflationary Theory

From earlier discussion of the *de Sitter solution*,

- A universe with *pure vacuum energy expands exponentially*,
$$a(t) = e^{Ht},$$
where *the Hubble parameter* H *is constant*.
- The basic idea of inflation is that shortly after birth the Universe was dominated by a *constant energy density*.
- This drove an *exponential expansion for a very short period* that caused *rapid cooling* because of the expansion.
- Then at the end of this period the Universe *exited from inflationary conditions and reheated*.
- The mechanism for reheating generally involves the rapid conversion of the constant energy density driving inflation into the mass–energy of more normal particles.
- This then produced a situation *dominated by radiation* rather than vacuum energy.

After inflation, the Universe began evolving in a standard (hot) big bang scenario *(power law dependence of the scale factor on time):*

$$\underbrace{a(t) \sim e^{Ht}}_{\text{inflation}} \longrightarrow \underbrace{a(t) \sim t^n}_{\text{hot big bang}}.$$

In various versions of inflation different reasons are assumed for the initial conditions triggering the exponential expansion.

- The original inflationary idea due to Alan Guth assumed that inflation was driven by a Lorentz scalar field associated with a first-order phase transition.
- This is conceptually simple, but proved to be incompatible with observations (as Guth himself realized).
- It was found that the resulting inflation could not halt in a manner that would give something that looks like the real Universe.
- In subsequent versions of inflation the inflation was often assumed to be driven by a scalar field having a time dependence of a particular form called a *slow rollover transition.*
- Although such theories often give a reasonably good account of data, they suffer from
  1. having little connection to scalar fields known already to exist in elementary particle physics, and
  2. requiring extremely fine empirical tuning of parameters to account well for data.

In keeping with the philosophy outlined above, we omit discussion of these different forms of inflation and instead concentrate on the consequences of inflationary expansion.

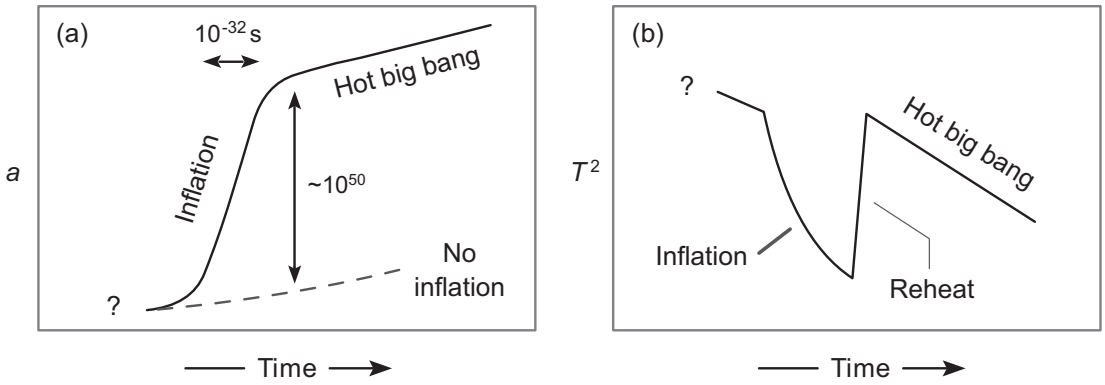


Figure 21.4: Inflationary cosmology. (a) In the inflationary epoch the Universe expands exponentially, which can increase the scale factor by factors of 10^{50} or larger in of order 10^{-32} s. (b) The Universe cools as it expands exponentially. At the end of inflation, some mechanism reheats the Universe, which then continues a standard hot big bang evolution.

Figure 21.4 illustrates the behavior of the scale factor and the temperature in highly schematic fashion during inflation and the following big bang evolution.

- During inflation the Universe *expanded at a much higher rate* than in normal big bang evolution.
- At the same time, the temperature dropped rapidly in the exponentially expanding Universe.
- Finally, when the period of inflation halted the Universe first rapidly reheated and then began to decrease in temperature according to the standard big bang scenario.

The question marks represent our substantial lack of knowledge concerning the Universe prior to the inflationary period.

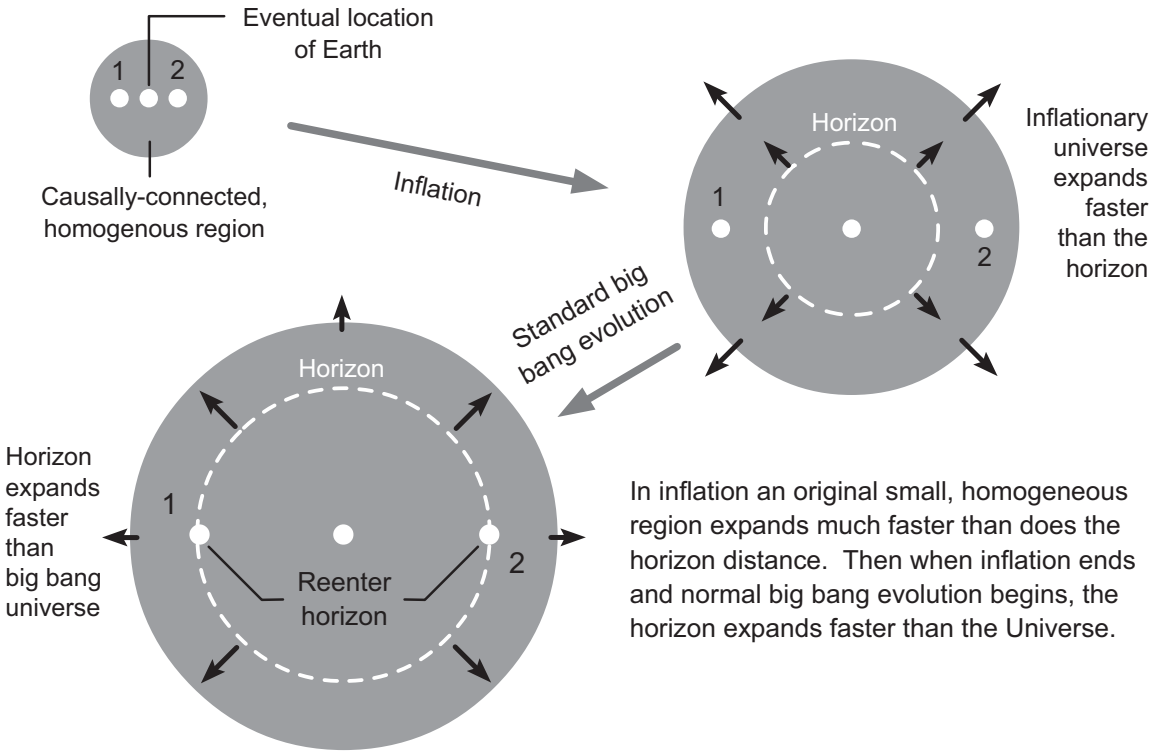


Figure 21.5: Solution of the horizon problem in the inflationary universe.

21.3.2 Taking the Inflationary Cure

The inflationary scenario provides (in principle) a *solution of the four fundamental problems posed above*.

Solution of the Horizon Problem

The *solution of the horizon problem* is illustrated in Fig. 21.5.

- The tremendous expansion means that regions that we see widely separated in the sky now at the horizon were *much closer together before inflation*.
- Thus, they *could have been in contact by light signals*.

Solution of the Flatness Problem

The tremendous *expansion greatly dilutes any initial curvature*.

- Think of standing on a basketball. It would be obvious that you are standing on a two-dimensional curved surface.
- Now imagine expanding the basketball to the size of the Earth.
- As you stand on it now, it will appear to be flat, even though it is actually curved on large scales.
- The same idea extended to four-dimensional spacetime accounts for the present flatness (lack of curvature) in the space of the Universe.

Out to the greatest distances that we can see the Universe looks flat on large scales, just as the Earth looks approximately flat out to our horizon.

Solution of the Monopole Problem

The rapid expansion of the Universe tremendously *dilutes the concentration of any magnetic monopoles* that are produced.

- They become so rare in any given volume of space that we would be very unlikely to ever encounter one in an experiment designed to search for them.
- Nor would they have sufficient density to alter the gravity and thereby the normal expansion of the Universe following inflation.

Two observational properties,

4: Smoothness of the CMB and

8: Evolved large-scale structure,

present a *compatibility issue:*

- The *remarkably high smoothness of the CMB* implies that the early Universe was largely devoid of density perturbations.
- Where then did the *density perturbations responsible for the growth of rich structure* in the present Universe on the supercluster and smaller scale originate?

The form of the density perturbations required to give the observed structure is known empirically, so they could be *postulated as initial conditions*, but again one would like to know *why.*

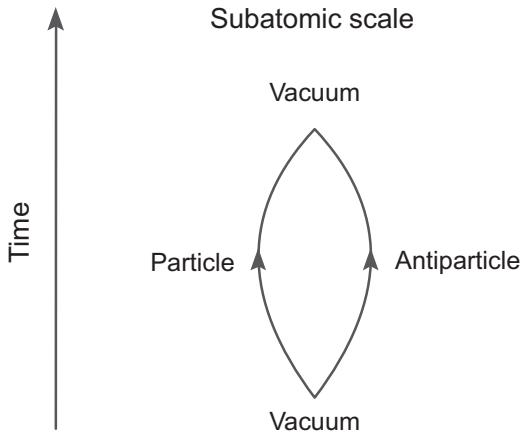
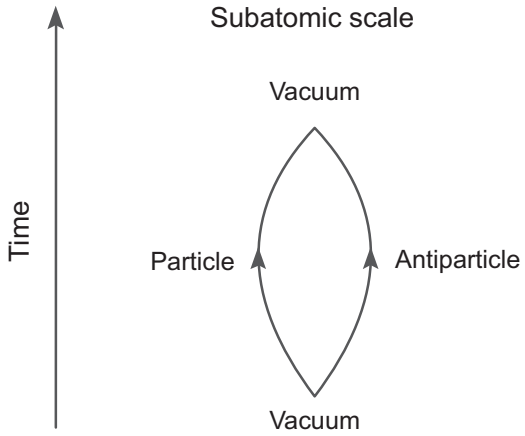


Figure 21.6: Quantum fluctuations can create a particle-antiparticle pair from the vacuum for a fleeting instant. In inflation such microscopic fluctuations can be stretched to macroscopic scales, producing density perturbations that can later seed the formation of large-scale structure.

The Structure and Smoothness Dichotomy

Perhaps the most important consequence of inflation is that it provides a possible *explanation for the origin of large-scale structure*.

- The inflationary explanation is in fact *rather remarkable*.
- During inflation *quantum fluctuations* (Fig. 21.6)
- are *stretched from microscopic to macroscopic size* by the exponential expansion.
- Because this process occurs during the entire time of inflation, we end up with
  - density fluctuations of *macroscopic size*
  - that *vary over many length scales*.



- Technically, this produces what is termed a *scale-invariant spectrum* of density fluctuations.
- This is known as a *Harrison–Zeldovich spectrum*.
- This is the spectrum of density perturbations is *most likely to give the observed large-scale structure* of the Universe.
- Because of the quantum nature of the fluctuations, they also are *gaussian*.
- "Gaussian" means that if fluctuations are expanded in a fourier series or in spherical harmonics, *different wavenumbers or multipole orders are not correlated*.
- These density perturbations will generally be *expanded beyond the horizon* during inflation.
- But after inflation is over and normal big bang evolution sets in, *the horizon grows with time*.
- Eventually the density perturbations *re-enter the horizon* and serve as *nucleation centers for structure formation*.

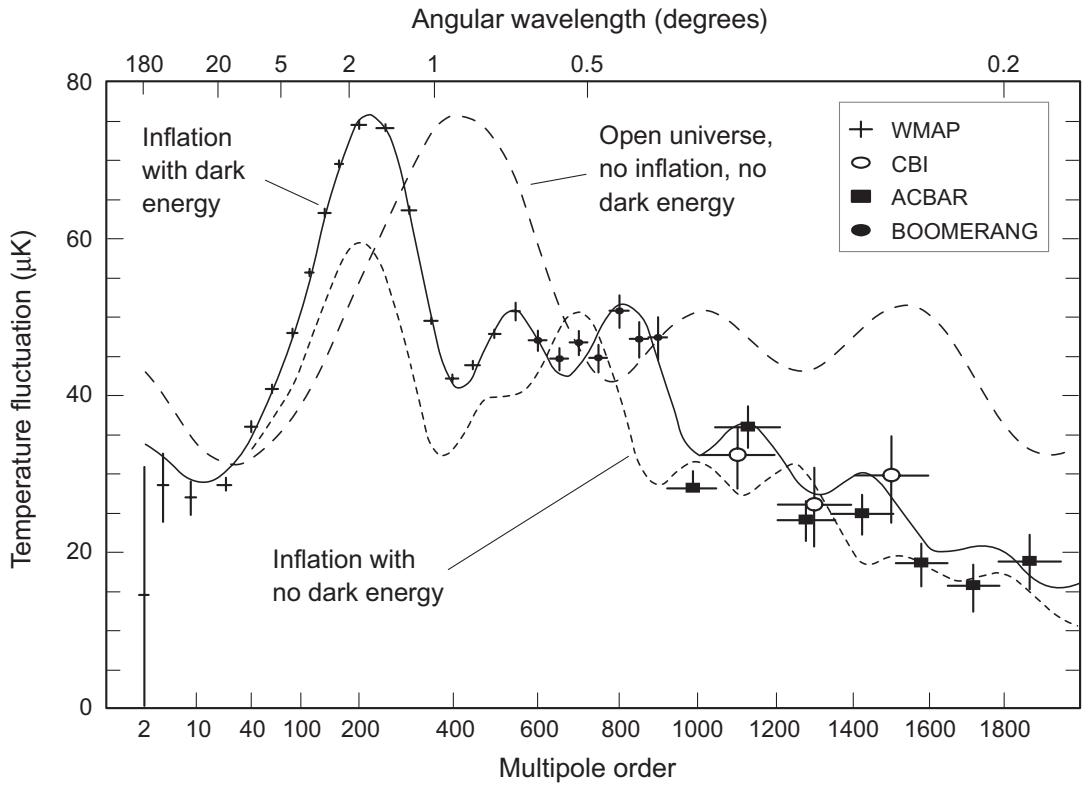


Figure 21.7: Temperature fluctuations in the CMB from satellite and high-altitude balloon observations compared with various theoretical models.

- Simulations of structure formation give reasonable results when effects of inflation are included.
- Furthermore, fluctuations in the CMB are described best by theories that include the effect of inflation.
  - Fig. 21.7 compares the angular fluctuations in temperature for the CMB with various models *with and without inflation*, and *with and without dark energy*.
  - Models with *both dark energy and inflation* are clearly favored.

21.3.3 Inflation Doesn't Replace the Big Bang

Inflation is *not a theory in competition with the big bang:*

- The theory of inflation modifies only the first tiny instants of creation.
- After the completion of the brief period of inflation, it is assumed that big bang evolution proceeds as described earlier.

Thus, inflation should be viewed as

- a *modified form of the big bang theory* that accounts for
- effects on *initial conditions*

not included in the standard big bang theory.

21.4 The Origin of the Baryons

In the standard big bang model

- the observed
  - preponderance of *matter over antimatter* and
  - preponderance of *photons over baryons*

  are introduced through *initial conditions* without fundamental justification.
- It would be highly desirable to *understand the origin of the baryons* from a deeper perspective.
- This has proved to be a *highly-elusive goal*.

Andrei Sakharov first enumerated the ingredients required to generate *baryon asymmetries* within the standard big bang model:

1. There must exist elementary particle interactions in the Universe that *do not conserve baryon number* N_{B}.
2. There must exist interactions that violate both charge conjugation symmetry (C) and the product of charge conjugation and parity (P) symmetries, denoted CP.
3. There must be *departures from thermal equilibrium* during the evolution of the Universe.

In these requirements

- *Baryon number* N_{B} takes the value $+1$ for a baryon and -1 for an antibaryon.
- Total baryon number is then the algebraic sum of these numbers for all the particles in a reaction.
- *Conservation of baryon number* (observed in every experiment so far) means that this number does not change in the interaction.
- *Charge conjugation* symmetry (C) is symmetry under *exchange of a particle with its antiparticle*.
- *Parity* symmetry (P) is symmetry under *inversion of the spatial coordinate system*.
- CP symmetry is symmetry under inversion of the coordinate system *and* exchange of particle with antiparticle.

Most interactions conserve these symmetries to high precision but the weak interactions are known to violate P, C, and CP.

- Departures from thermal equilibrium are likely to have occurred at various times in the evolution of the Universe.
- At least the weak interactions are known to violate both C and CP symmetry.

Thus (in principle) all ingredients exist to account for baryon asymmetry except for *baryon non-conserving reactions*.

- Experimentally, *baryon non-conservation has never been observed*.
- However, there are theoretical reasons to believe that *baryon conservation might not be an exact symmetry* but just one that has not yet been observed to be violated.
- For example baryon non-conservation may occur only on a energy scale not yet reached in laboratory experiments but that could have occurred in the early Universe.

One class of elementary particle theories that could have played a role in producing the baryons is that of *Grand Unified Theories (GUTs).*

- In the *Standard Electroweak Theory* of elementary particle physics the electromagnetic interactions and the weak interactions have been (partially) unified.
- This means that at high enough energy (in this case a scale of about 100 GeV, (where $1\,\text{GeV} = 10^9\,\text{eV}$) the weak and electromagnetic interactions *take on the same properties*.
- A GUT attempts to extend this idea to unify weak, electromagnetic, and strong interactions into a single unified theory.
- The characteristic *GUT energy scale* is very high (10^{14-15} GeV is a common estimate), but on that scale GUTs typically *violate baryon number* strongly.
- At one time GUTs were favored to account for baryogenesis but there have since been shown to be difficulties with this approach.
- Thus a viable theory of baryon asymmetry may require
  - a *baryon-violating phase transition* at
  - a *lower energy than the GUTs scale*.

A possible alternative mechanism is *leptogenesis*.

- *Leptogenesis* postulates that perhaps the baryon asymmetry was generated by *strongly CP-violating processes in the neutrino sector*.
- But it is not clear that this can account for the observed baryon asymmetry of the Universe because
- the known electroweak interactions do not exhibit interactions with the required characteristics.
- However, the correct electroweak theory could be more general than the present one,
  - which is *not tested exhaustively above 100 GeV*, and
  - is now known to be *contradicted by the small but finite masses* observed for neutrinos.

Thus, an *improved electroweak theory* eventually might be able to account for the baryon asymmetry, but that is *mostly speculation* at this point.

Part IV

Gravitational Wave Astronomy

Chapter 22

Gravitational Waves

- *Maxwell's theory* permits the existence of *propagating waves in the electromagnetic field*.
- Likewise, Einstein's theory of general relativity predicts that *fluctuations in the metric of spacetime* can propagate as *gravitational waves.*
- Until 2015 gravitational waves could only be inferred indirectly.
- That changed dramatically with the detection in September, 2015, of *GW150914* by the Laser Interferometry Gravitational Wave Observatory (LIGO).
- GW150914 was interpreted as a gravitational wave generated by *merger of two* $\sim 30 M_\odot$ *black holes*.
- Thus was *gravitational wave astronomy* born.

Gravitational waves are of considerable current interest for several reasons.

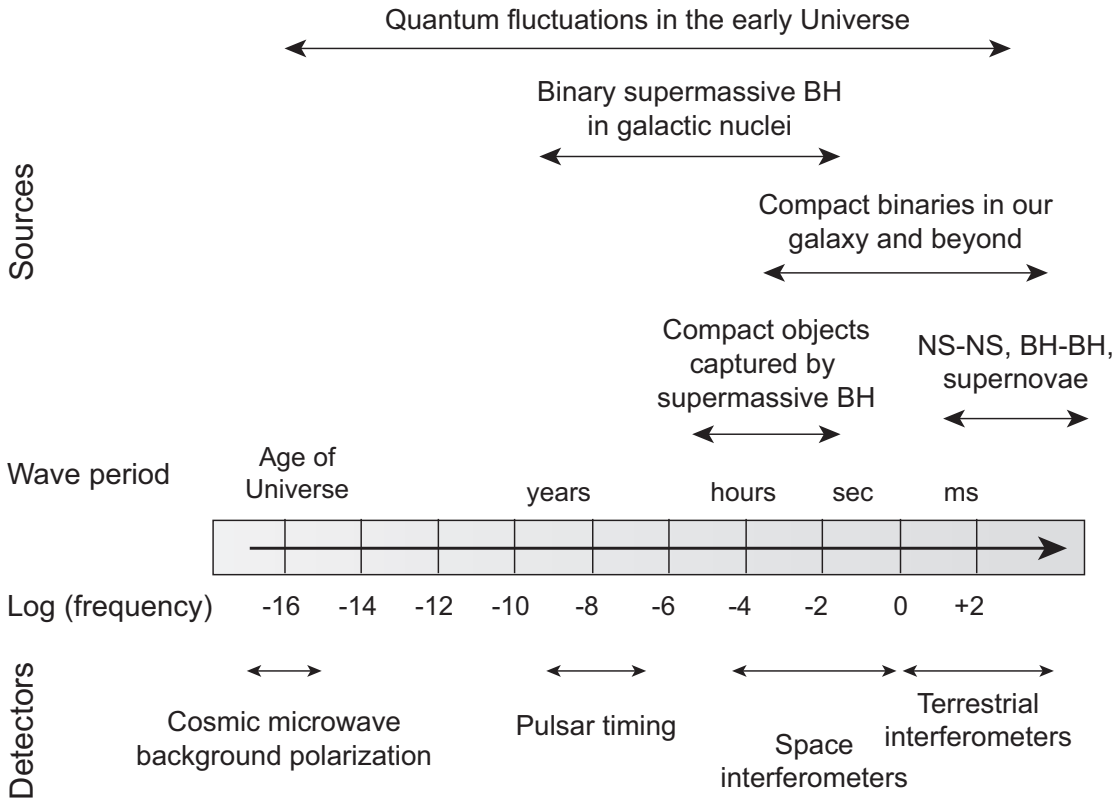


Figure 22.1: The gravitational wave spectrum with potential sources and methods of detection. Black holes are denoted by BH and neutron stars by NS. Terrestrial interferometers, space-based interferometers, and pulsar timing arrays for detecting gravitational waves will be discussed later. Cosmic microwave background polarization refers to expected signatures imprinted on the CMB by gravitational waves.

22.1 Significance of Gravitational Waves

The schematic *spectrum of gravitational waves*, their possible *origin* in various astronomical events, and potential *methods of detection* are illustrated in Fig. 22.1.

Unprecedented Tests of General Relativity

Before the observation of GW150914,

- gravitational waves represented the *last classical prediction of general relativity not tested directly*.
- The *Binary Pulsar* provides strong *indirect evidence* for emission of gravitational waves from that system, but no gravitational wave had been observed directly.
  - This is *not because gravitational waves are rare;*
  - it is because they *interact so weakly with matter*.
- General relativity is tested by the mere *existence of gravitational waves* and by their *detailed properties*.
- For example, GR predicts that gravitational waves
  - can *have only two states of polarization* and
  - must *travel at light velocity*.
- Some alternative theories of gravity predict gravitational waves with *more polarization states and with* $v \neq c$.
- Furthermore, detection of GW150914 and its interpretation as a *binary black hole merger* provides the first test of GR in the *strong gravity, high-velocity, nonlinear limit*.
- All prior direct tests: deflection and redshift of light, precession of orbits, and so on, correspond to the relatively *weak gravity limit* of general relativity.

A Probe of Dark Events

Gravitational waves can signal events that release *large gravitational energy* but *little electromagnetic radiation*.

- Indeed, the evidence is that GW150914 was
  - an enormously violent event
  - that converted $\sim 200 M_\odot$ per second into gravitational waves at peak luminosity, for which
  - *no obvious electromagnetic counterpart* was observed.
- Thus, gravitational waves may provide an alternative—or perhaps the only—probe of some events releasing large gravitational energy.

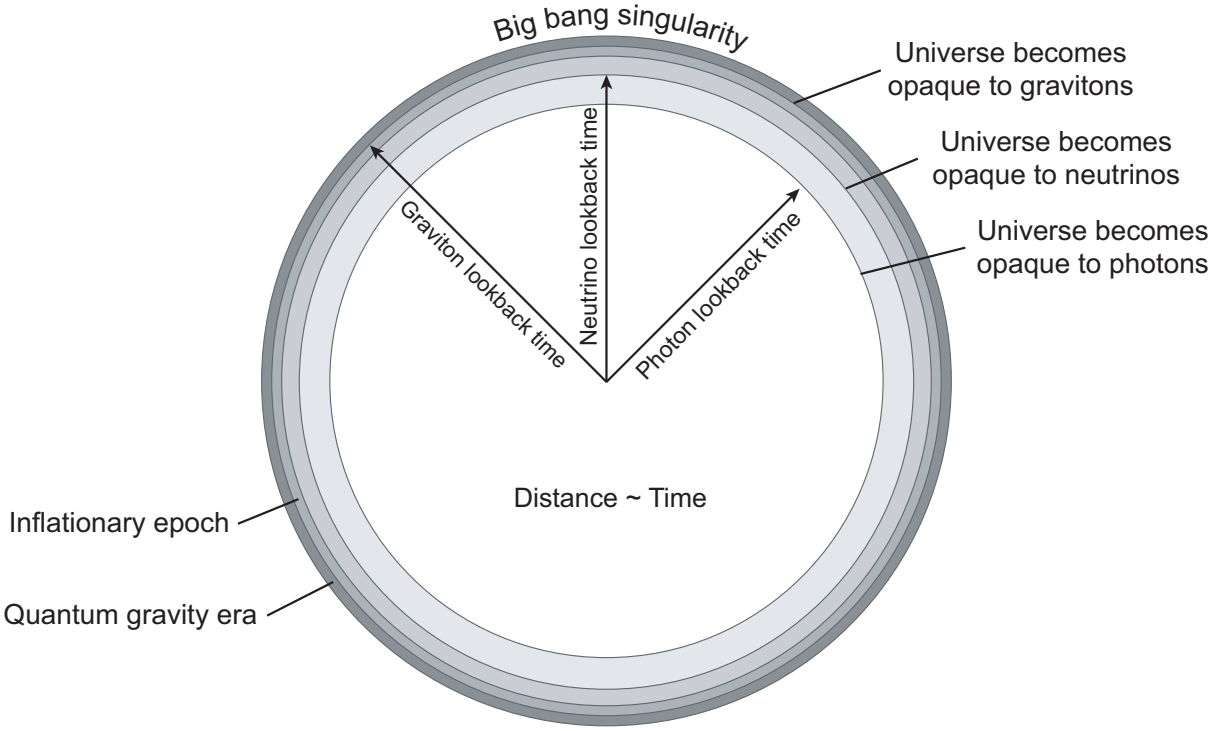


Figure 22.2: Maximum lookback times (not to scale) represent the earliest time since the big bang from which a particular probe could be seen.

The Deepest Probe

Their weakness makes gravitational waves hard to detect but

- That same weakness means that gravitational waves can in principle be seen from earlier epochs (Fig. 22.2).
- *Photons decouple* $\sim 10^5$ *yr after the big bang*, so we can't see direct photons from earlier periods.
- *Neutrinos decouple* ~ 1 *s after the big bang*, so neutrinos can probe only back to that time.
- Gravitational waves could probe *near the Planck scale*.

22.2 Linearized Gravity

The *Einstein equation* may be expressed in the form (Problem)

$$R_{\mu\nu} = -8\pi G(T_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}T^{\lambda}_{\lambda}).$$

- Because this equation is *non-linear*, general solutions that correspond to gravitational waves are difficult to obtain.
- If we *assume the gravitational waves to be weak*, the metric may be approximated as

$$g_{\mu\nu}(x) = \eta_{\mu\nu} + h_{\mu\nu}(x) \qquad \text{(weak gravity)}$$

  - where $\eta_{\mu\nu}$ is the *metric of (flat) Minkowski space* and
  - the rank-2 Minkowski tensor $h_{\mu\nu}$ *is small*.

- The *linearized vacuum Einstein equation* then results from
  - inserting this approximate metric into the *vacuum Einstein equation*,

$$R_{\mu\nu} = 0$$

    (obtained by setting the stress–energy tensor $T_{\mu\nu}$ to zero) and
  - *expansion* of the resulting equations to *first order in the small quantity* $h_{\mu\nu}$.

22.2.1 Linearized Curvature Tensor

- The *Ricci curvature tensor* $R_{\mu\nu}$ is given by

$$R_{\mu\nu} = \Gamma^{\gamma}_{\mu\nu,\gamma} - \Gamma^{\gamma}_{\mu\gamma,\nu} + \Gamma^{\gamma}_{\mu\nu}\Gamma^{\sigma}_{\gamma\sigma} - \Gamma^{\sigma}_{\mu\gamma}\Gamma^{\gamma}_{\nu\sigma},$$

  where the *Christoffel symbols* $\Gamma^{\gamma}_{\mu\nu}$ are related to the *metric tensor* $g_{\mu\nu}$ by

$$\Gamma^{\gamma}_{\mu\nu} = \frac{1}{2}g^{\gamma\delta}\left(\frac{\partial g_{\nu\delta}}{\partial x^{\mu}} + \frac{\partial g_{\mu\delta}}{\partial x^{\nu}} - \frac{\partial g_{\mu\nu}}{\partial x^{\delta}}\right).$$

- To *zeroth order in* $h_{\mu\nu}$, the Christoffel coefficients vanish and so does $R_{\mu\nu}$, since $\partial g/\partial x = 0$ for $g_{\mu\nu} \to \eta_{\mu\nu}$.

- To *first order in* $h_{\mu\nu}$,

$$\delta\Gamma^{\gamma}_{\mu\nu} = \frac{1}{2}\eta^{\gamma\delta}\left(\frac{\partial h_{\nu\delta}}{\partial x^{\mu}} + \frac{\partial h_{\mu\delta}}{\partial x^{\nu}} - \frac{\partial h_{\mu\nu}}{\partial x^{\delta}}\right).$$

- The last two terms in the Ricci tensor are *quadratic in* ∂h,

$$R_{\mu\nu} = \Gamma^{\gamma}_{\mu\nu,\gamma} - \Gamma^{\gamma}_{\mu\gamma,\nu} \underbrace{+ \Gamma^{\gamma}_{\mu\nu}\Gamma^{\sigma}_{\gamma\sigma} - \Gamma^{\sigma}_{\mu\gamma}\Gamma^{\gamma}_{\nu\sigma}}_{\text{quadratic in } \partial h},$$

  and may be *discarded to first order in* h, giving

$$\delta R_{\mu\nu} = \frac{\partial(\delta\Gamma^{\lambda}_{\mu\nu})}{\partial x^{\lambda}} - \frac{\partial(\delta\Gamma^{\lambda}_{\mu\lambda})}{\partial x^{\nu}} + \mathcal{O}\left(h^2\right),$$

  for the *linearized Ricci tensor*.

22.2.2 Wave Equation

Substitution and some algebra yields (Problem)

$$\delta R_{\mu\nu} = \tfrac{1}{2}(-\Box h_{\mu\nu} + \partial_\mu V_\nu + \partial_\nu V_\mu),$$

where the *4-dimensional Laplacian (d'Alembertian operator)* is

$$\Box \equiv \eta^{\mu\nu}\partial_\mu\partial_\nu = -\frac{\partial^2}{\partial t^2} + \boldsymbol{\nabla}^2,$$

with the definitions

$$\partial_\mu \equiv \frac{\partial}{\partial x^\mu} \qquad \partial^\mu \equiv \frac{\partial}{\partial x_\mu},$$

and where the V_ν are given by

$$\begin{aligned} V_\nu &\equiv \partial_\lambda h^\lambda_\nu - \tfrac{1}{2}\partial_\nu h^\lambda_\lambda \\ &= \partial_\lambda \eta^{\lambda\delta} h_{\delta\nu} - \tfrac{1}{2}\partial_\nu \eta^{\lambda\delta} h_{\delta\lambda}. \end{aligned}$$

Note that *raising and lowering of indices in linearized gravity* uses *contraction with the flat-space metric* $\eta_{\mu\nu}$,

$$h^\gamma_\nu \equiv \eta^{\gamma\delta} h_{\delta\nu}$$

rather than through contraction with

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}.$$

Thus the vacuum Einstein equation to this order yields the *wave equation* (Problem)

$$\Box h_{\mu\nu} - \partial_\mu V_\nu - \partial_\nu V_\mu = 0.$$

- Since $h_{\mu\nu}$ *is symmetric*, this constitutes a set of 10 *linear partial differential equations* for the metric perturbation $h_{\mu\nu}$.
- One refers to these as the *linearized vacuum Einstein equations* and to the resulting theory as *linearized gravity*.

As is clear from the derivation, we may expect this to be a valid approximation to the full gravitational theory when the metric departs only slightly from that of flat spacetime.

22.2.3 Degrees of Freedom and Gauge Transformations

The equation

$$\Box h_{\mu\nu} - \partial_\mu V_\nu - \partial_\nu V_\mu = 0.$$

cannot yield unique solutions in its present form because of the freedom of coordinate transformations:

- Given one solution (metric), we may generate another by a coordinate transformation.
- This is related to a *similar ambiguity in electromagnetism (EM)* associated with *freedom to make gauge transformations* without altering the electric and magnetic fields.
- If $h_{\mu\nu}$ solves the linearized Einstein equations,
  - then so does $h'_{\mu\nu}$ where
  - $h'_{\mu\nu}$ is related to $h_{\mu\nu}$ through a *general coordinate transformation* $x \to x'$.
- This is reminiscent of the gauge ambiguity in EM, which can be removed by *fixing (choosing) a specific gauge*.
- (If you are unfamiliar with this concept, see the discussion of the Maxwell equations in Ch. 4.)
- The "gauge ambiguity" in linearized gravity can be removed in a similar way by *fixing the coordinate system*.

The *coordinate independence of GR* is analogous to the *gauge ambiguity in electromagnetism*.

- The symmetric tensor $h_{\mu\nu}$ has 10 *independent components,* but
- gravitational waves have only *2 independent degrees of freedom*.
- However, a judicious choice of coordinate system yields *8 constraints allowing unique solutions* of

$$\Box h_{\mu\nu} - \partial_\mu V_\nu - \partial_\nu V_\mu = 0.$$

that have only *2 physical degrees of freedom.*

The *freedom to make gauge (coordinate) transformations* is of crucial importance in understanding gravitational waves, so let's now examine this in a little more detail.

22.2.4 Choice of Gauge

In electromagnetism (EM) we can make *different choices of the (vector and scalar) potentials that give the same electric and magnetic fields* and thus the same classical observables.

- Something similar is possible in linearized gravity.
- *Small coordinate changes can be made* that don't change $\eta_{\mu\nu}$ in
$$g_{\mu\nu}(x) = \eta_{\mu\nu} + h_{\mu\nu}(x),$$
but that alter the functional form of $h_{\mu\nu}$.
- Under such *small changes in the coordinates*
$$x^\mu \to x'^\mu = x^\mu + \varepsilon^\mu(x),$$
where $\varepsilon^\mu(x)$ is similar in size to $h_{\mu\nu}$.
- Then the metric is changed to,
$$\begin{aligned} g_{\mu\nu}(x) = \eta_{\mu\nu} + h_{\nu\mu}(x) &\to \eta_{\mu\nu} + h'_{\mu\nu}(x) \\ &= \eta_{\mu\nu} + h_{\mu\nu}(x) - \partial_\mu \varepsilon_\nu - \partial_\nu \varepsilon_\mu. \end{aligned}$$
- The transformation $h_{\mu\nu} \to h'_{\mu\nu}$, with
$$h'_{\mu\nu} = h_{\mu\nu}(x) - \partial_\mu \varepsilon_\nu - \partial_\nu \varepsilon_\mu,$$
is termed a *gauge transformation*, by analogy with EM.
- If $h_{\mu\nu}$ is a solution of
$$\Box h_{\mu\nu} - \partial_\mu V_\nu - \partial_\nu V_\mu = 0.$$
then *so is* $h'_{\mu\nu}$.

Table 22.1: Gauge invariance in linearized gravity and electromagnetism

	Linearized gravity	Electromagnetism[†]
Potentials	$h_{\mu\nu}$	Vector potential: $\boldsymbol{A}(t,\boldsymbol{x})$ Scalar potential : $\Phi(t,\boldsymbol{x})$
Fields	Linearized Riemann curvature: $\delta R_{\alpha\beta\delta\gamma}(x)$	Electric field: $\boldsymbol{E}(t,\boldsymbol{x})$ Magnetic field: $\boldsymbol{B}(t,\boldsymbol{x})$
Gauge transformation	$h_{\mu\nu} \to h_{\mu\nu} - \partial_\mu \varepsilon_\nu - \partial_\nu \varepsilon_\mu$	$A^\mu \to A^\mu - \partial^\mu \chi$
Example of a gauge condition ("Lorentz")	$\partial_\nu \bar{h}^{\mu\nu} = 0$	$\partial^\mu A_\mu = 0$
Field equations in Lorentz gauge	$\Box \bar{h}_{\mu\nu} = 0$	$\Box A^\mu = 0$

[†]The 4-vector potential is $A^\mu \equiv (\Phi, \boldsymbol{A})$ and χ is an arbitrary scalar function.

- *Gauge transformations in E&M* lead to
  - *new potentials* A^μ but
  - the *same electric fields* $\boldsymbol{E}$ *and magnetic fields* $\boldsymbol{B}$.
- Likewise, *gauge transformations in gravity* lead to
  - *new potentials* $h_{\mu\nu}$ but
  - the *same fields* [linearized form $\delta R_{\mu\nu\beta\gamma}(x)$ of the Riemann curvature tensor].

Analogies between

- coordinate ("gauge") transformations in *linearized gravity* and
- gauge transformations in *classical E&M*

are *summarized in Table 22.1*.

A standard gauge choice *analogous to choosing Lorentz gauge in electromagnetism* permits the linearized vacuum gravitational equations to be replaced by the two equations

$$\Box \bar{h}_{\mu\nu}(x) = \left(-\frac{\partial^2}{\partial t^2} + c^2\boldsymbol{\nabla}^2\right)\bar{h}_{\mu\nu} = 0,$$
$$\partial_\nu \bar{h}^{\mu\nu}(x) = 0,$$

- where the first is a *wave equation* corresponding to the linearized Einstein equation,
- the second is a (Lorentz) *gauge constraint*, and
- the *trace-reversed amplitude* is defined by

$$\bar{h}_{\mu\nu} \equiv h_{\mu\nu} - \tfrac{1}{2}\eta_{\mu\nu}h,$$

where $h \equiv h^\alpha_\alpha$ is the *trace (sum of diagonal elements)*.

As in electromagnetism, the "Lorentz" gauge is really *a family of gauges*. This will be used to simply things further below.

22.3 Weak Gravitational Waves

Let us now seek *solutions of*

$$\Box \bar{h}_{\mu\nu}(x) = 0,$$
$$\partial_\nu \bar{h}^{\mu\nu}(x) = 0.$$

We expect the solution to be a superposition of components in the form

$$\bar{h}_{\mu\nu}(x) = \alpha_{\mu\nu} e^{ik\cdot x} = \alpha_{\mu\nu} e^{ik_\alpha x^\alpha},$$

where

- $\alpha_{\mu\nu}$ is a constant, symmetric 4×4 matrix called the *polarization tensor* and
- k is the *4-wavevector*,

$$k^\mu = \left(\frac{\omega}{c}, \boldsymbol{k}\right),$$

with ω the *wave angular frequency*.

22.3.1 States of Polarization

The polarization tensor has *10 independent components*.

- However, requiring that

$$\bar{h}_{\mu\nu}(x) = \alpha_{\mu\nu} e^{ik\cdot x},$$

satisfy the gauge condition $\partial_\nu \bar{h}^{\mu\nu}(x) = 0$ implies that

$$ik^\mu \alpha_{\mu\nu} \exp(ik\cdot x) = 0,$$

which is true generally only if

$$k^\mu \alpha_{\mu\nu} = 0 \qquad (\nu = 0,1,2,3).$$

- These four equations *reduce the independent components of* α^ν_μ *from ten to six*.
- But we have *not yet exhausted the gauge (coordinate) degree of freedom* because any coordinate transform

$$x^\mu \to x'^\mu = x^\mu + \varepsilon^\mu(x)$$

that leaves

$$\partial_\nu \bar{h}^{\mu\nu}(x) = 0$$

valid *does not alter linearized gravity physically*.

We now show that this freedom can be used to *set any four linear combinations of the* $\bar{h}_{\mu\nu}$ *to zero*.

22.3.2 Polarization Tensor in Transverse–Traceless Gauge

The remaining gauge freedom alluded to at the end of the preceding section may be used to transform to *transverse traceless (TT) gauge* by choosing

$$\bar{h}_{0i} = \bar{h}_{ti} = 0 \quad (i = 1,2,3) \qquad \mathrm{Tr}\,\bar{h} \equiv \bar{h}^{\beta}_{\beta} = 0.$$

In terms of the polarization tensor $\alpha_{\mu\nu}$, this corresponds to

$$\alpha_{0i} = 0 \qquad \mathrm{Tr}\,\alpha = \alpha^{\mu}_{\mu} = 0.$$

The gauge conditions

$$k^{\mu}\alpha_{\mu\nu} = 0$$

with $\nu = 0$ and $\alpha_{0i} = 0$ then require that $\alpha_{00} = 0$, so *four of the* $\alpha_{\mu\nu}$ *vanish:*

$$\alpha_{0\mu} = 0.$$

Furthermore, for $i = 1,2,3$ the gauge conditions require the *transversality condition* to be satisfied,

$$k^{j}\alpha_{ij} = 0.$$

> *Note:* In TT gauge $h = \bar{h}$, so *the bar may be dropped on* h *as long as we work in TT gauge.*

Let's take stock. We started with 10 independent components of the symmetric polarization tensor $\alpha_{\mu\nu}$. We then found that

- The condition $\bar{h}_{0i} = 0$ requires $\alpha_{01} = \alpha_{02} = \alpha_{03} = 0$.
- The requirement
$$\partial_\nu \bar{h}^{\mu\nu}(x) = 0$$
yields $\alpha_{00} = 0$.
- Therefore, *the four components $\alpha_{0\mu}$ vanish in TT gauge*.
- The trace condition
$$\mathrm{Tr}\, h \equiv h^\beta_\beta = 0.$$
gives *one constraint* and the transversality condition
$$k^j \alpha_{ij} = 0 \qquad (i = 1,2,3)$$
gives *three additional constraints* for a *total of four constraints*.

In summary, in TT gauge,

- ten $\alpha_{\mu\nu}$
- minus four $\alpha_{\mu\nu}$ that are identically zero,
- minus four constraints on $\alpha_{\mu\nu}$,

leave *two independent physical polarizations* for gravitational waves.

22.3.3 Helicity Components

Further insight comes from asking how the $\alpha_{\mu\nu}$ change under rotations of the coordinate system about the z axis.

- Helicity is the *projection of the angular momentum on the direction of motion.*
- Generally a gravitational plane wave can be decomposed into helicity components ± 2, ± 1, 0, and 0.
- However, the components with helicity 0 and ± 1 vanish under a suitable choice of coordinates.
- Thus, *only the helicity components* ± 2 *are physically relevant* for gravitational waves.
- This explains why there are *two independent physical states of polarization*.

It is in this sense that quantum field theory associates gravity with a *spin-2 field*.

Compare with the analogous situation in electromagnetism, which is described by a *4-vector field* A^μ.

- This suggests that the Maxwell field should have *4 independent states of polarization* α_μ.
- However, $k_\mu \alpha^\mu = 0$ reduces this to 3 and
- the freedom to make gauge transformations that leave the $\boldsymbol{E}$ and $\boldsymbol{B}$ fields unchanged demonstrates explicitly that the number of independent polarizations is actually only 2.
- Furthermore, a decomposition under rotations about the z axis yields helicities 0 and ± 1,
- but *only the helicities* ± 1 *are physically relevant*.
- These correspond to the *2 independent states of polarization* for a *massless vector (spin-1) field*.

That there are only two physical states of polarization for the photon is tied intimately to its *masslessness* and associated *local gauge invariance*.

- A *massive* vector field has 3 states of polarization and is not locally gauge invariant.
- Likewise, that the gravitational field exhibits only 2 physical states of polarization is a consequence of the *masslessness of the graviton*.
- A massive spin-2 field would have additional physical polarization states.

22.3.4 General Solution in TT Gauge

Assume the wave to propagate on the z axis with frequency ω.

- The requirement that

$$\bar{h}_{\mu\nu}(x) = \alpha_{\mu\nu} e^{ik\cdot x},$$

satisfy the wave equation

$$\Box \bar{h}_{\mu\nu}(x) = 0,$$

implies that the wavevector satisfies

$$k_\mu k^\mu = 0.$$

Thus k is a *null vector* with

$$k^\mu = (\omega, 0, 0, \omega)$$

- From the preceding two equations with c restored,

$$k^\mu k_\mu = -\frac{\omega^2}{c^2} + \boldsymbol{k}^2 = 0 \quad \longrightarrow \quad \frac{\omega}{|\boldsymbol{k}|} = c.$$

- Thus the *group and phase velocities are equal to* c, just as for electromagnetic waves.

> The *speed of gravity* is exactly *equal to the speed of light*.

As you are asked to show in a Problem, for TT gauge

- The transversality condition $k^j \alpha_{ij} = 0$
- the null wavevector $k^\mu = (\omega, 0, 0, \omega)$,
- the earlier result that $\alpha_{0\mu} = 0$, and
- the trace condition $\mathrm{Tr}\,\alpha = 0$,

mean that the *only nonvanishing components* of α are

$$\alpha_{11} \qquad \alpha_{12} = \alpha_{21} \qquad \alpha_{22} = -\alpha_{11}.$$

Furthermore, from $k^\mu = (\omega, 0, 0, \omega)$ we have

$$ik \cdot x = -i(\omega t - \omega z)$$

and the general solution for z axis propagation with fixed frequency ω in TT gauge is

$$h_{\mu\nu}(t,z) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \alpha_{11} & \alpha_{12} & 0 \\ 0 & \alpha_{12} & -\alpha_{11} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} e^{i\omega(z-t)},$$

which exhibits explicitly the *transverse and traceless properties*, with *two independent polarization states*.

Furthermore, since k is a *null vector*, the gravitational wave *propagates at the speed of light*.

- (Note that since we are in TT gauge we have *dropped the bar on* $\bar{h}_{\mu\nu}$.)
- The part of the wave that is proportional to $\alpha_{xx} = \alpha_{11}$ is called the *plus polarization* (denoted by $+$) and
- the part proportional to $\alpha_{xy} = \alpha_{12} = \alpha_{21}$ is called the *cross polarization* (denoted by $\times$).
- For example a *purely cross-polarized plane wave* propagating in the z direction may be represented as

$$h^{\times}_{\mu\nu}(t,z) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & \alpha_{12} & 0 \\ 0 & \alpha_{12} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} e^{i\omega(z-t)}.$$

- The *most general gravitational wave in TT gauge* is a superposition of waves having the form

$$h_{\mu\nu}(t,z) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \alpha_{11} & \alpha_{12} & 0 \\ 0 & \alpha_{12} & -\alpha_{11} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} e^{i\omega(z-t)},$$

  with different ω, directions of propagation, and amplitudes for the two polarizations.

- It may be expressed as

$$h_{\mu\nu}(t,z) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & f_+(t-z) & f_\times(t-z) & 0 \\ 0 & f_\times(t-z) & -f_+(t-z) & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

  - where we assume propagation of the gravitational wave along the z axis, and where
  - $f_+(t-z)$ and $f_\times(t-z)$ are dimensionless functions characterizing the shape and amplitude of the wave.

22.4 Gravitational Waves versus Electromagnetic Waves

Much of the discussion in this chapter has emphasized *similarities between gravitational waves and electromagnetic waves*. However, it is important to point out also that there are some *fundamental differences*.

22.4.1 Interaction with Matter

Electromagnetic waves interact strongly with matter but gravitational waves interact extremely weakly with matter. This has two consequences:

- Gravitational waves are *much harder to detect* than electromagnetic waves.
- Gravitational waves are *not significantly absorbed* by intervening matter.

22.4.2 Wavelength Relative to Source Size

- Electromagnetic waves can be used to form an image.
- This is because their wavelength is typically smaller than the size of the emitting system.
- In contrast, the wavelength of gravitational waves is typically greater than or equal to the source size.
- Thus, they generally cannot be used to form an image.
- Hence *detecting gravitational waves has been likened to hearing sound*.

Ultimately this is because

- electromagnetic waves are generated by *local moving charges* within a larger source, while
- gravitational waves are generated by *bulk motion of the entire source*.

22.4.3 Phase Coherence

Photons are emitted *incoherently* from a typical astronomical source.

- In contrast, gravitons for astrophysical gravitational waves strong enough to detect are generated by bulk motion of large masses, so they are typically *phase-coherent when emitted*.
- In this sense, gravitational waves are similar to laser light.
- This coherence has two important observational consequences.
  - If the waveform is well modeled, *matched filtering techniques* (later chapter) can extend the distance at which sources can be detected.
  - The direct observable for a gravitational wave is the *strain, falling off as* $1/r$ for a source at distance r.

This may be contrasted with observables for incoherent electromagnetic radiation, which are typically *energy fluxes falling off as* $1/r^2$.

22.4.4 Field of View

Electromagnetic astronomy is typically based on

- deep images with
- small fields of view.
- This allows observers to mine *large amounts of information* from a *small region of the sky*.

In contrast, *gravitational wave astronomy*

- *sees essentially the entire sky* at once, but
- with *relatively low resolution* compared with that typical of electromagnetic astronomy.

> The difference has been likened to the angular resolution contrast between *hearing and seeing*.

22.5 Response of Test Particles to Gravitational Waves

We cannot detect a gravitational wave locally

- In a local enough region the effects of gravity may be transformed away *(equivalence principle).*
- Thus the effect of a gravitational wave on a single point test particle has *no measurable consequences.*

Gravitational waves may be detected only by their *influence on two or more test particles at different locations*.

22.5.1 Response of Two Test Masses

- Assume a linearized gravitational wave of + polarization propagating on the z axis in TT gauge,

$$h_{\mu\nu}(t,z) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} f(t-z),$$

  with a corresponding time-dependent metric

$$ds^2 = -dt^2 + (1+f(t-z))dx^2 + (1-f(t-z))dy^2 + dz^2.$$

- Consider two test masses with
  - *Mass A* initially at rest at the origin, $x^i_{\rm A} = (0,0,0)$,
  - *Mass B* initially at rest at a point $x^i_{\rm B} = (x_{\rm B}, y_{\rm B}, z_{\rm B})$,

  with a gravitational wave propagating on the z axis.
- Since the particles are at rest before the gravitational wave arrives, the *initial 4-velocities* are

$$u_{\rm A} = u_{\rm B} = (1,0,0,0).$$

- For the test masses the *geodesic equation* is given by

$$\frac{d^2x^i}{d\tau^2} = -\Gamma^i_{\mu\nu}u^\mu u^\nu = -\Gamma^i_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau}.$$

The undisturbed spacetime is assumed flat and $\Gamma^i_{\mu\nu}$ vanishes.

- From

$$\delta\Gamma^\gamma_{\mu\nu} = \frac{1}{2}\eta^{\gamma\delta}\left(\frac{\partial h_{\delta\mu}}{\partial x^\nu} + \frac{\partial h_{\delta\nu}}{\partial x^\mu} - \frac{\partial h_{\mu\nu}}{\partial x^\delta}\right).$$

  we have to first order

$$\begin{aligned}\frac{d^2(\delta x^i)}{d\tau^2} &= -\delta\Gamma^i_{\mu\nu}u^\mu u^\nu = -\delta\Gamma^i_{00} \\ &= -\frac{1}{2}\eta^{i\delta}\left(\frac{\partial h_{\delta 0}}{\partial x^0} + \frac{\partial h_{\delta 0}}{\partial x^0} - \frac{\partial h_{00}}{\partial x^\delta}\right),\end{aligned}$$

  where

$$u_{\rm A} = u_{\rm B} = (1,0,0,0).$$

  has been used.

- But in TT gauge $h_{\delta 0} = 0$, so

$$\frac{d^2(\delta x^i)}{d\tau^2} = -\delta\Gamma^i_{00} = 0$$

To this order

- the *coordinate distance* between A and B is not changed by the gravitational wave.
- However, the *proper distance* between A and B is changed.

For example, assume A and B to lie on the x axis and to be separated by a distance L_0. From

$$ds^2 = -dt^2 + (1 + f(t-z))dx^2 + (1 - f(t-z))dy^2 + dz^2.$$

the relevant line element is then

$$ds^2 = -dt^2 + (1 + f(t-z))dx^2,$$

corresponding to the metric

$$g_{\mu\nu} = \begin{pmatrix} -1 & 0 \\ 0 & 1 + h_{11}(t,0) \end{pmatrix},$$

where $h_{xx} = h_{11}$ and we have chosen $z = 0$ at time t.

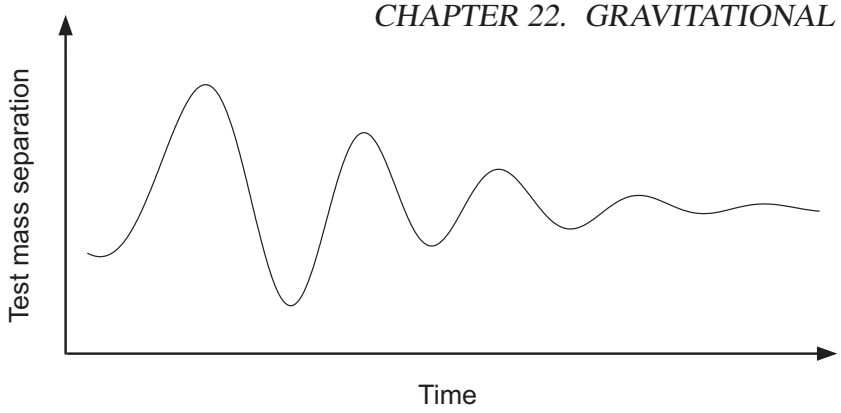


Figure 22.3: Test mass separation perturbed by a gravitational wave.

Then the *proper distance between A and B* is

$$\begin{aligned} L(t) &= \int_0^{L_0} (-\det g)^{1/2}\, dx \\ &= \int_0^{L_0} (1+h_{11}(t,0))^{1/2}\, dx \\ &\simeq \int_0^{L_0} (1+\tfrac{1}{2}h_{11}(t,0))\, dx = \left(1+\tfrac{1}{2}h_{11}(t,0)\right) L_0. \end{aligned}$$

Therefore the *change in distance* between the masses is

$$\frac{\delta L(t)}{L_0} \simeq \tfrac{1}{2}h_{11}(t,0),$$

which oscillates as indicated schematically in Fig. 22.3.

The *amplitude is very small, however!* We expect

$$\frac{\delta L}{L_0} \sim 10^{-21}$$

for typical gravitational waves.

If spacetime is viewed as *an elastic medium*, then from the theory of elastic solids

- the dimensionless quantity $\delta L/L_0$ may be termed the (fractional) *strain* associated with the deforming wave.
- However, this strain is very small, with $\delta L/L_0 \sim 10^{-21}$ for a typical gravitational wave.
- This implies that *spacetime is an extremely stiff medium,* since a large energy is required to create a perceptible distortion of the medium.

Interferometers are designed to measure the strain produced by gravitational waves. This has *two important implications* for such detectors:

1. The gravitational-wave strain produced by astronomical events
   - will be *extremely small*, and this requires
   - measurements of *exacting precision*.
2. Because the strain is *proportional to the amplitude* of the wave,
   - the strain measured on Earth for events at distance r will fall off as r^{-1},
   - not as r^{-2}, as would an energy flux characteristic of most kinds of astronomical observations.

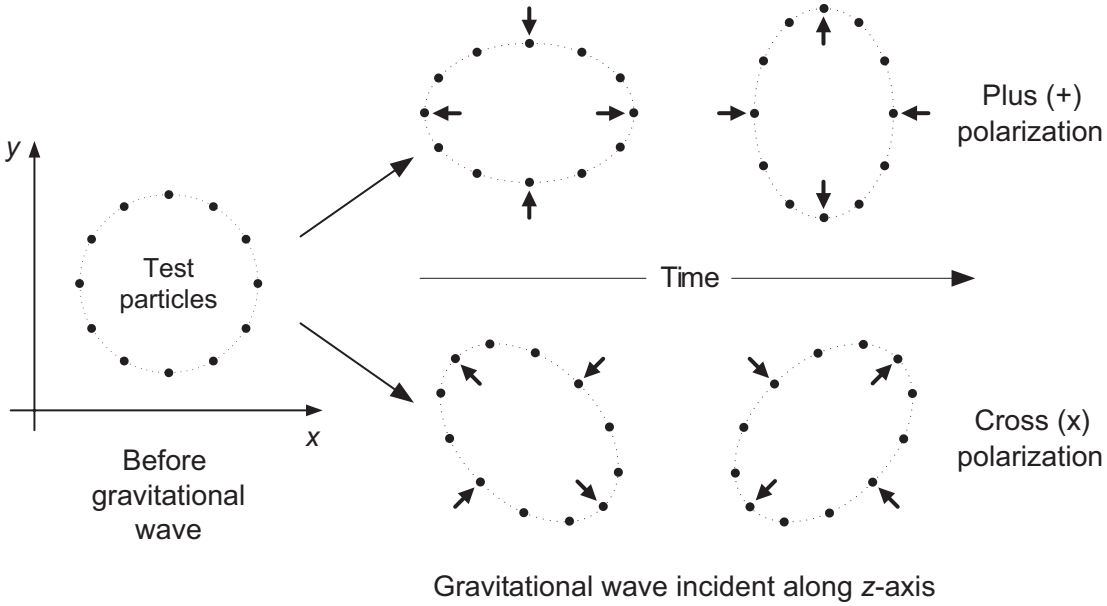


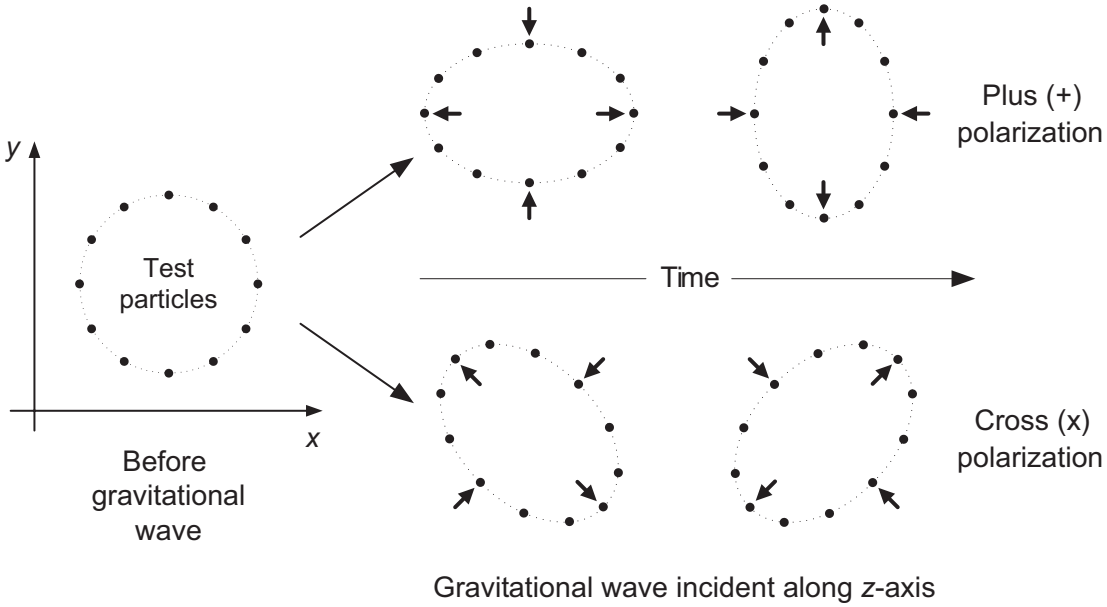
Figure 22.4: Effect of a gravitational wave incident along the z axis on an array of test masses in the x–y plane.

22.5.2 The Effect of Polarization

The influence of polarization may be illustrated by the effect of a gravitational wave on an planar test mass array.

- Like electromagnetic waves, *gravitational waves are transverse*.
- Thus only separations in the transverse directions are changed by the gravitational wave.
- The effect of *purely plus-polarized* and *purely cross-polarized* gravitational waves on an initially circular array of test masses is illustrated in Fig. 22.4.

As in Fig. 22.3, the magnitude is *greatly exaggerated*.



Each polarization is seen to give rise to an elliptical oscillation in the distribution of the test masses.

- The *cross polarization* pattern is rotated by $\frac{\pi}{4}$ relative to the corresponding *plus polarization* pattern.
- The 45^{o} relative rotation results from gravity being described by a *rank-2 tensor field (spin-2 field)*.
- An electromagnetic field corresponds to a *rank-1 tensor* A^{μ} *(spin-1 or vector field)* and
- the rotation angle between the two independent states of polarization is instead 90^{o}.

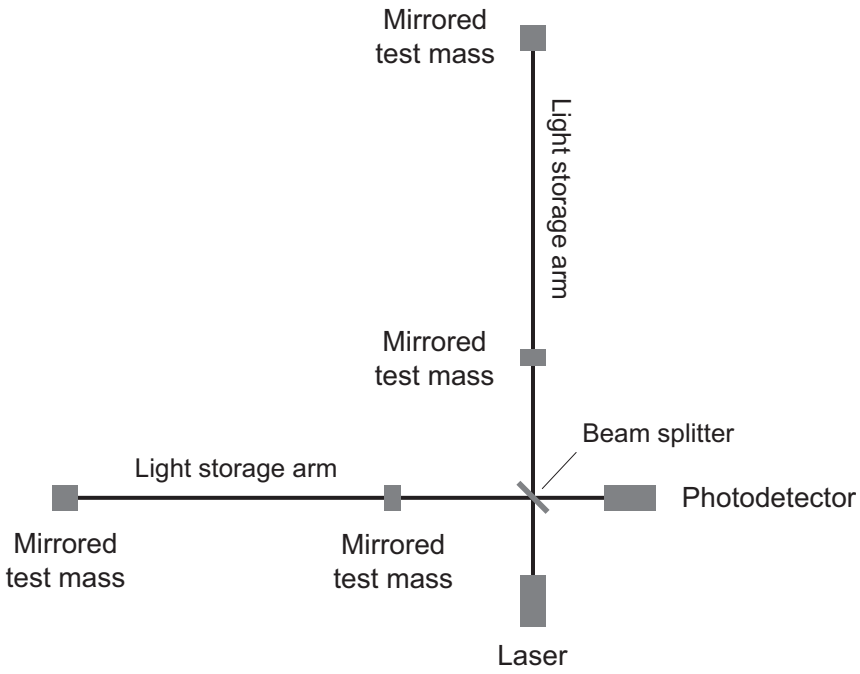


Figure 22.5: Schematic laser interferometer gravity wave detector. In the light storage arms light is multiply reflected, increasing the effective length of the arms.

22.6 Gravitational Wave Detectors

Modern gravitational wave detectors use laser interferometers with kilometer or longer arms (Fig. 22.5).

- Laser light is split and directed down two arms.
- Suspended, mirrored test masses reflect the light at the ends of the arms.
- The reflected light is recombined and
- interference fringes are analyzed for evidence indicating changes in the distances to the test masses.

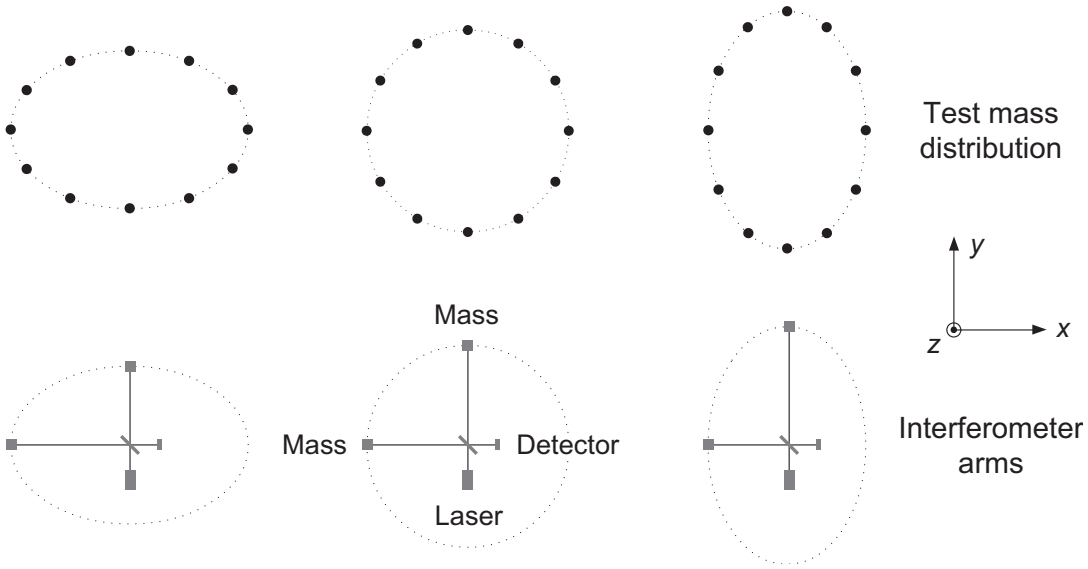


Figure 22.6: Analogy between interaction of a gravitational wave with a test mass distribution and with an interferometer.

Because light is interfering over long path lengths, it is possible to detect extremely small changes in distance.

- This is necessary, since to detect gravitational waves from expected astronomical events fractional changes in distance of order 10^{-21} or smaller must be measured.
- 10^{-21} is approximately the *ratio of the width of a human hair to the distance to Alpha Centauri!*

Laser interferometers may be viewed as extremely precise ways to measure the distortions for a small array of test masses (Fig. 22.6).

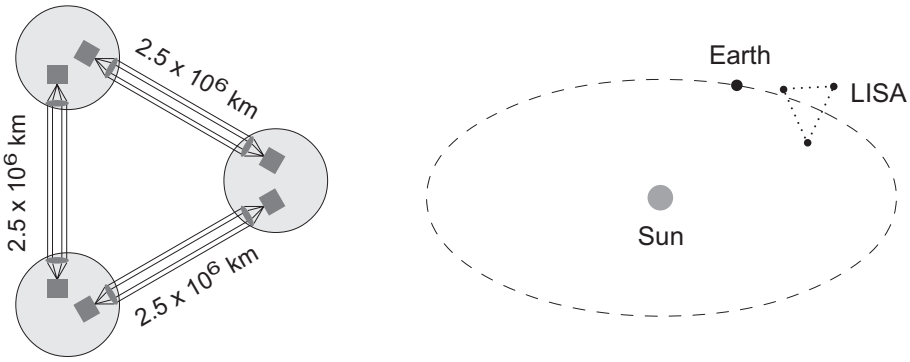


Figure 22.7: The proposed LISA space-based gravitational wave interferometer. This system consists of three drag-free satellites that implement Michelson laser interferometers having 2.5 million kilometer arms. (Unlike for LIGO, the laser arms will not be multiple-reflection cavities.) Space-based interferometers like LISA would be sensitive to much lower frequencies than those on Earth because (1) they can have very long interferometer arms, and (2) because there is little interference from environmental noise (in particular seismic noise), which limits the low-frequency performance of Earth-based systems.

Several gravitational wave laser interferometers are now operational, under construction, or proposed. For example,

- Ground-based detectors such as LIGO and Virgo are now in operation.
- A proposed space-based array, LISA, would have 2.5 million kilometer interferometer arms and could be launched by the 2030s.
- The schematic arrangement for LISA, and its proposed orbit, are illustrated in Fig. 22.7.

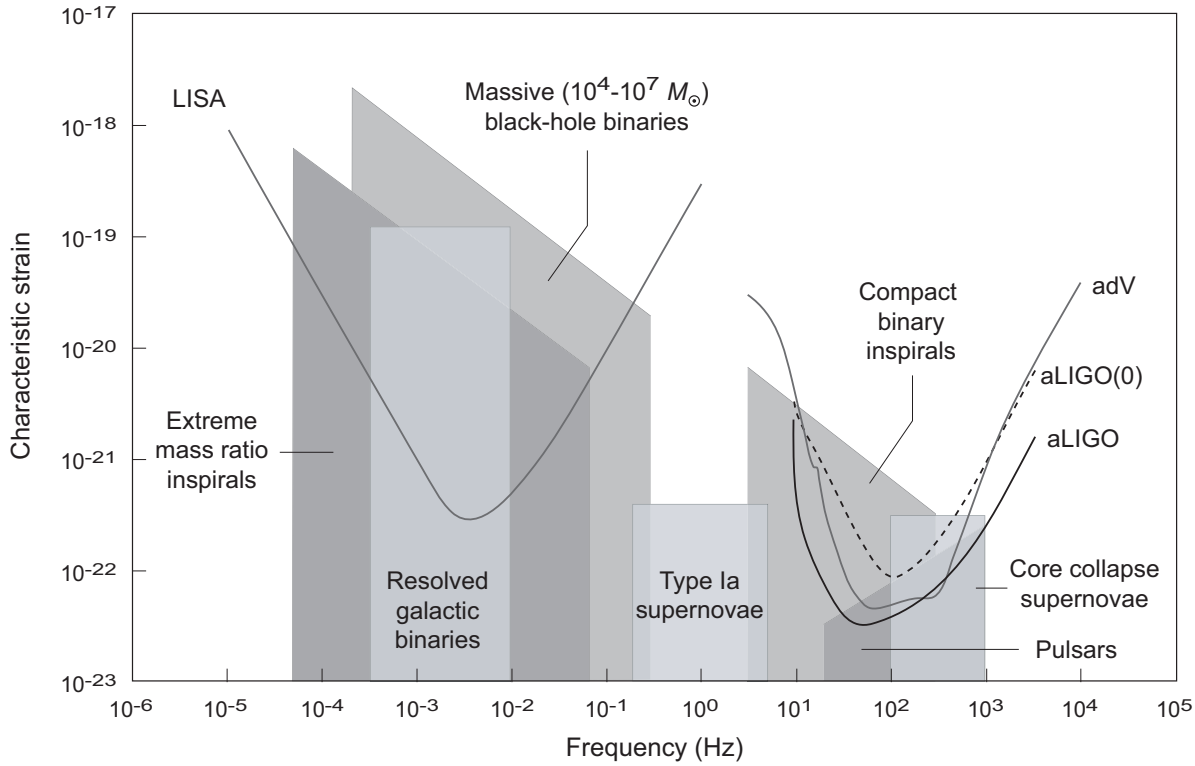
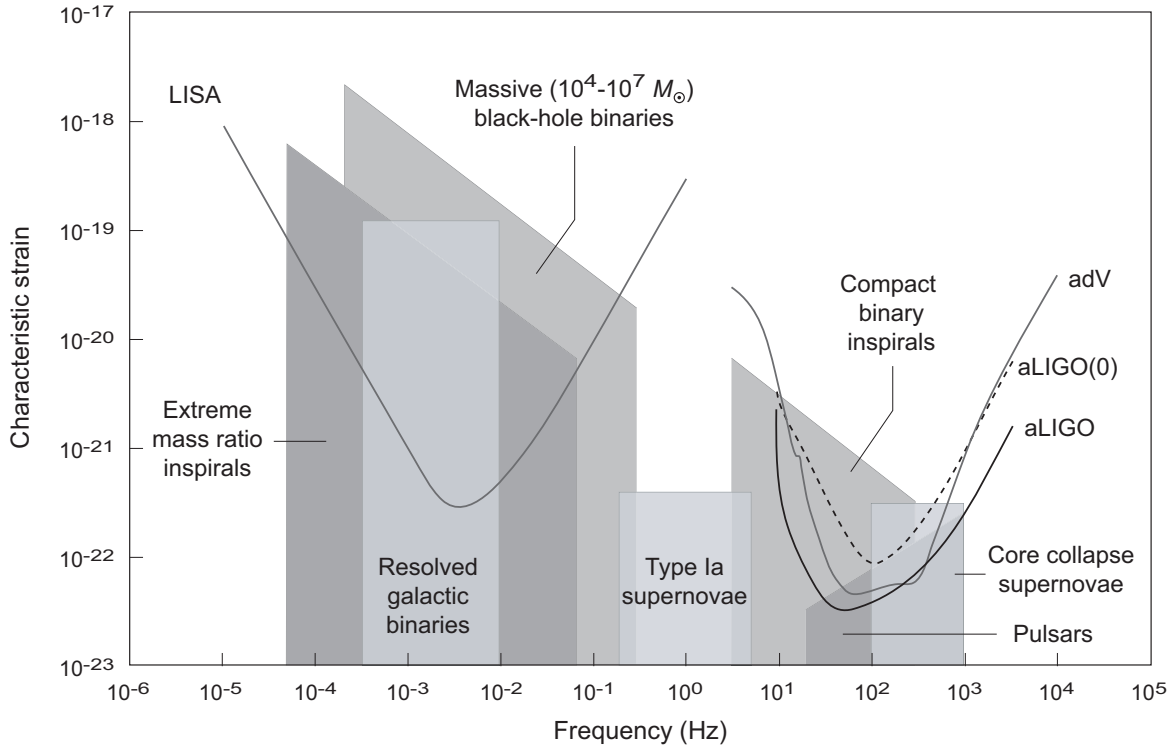


Figure 22.8: Strain amplitude and frequency ranges expected for gravitational waves from various astronomical sources. The minimum strain detection bounds for advanced LIGO (aLIGO) at full design capacity (∼2020), advanced Virgo (adV) at full design capability (∼ 2020), advanced LIGO in the first observing run after the upgrade [aLIGO(0), indicated by the dashed curve], during which the gravitational wave GW150914 was observed in 2015, and the proposed space-based array LISA are indicated.

- Amplitude and frequency windows for LIGO, Virgo, and LISA, and ranges expected for important sources of gravitational waves, are illustrated in Fig. 22.8.
- Space-based interferometers like LISA can go to much lower frequencies because they can have *very long arms* and *little environmental noise*.



22.6.1 Detecting Very Long Wavelengths

The figure shown above omits long-wavelength sources with frequencies below 10^{-6} Hz. Sources of considerable astronomical interest in this frequency range include

1. Mergers of very massive black holes found in the centers of galaxies ($\sim 10^9 M_\odot$), which are expected to occur in the 10^{-7} to 10^{-9} Hz range with strains as large as 10^{-16}.
2. A stochastic background of supermassive black hole mergers related to large-scale structure formation.
3. A stochastic background associated with cosmic inflation and first-order phase transitions in the early Universe.

Detection of some events with very long wavelengths may be feasible with a *pulsar timing array (PTA).*

- In a PTA millisecond pulsars are viewed as precise clocks and a set of them is monitored for timing changes indicating the passage of a gravitational wave.
- Millisecond pulsars are favored because they are faster and more stable than average pulsars.
- These PTAs may be thought of as *natural interferometers with arms of galactic scale (kiloparsecs)*.

Pulsar timing arrays are sensitive to gravitational waves in the frequency range 10^{-9} to 10^{-6} Hz.

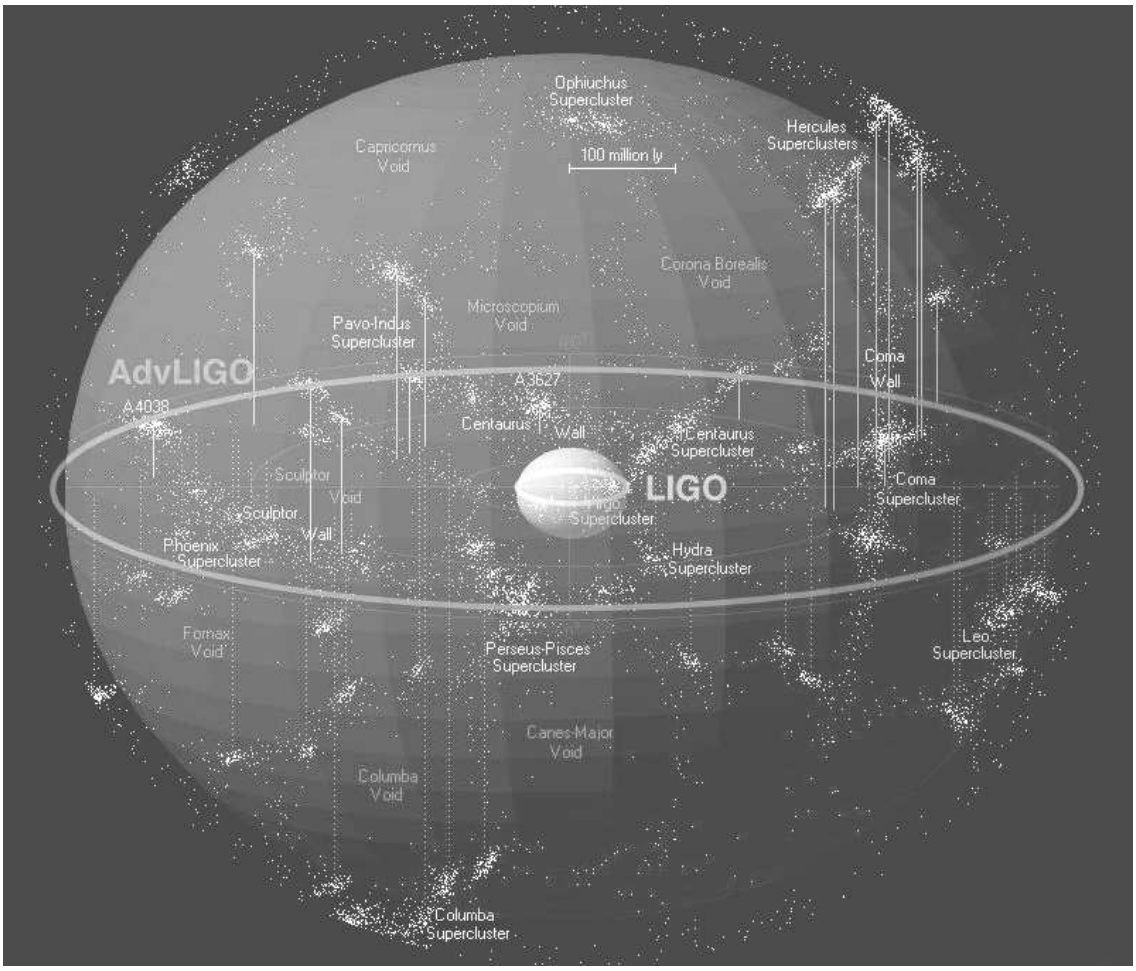


Figure 22.9: The expected reach of Advanced LIGO for detection of a neutron star merger event when at full design capability in 2020.

22.6.2 Reach of Advanced LIGO and Advanced VIRGO

Fig. 22.9 illustrates the reach expected for Advanced LIGO.

- The outer sphere has an approximate radius of 150 Mpc.
- The inner sphere indicates the corresponding reach of LIGO before the upgrade to Advanced LIGO.

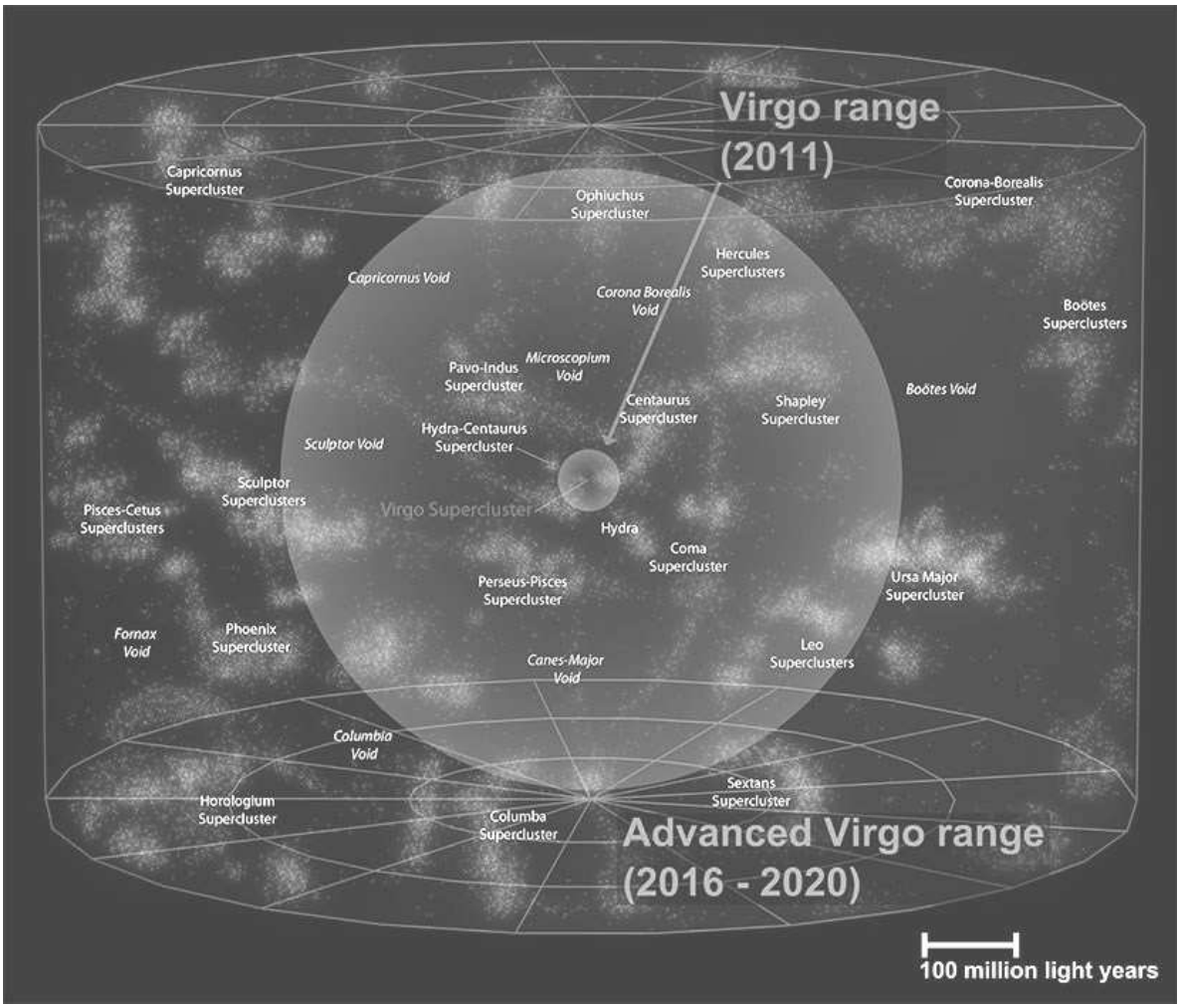


Figure 22.10: The expected reach of Advanced Virgo by 2020.

Figure 22.10 illustrates the projected reach of Advanced Virgo at full design capacity (2020).

Chapter 23

Weak Sources of Gravitational Waves

The preceding chapter introduced the basic idea of gravitational waves in terms of a linearized approximation to gravity.

- In this chapter and the next we consider potential sources of detectable gravitational waves.
- We begin with the simpler case of weak gravitational waves, described by *linearized gravity with weak sources*.

The next chapter will give a general introduction to the more difficult problem of describing strong sources of gravitational waves.

23.1 Production of Weak Gravitational Waves

To study the production of gravitational waves, we must in general solve the *full non-linear Einstein equations with source terms* $T_{\mu\nu}$.

- This is a formidable problem, generally only tractable for large-scale computation (*numerical relativity*).
- However, we can gain considerable insight by studying a less complex situation, the *linearized Einstein equation with sources*.
- This is an analytically accessible problem that has *many parallels with the study of sources for electromagnetic waves*.
- It has been shown by numerical simulation that many key features for the production of weak gravitational waves carry over in recognizable form for the production of gravitational waves in strong-gravity environments.
- Therefore, our approach in this chapter will be to concentrate on a more quantitative treatment of sources for weak gravitational waves

We shall then conclude in the following chapter with *qualitative order–of–magnitude remarks* and some *numerically computed examples* for strong-gravity wave sources.

23.1.1 Energy Densities

In an electromagnetic field or a Newtonian gravitational field it makes sense to talk about *local energy densities*.

- In a Newtonian field the energy density at a point $\boldsymbol{x}$ is

$$\varepsilon(\boldsymbol{x}) = -\frac{1}{8\pi G}(\nabla\Phi(\boldsymbol{x}))^2,$$

  where $\Phi(\boldsymbol{x})$ is the Newtonian gravitational potential.

- There is *no corresponding local energy density* in GR.
- Such a local density would *contradict the equivalence principle* (gravity *vanishes* in a local inertial frame).
- However, it makes sense to speak of an approximate energy density associated with a weak gravitational wave of wavelength λ, if λ *is much shorter than the curvature* R *of the background spacetime*.

Such approximations become very good at *large distances from the source* of a gravitational wave, where curvature associated with the source becomes negligible and $\lambda/R \to 0$.

Therefore, we may formulate a description of energy loss from gravitational wave sources at large distances from the source where we may associate an *approximate energy density* with a wave by averaging over several wavelengths.

23.1.2 Multipolarities

The lowest-order contribution from a source to electromagnetic radiation corresponds to *dipole motion of the source.*

- The gravitational field is a *tensor rather than vector field.*
- Like the electromagnetic field, the production of gravitational waves requires *non-spherical motion of the charge*, which is
  - *electrical charge* for the electromagnetic field and
  - *inertial mass* for the gravitational field.
- However, for the gravitational field *no monopole or dipole component contributes* to the generation of gravitational waves.
- The lowest order gravitational wave generation that is permitted corresponds to time-dependent *quadrupole distortions of the source mass.*

As a result, many of the formulas for sources of electromagnetic waves and for weak gravitational waves are *similar but not identical*.

23.1.3 Linearized Einstein Equation with Sources

Adding a source to the linearized Einstein equation in Lorentz gauge gives the wave equation

$$\Box \bar{h}_{\mu\nu} = -16\pi G T_{\mu\nu},$$

- where the stress–energy tensor $T_{\mu\nu}$ describing the source is assumed small, consistent with the linear approximation for the metric,
- $\Box$ is defined by

$$\Box \equiv \eta^{\alpha\beta}\partial_\alpha\partial_\beta = -\frac{\partial^2}{\partial t^2} + \boldsymbol{\nabla}^2$$

- and $\bar{h}_{\mu\nu}(t,\boldsymbol{x})$ is the *trace-reversed amplitude* given by

$$\bar{h}_{\mu\nu} \equiv h_{\mu\nu} - \tfrac{1}{2}\eta_{\mu\nu}h^\lambda_\lambda.$$

Solutions of the wave equation with sources can be found using

- Green's function methods similar to those used to solve wave equations in electromagnetism.
- We skip the details and jump to the results.

The metric perturbation for

- long-wavelength gravitational waves
- far from a nonrelativistic source

is found to be

$$\bar{h}^{ij}(t,\boldsymbol{x})_{r\to\infty} \simeq \frac{2}{r}\ddot{I}^{ij}(t-r),$$

- where double dots denote the *second time derivative* and
- The *second mass moment* $I^{ij}(t)$ is given by

$$I^{ij}(t) \equiv \int \rho(t,\boldsymbol{x}) x^i x^j d^3x,$$

where $\rho(t,\boldsymbol{x})$ is the *mass density of the source*.

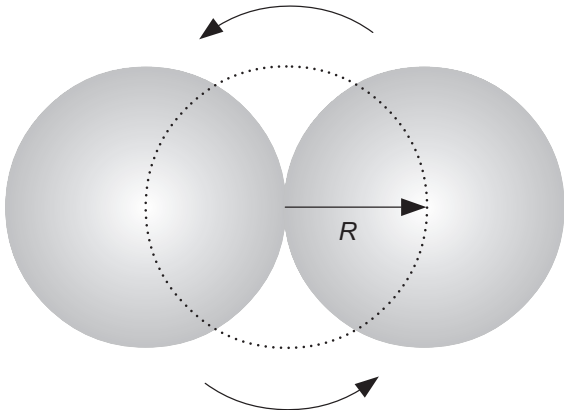


Figure 23.1: A contact binary as a source of gravitational waves.

23.1.4 Gravitational Wave Amplitudes

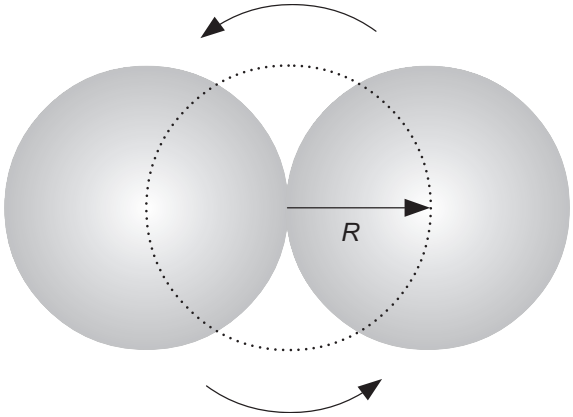
The gravitational wave amplitude $\bar{h}^{ij}$ is given by

$$\bar{h}^{ij}(t,\boldsymbol{x})_{r\to\infty} \simeq \frac{2}{r}\ddot{I}^{ij}(t-r),$$

in linear approximation.

- Let us make some estimates based on this formula using as a simple model for gravitational wave emission a binary star system.
- For simplicity, we shall assume the two stars
  - to be of the *same radius and mass*,
  - to revolve in *circular orbits* about their center of mass, and
  - to be in an orbit such that the surfaces of the two stars touch (*contact binary*).

Figure 23.1 illustrates.



The second mass moment is

$$I^{ij}(t) = \int \rho(t,\boldsymbol{x}) x^i x^j d^3x = MR^2 + MR^2 = 2MR^2.$$

The system revolves with a period P and taking the derivative twice with respect to time gives a factor $\omega^2 \sim 1/P^2$

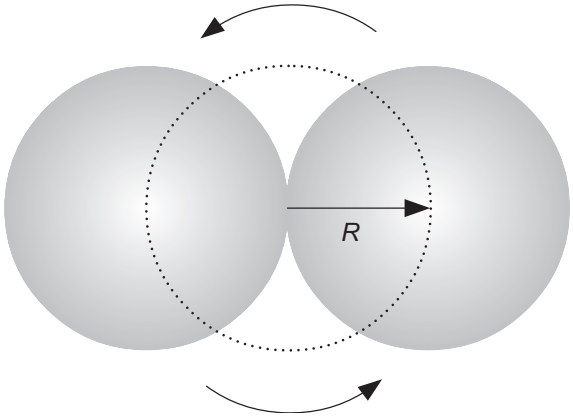
$$\ddot{I}^{ij} \simeq 2\frac{MR^2}{P^2}.$$

Insertion of this approximation in

$$\bar{h}^{ij}(t,\boldsymbol{x})_{r\to\infty} \simeq \frac{2}{r}\ddot{I}^{ij}(t-r),$$

leads to

$$\bar{h}^{ij} \simeq \frac{2}{r}\ddot{I}^{ij} = \frac{4MR^2}{rP^2}.$$



The relation

$$\bar{h}^{ij} \simeq \frac{2}{r}\ddot{I}^{ij} = \frac{4MR^2}{rP^2}.$$

can be expressed *in terms of source velocities* by noting that

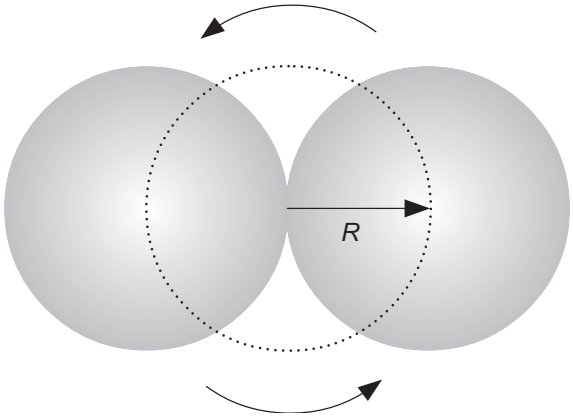
- In one period each star travels a distance $C = 2\pi R$.
- Therefore, the velocity of the center of mass for a star is

$$v = \frac{2\pi R}{P},$$

- This can be solved for the ratio R/P and used to rewrite $\bar{h}^{ij}$ as

$$\bar{h}^{ij} \simeq \frac{4MR^2}{rP^2} \quad \longrightarrow \quad \bar{h}^{ij} \simeq \frac{Mv^2}{\pi^2 r}.$$

This is a specialized result obtained assuming crude approximations, but more careful derivations indicate that it has broader validity than its derivation would suggest.



If we note that

- the *Schwarzschild radius and mass* are related by $r_S = 2M$,
- reinsert factors of c and G, and
- drop the numerical factors (note that the numerical factors dropped are larger than order unity),

the expression

$$\bar{h}^{ij} \simeq \frac{Mv^2}{\pi^2 r}.$$

may be rewritten as

$$\bar{h} \simeq \frac{r_S}{r}\frac{v^2}{c^2}.$$

The amplitude of the metric perturbation is

- largest for *compact sources* that have
- *radii comparable to* r_S.

As noted above, the result

$$\bar{h} \simeq \frac{r_S}{r}\frac{v^2}{c^2}.$$

is *more general than its derivation might suggest*.

We shall take it as a qualitative guide to gravitational wave amplitudes far from a weak source.

Weak gravitational waves are generated primarily by systems that are *gravitationally bound or nearly so*.

- Therefore *the virial theorem is applicable* and we may expect kinetic and potential energies to be comparable,

$$\frac{1}{2}Mv^2 \sim \frac{GM^2}{R},$$

- From this we may write

$$\frac{v^2}{c^2} \simeq \frac{2GM^2}{MRc^2} = \frac{r_S}{R} = \varepsilon^{2/7},$$

where we have defined a *gravitational wave emission efficiency factor*

$$\varepsilon \equiv \left(\frac{r_S}{R}\right)^{7/2}$$

(the justification for this terminology will be given below).

- Therefore, $\bar{h}$ may be expressed in the form

$$\begin{aligned}\bar{h} \simeq \frac{r_S}{r}\frac{v^2}{c^2} &\simeq \frac{r_S^2}{rR} = \varepsilon^{2/7}\frac{r_S}{r}, \\ &= 9.6\times 10^{-17}\varepsilon^{2/7}\left(\frac{M}{M_\odot}\right)\left(\frac{\text{kpc}}{r}\right),\end{aligned}$$

which is dimensionless.

23.1.5 Amplitudes and Event Rates

The preceding results may be used to estimate amplitudes and event rates for candidate gravitational wave events.

- Assume $\varepsilon = (r_S/R)^{7/2} \sim 0.1$ and that the mass participating in gravitational wave generation is $\sim 1 M_\odot$
- (These are reasonable guesses for merging neutron stars or asymmetric core collapse supernovae).
- For events in the galaxy, take $r \sim 10\,\mathrm{kpc}$.
- Then we may estimate

$$\begin{aligned} \bar{h} &= 9.6 \times 10^{-17} \varepsilon^{2/7} \left(\frac{M}{M_\odot}\right) \left(\frac{\mathrm{kpc}}{r}\right) \\ &\simeq 5 \times 10^{-18}. \end{aligned}$$

- But within our galaxy, on average,
  - supernova explosions occur only *once every 50 years*
  - neutron star mergers occur only *once every* 10^5 *or* 10^6 *years*.

Therefore, to obtain a reasonable event rate we must look for gravitational waves from *sources at much larger distances*.

The nearest rich cluster containing thousands of galaxies is Virgo, at a distance of about 16.5 Mpc.

- Therefore, if we go out to 100 Mpc or more, we may expect event rates for gravitational waves from merging neutron stars and asymmetric core-collapse supernovae to be much higher because
- we are now surveying many thousands or tens of thousands of galaxies.
- But the average observed metric perturbation (which is directly related to the strain detected by the detectors) becomes $\bar{h} \sim 10^{-21} - 10^{-22}$.

Thus, we expect that to detect systematic gravitational wave events the detectors must be able to sample strains reliably at the $\Delta L/L_0 \simeq 10^{-21}$ level or better.

23.1.6 Power in Gravitational Waves

The power radiated in gravitational waves for a system that

- has velocities well below c and
- weak internal gravity

is given by the *quadrupole formula,*

$$L = \frac{dE}{dt} = \frac{1}{5}\left\langle \dddot{\mathbf{I}}_{ij}\dddot{\mathbf{I}}^{ij}\right\rangle = \frac{1}{5}\frac{G}{c^5}\left\langle \dddot{\mathbf{I}}_{ij}\dddot{\mathbf{I}}^{ij}\right\rangle,$$

- $\langle\ \rangle$ denotes a *time average over a period*,
- the triple dot means a *third time derivative*, and
- the *reduced quadrupole tensor* $\mathbf{I}$ is defined by

$$\mathbf{I}^{ij} \equiv I^{ij} - \tfrac{1}{3}\delta^{ij}\mathrm{Tr}\, I,$$

where $\mathrm{Tr}\, I \equiv I^k_k$.

This formula is the gravitational analog for *radiated power in electromagnetism* but it has

- a different factor (1/5 instead of 1/20) and
- corresponds to *quadrupole radiation* rather than *dipole radiation*.

As shown in a Problem,

$$L = \frac{dE}{dt} = \frac{1}{5}\frac{G}{c^5}\left\langle \dddot{I}_{ij}\dddot{I}^{ij} \right\rangle,$$

may be reduced to

$$L \simeq L_0 \frac{r_S^2}{R^2}\left(\frac{v}{c}\right)^6,$$

where the *characteristic scale for radiated gravitational wave power* is set by

$$L_0 \equiv \frac{c^5}{G} = 3.6 \times 10^{59}\ \text{erg s}^{-1},$$

and the total energy ΔE emitted in one period P can then be calculated as

$$\Delta E \simeq LP \simeq Mc^2\left(\frac{r_S}{R}\right)^{7/2} = \varepsilon Mc^2.$$

Therefore, we conclude that

$$\varepsilon = \left(\frac{r_S}{R}\right)^{7/2}$$

parameterizes the *efficiency of converting mass to energy by gravitational wave emission*.

23.2 Gravitational Radiation from Binary Systems

- In preceding sections we have made some qualitative estimates for gravitational wave emission from binary stars.
- In this section we derive in a somewhat more rigorous fashion a formalism applicable for such systems.

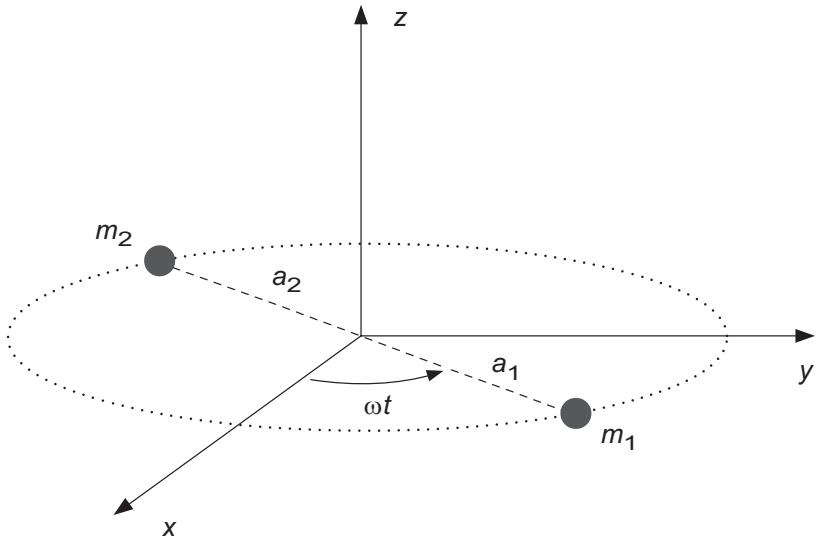


Figure 23.2: Coordinate system for binary stars in circular orbits.

23.2.1 Gravitational Wave Luminosity

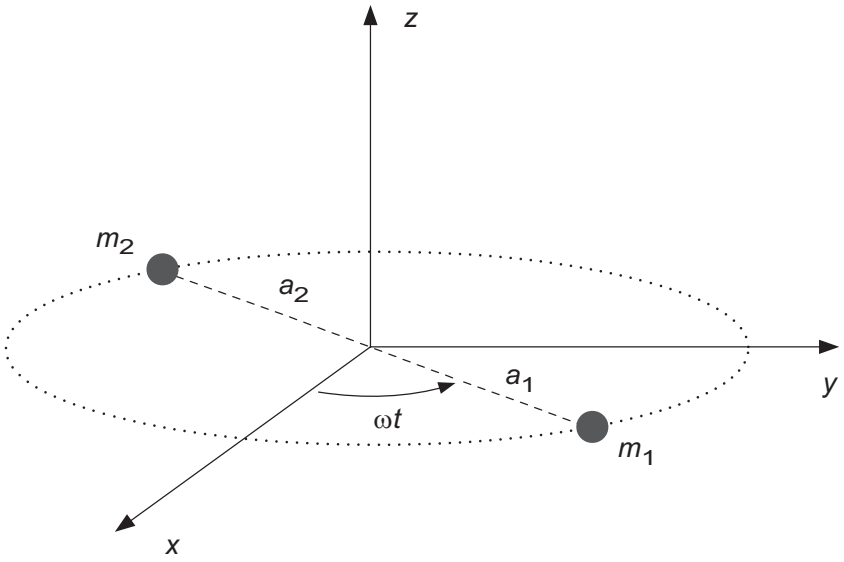
A binary system with circular orbits is illustrated in Fig. 23.2

- a_1 and a_2 are the distances of the masses m_1 and m_2, respectively, from the center of mass, and
- $\varphi = \omega t$ is the azimuthal angle between the line joining the stars and the x axis.

The *components of the second mass moment* reduce to a sum,

$$I^{ij}(t) \equiv \int \rho(t,\boldsymbol{x}) x^i x^j d^3x \quad \longrightarrow \quad I^{ij} = m_1 x^i x^j + m_2 x^i x^j$$

for two discrete masses (superscripts label coordinates and subscripts label objects).



Introduce *polar coordinates for each star* i,

$$x_i(t) = a_i \cos \omega t \qquad y_i(t) = a_i \sin \omega t \qquad z_i(t) = 0,$$

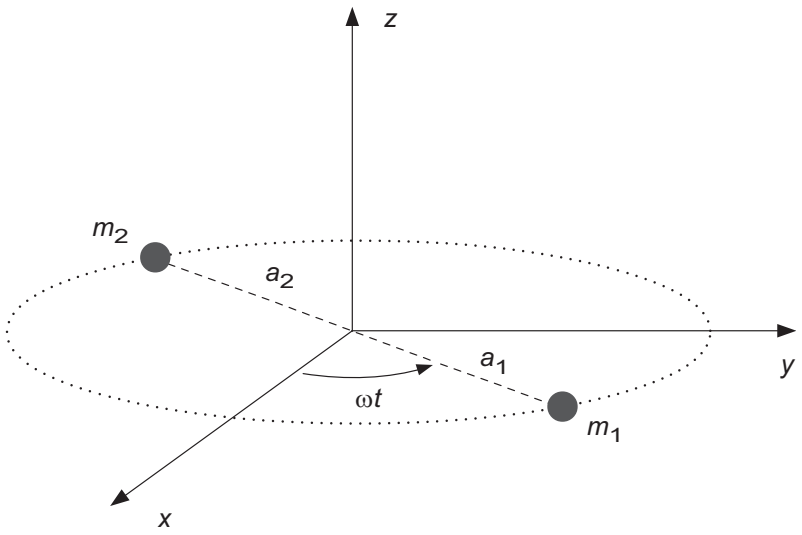
and *evaluate the mass moments*. For example,

$$\begin{aligned} I^{xx} = I^{11} &= m_1 a_1^2 \cos^2 \omega t + m_2 a_2^2 \cos^2 \omega t \\ &= (m_1 a_1^2 + m_2 a_2^2) \cos^2 \omega t \\ &= \mu a^2 \cos^2 \omega t \\ &= \tfrac{1}{2} \mu a^2 (1 + \cos 2\omega t), \end{aligned}$$

- where $a \equiv a_1 + a_2$,
- the *reduced mass* is $\mu = m_1 m_2/(m_1 + m_2)$, and
- the identity

$$m_1 a_1 = m_2 a_2 = \mu a$$

was used.



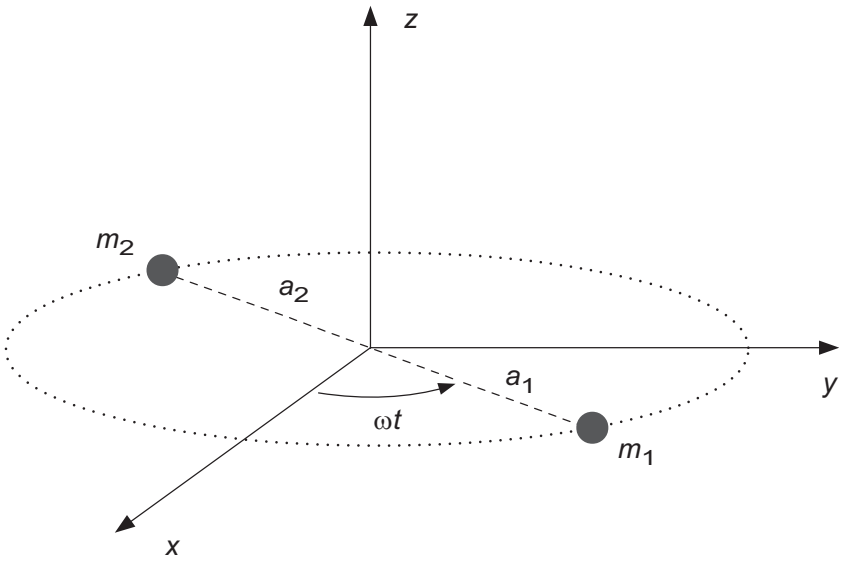
Carrying out a similar steps for all components gives

$$I^{xx} = I^{11} = \tfrac{1}{2}\mu a^2(1+\cos 2\omega t),$$
$$I^{xy} = I^{yx} = I^{12} = \tfrac{1}{2}\mu a^2 \sin 2\omega t,$$
$$I^{yy} = I^{22} = \tfrac{1}{2}\mu a^2(1-\cos 2\omega t).$$

Computing second time derivatives of I^{ij} and inserting in the amplitude equation gives

$$\bar{h}^{ij} = \frac{4\omega^2\mu a^2}{r}\begin{pmatrix} -\cos 2\omega(t-r) & -\sin 2\omega(t-r) & 0 \\ -\sin 2\omega(t-r) & \cos 2\omega(t-r) & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

- The appearance of 2ω in the arguments means that the frequency f of the emitted gravitational wave radiation will be *twice the orbital frequency* ω.
- This is because the mass distribution varies with a period equal to half the rotational period.



From above,

$$\mathrm{Tr}\, I = I^{xx} + I^{yy} = \mu a^2,$$

and $\dot{I}^{ij} = \dot{I}^{ij}$, so that the radiated luminosity is given by

$$L = \frac{1}{5}\left\langle \dddot{I}_{ij}\dddot{I}^{ij}\right\rangle = \frac{1}{5}\left\langle \dddot{I}_{ij}\dddot{I}^{ij}\right\rangle .$$

From the second time derivatives found above, the triple time derivatives are

$$\dddot{I}^{xx}(t) = 4\omega^3\mu a^2 \sin 2\omega t \qquad \dddot{I}^{yy}(t) = -4\omega^3\mu a^2 \sin 2\omega t$$
$$\dddot{I}^{xy}(t) = \dddot{I}^{yx}(t) = -4\omega^3\mu a^2 \cos 2\omega t \quad ,$$

Thus the radiated gravitational wave power is

$$\begin{aligned} L &= \frac{1}{5}\left\langle (\dddot{I}^{xx})^2 + 2(\dddot{I}^{xy})^2 + (\dddot{I}^{yy})^2 \right\rangle \\ &= \frac{1}{5}\left[\frac{1}{P}\int_0^P \left((\dddot{I}^{xx})^2 + 2(\dddot{I}^{xy})^2 + (\dddot{I}^{yy})^2\right) dt\right] \\ &= \frac{16\omega^6\mu^2 a^4}{5P}\int_0^P \left(\sin^2 2\omega t + 2\cos^2 2\omega t + \sin^2 2\omega t\right) dt \\ &= \frac{32}{5}\omega^6\mu^2 a^4 \\ &= \frac{32}{5}\frac{G^4}{c^5}\frac{M^3\mu^2}{a^5}, \end{aligned}$$

where $M \equiv m_1 + m_2$ and

- the second step averages over one period,
- the integral in line three evaluates to $2P$,
- in the last step Kepler's 3rd law was used to eliminate ω,
- and factors of G and c have been restored.

Often it is convenient to express the power in terms of the period P rather than the separation a. Using *Kepler's 3rd law* again,

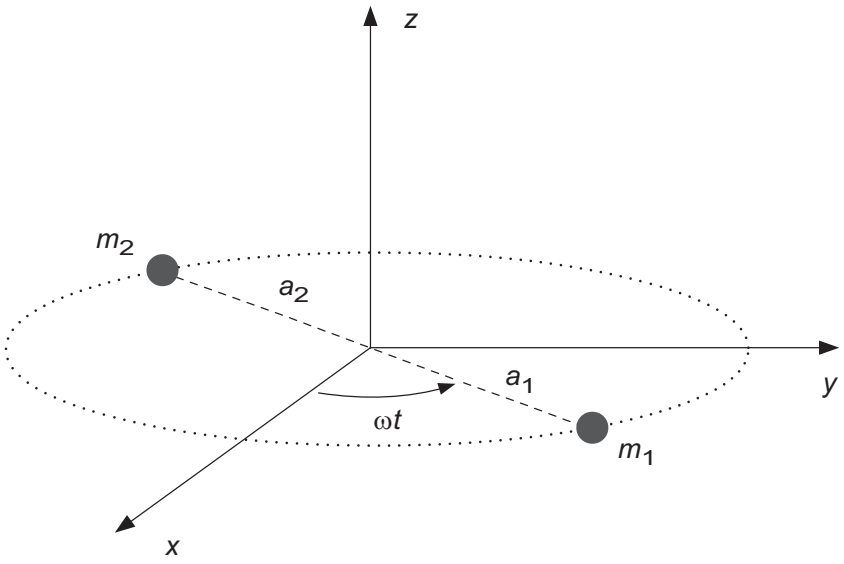
$$L = \frac{128}{5}4^{2/3}\frac{G^{7/3}}{c^5}M^{4/3}\mu^2\left(\frac{\pi}{P}\right)^{10/3}.$$

The preceding expressions have assumed *circular orbits*.

- The corresponding luminosity for *eccentric binary orbits* can be approximated as

$$L = f(e)\bar{L} \qquad f(e) = \frac{1 + \frac{73}{24}e^2 + \frac{37}{96}e^4}{(1-e^2)^{7/2}},$$

- where the orbital *eccentricity is* e and
- the *luminosity assuming circular orbits* is $\bar{L}$.



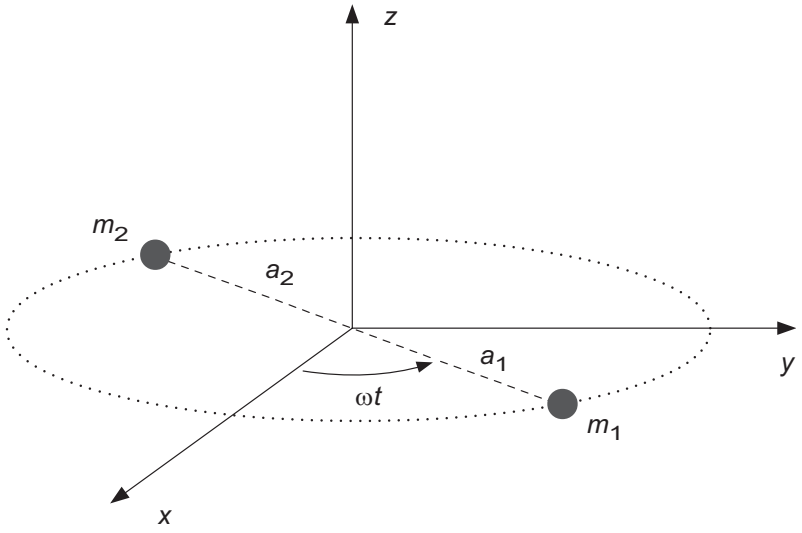
Example: Assume that in the binary above $m_1 = m_2 = 1\,M_\odot$, and that the period is 6 hours. Restoring factors of G and c,

$$L = \frac{128}{5} 4^{2/3} \frac{G^{7/3}}{c^5} M^{4/3} \mu^2 \left(\frac{\pi}{P}\right)^{10/3}$$
$$= 2.3 \times 10^{45} \left(\frac{M}{M_\odot}\right)^{4/3} \left(\frac{\mu}{M_\odot}\right)^2 \left(\frac{1\,\mathrm{s}}{P}\right)^{10/3} \mathrm{erg\ s^{-1}}.$$

Inserting the masses and period gives a luminosity of

$$L = 5.2 \times 10^{30}\,\mathrm{erg\ s^{-1}},$$

if circular orbits are assumed.



23.2.2 Gravitational Radiation and Binary Orbits

The *total energy is negative* (it is a bound system), so if the energy is reduced by gravitational wave emission

- the period P (and size) of a binary orbit must decrease.
- The emission of gravitational wave radiation also tends to *circularize elliptical orbits*, so assume orbits to be circular.

The *total orbital energy* of the system is given by

$$E = -\frac{Gm_1m_2}{2a} = -\frac{G\mu M}{2a},$$

where

- $M = m_1 + m_2$ is the total mass,
- $\mu = m_1m_2/M$ is the reduced mass, and
- $a = a_1 + a_2$ is the average separation of the binary pair.

Utilizing the preceding result and Kepler's 3rd law for the period P

$$E = -\frac{Gm_1m_2}{2a} = -\frac{G\mu M}{2a} \qquad P^2 = \frac{4\pi^2}{GM}a^3,$$

and assuming that $L = -dE/dt$ where

- L is the *gravitational wave luminosity* and
- dE/dt is the *rate of energy loss* from the orbital motion,

the *rate of change for the period* is given by

$$\frac{dP}{dt} = -\frac{96}{5}\frac{G^3}{c^5}\frac{M^2\mu}{a^4}P.$$

Utilizing Kepler's 3rd law to eliminate a in favor of P,

$$\frac{dP}{dt} = -\frac{192\pi}{5}\frac{G^{5/3}}{c^5}\frac{m_1m_2}{M^{1/3}}\left(\frac{2\pi}{P}\right)^{5/3}.$$

This is a dimensionless measure of how the binary orbit is altered over time by emission of gravitational waves.

The preceding formulas for dP/dt assumed *circular orbits*. If the binary has an *eccentric orbit* we may approximate

$$\frac{dP}{dt} = f(e)\left(\frac{dP}{dt}\right)_0,$$

with the definitions

$$\left(\frac{dP}{dt}\right)_0 \equiv -\frac{192\pi}{5}\frac{G^{5/3}}{c^5}\frac{m_1 m_2}{M^{1/3}}\left(\frac{2\pi}{P}\right)^{5/3}$$

$$f(e) = \frac{1+\frac{73}{24}e^2+\frac{37}{96}e^4}{(1-e^2)^{7/2}}.$$

for an *orbital eccentricity* e.

23.2.3 Gravitational Waves from the Binary Pulsar

If the orbit for the *Binary Pulsar* were circular, then after evaluating constants in the preceding equations

$$\left(\frac{dP}{dt}\right)_0 = -3.66\times10^{-6}\frac{(m_1/M_\odot)(m_2/M_\odot)}{(M/M_\odot)^{1/3}}\left(\frac{1\,\mathrm{s}}{P}\right)^{5/3}$$
$$= -2\times10^{-13},$$

- where $M \equiv m_1 + m_2$,
- the observed masses of $m_1 \sim m_2 \sim 1.4M_\odot$, and
- the orbital period of 7.75 hours were used.

But the Binary Pulsar orbit is *eccentric with* $e = 0.617$, so

$$\frac{dP}{dt} = f(e)\left(\frac{dP}{dt}\right)_0$$
$$= (11.84)\times(-2\times10^{-13})$$
$$= -2.37\times10^{-12},$$

where an eccentricity factor $f(e) = 11.84$ was computed.

- Assuming this rate to be constant, the total change in period over one year is $\Delta P \simeq -7.5\times10^{-5}$ s.
- This decrease is *tiny but easy to measure* because of the *precise pulsar clock*.

From this rate of orbital decay the *inspiral time to merger* is about 3×10^8 years.

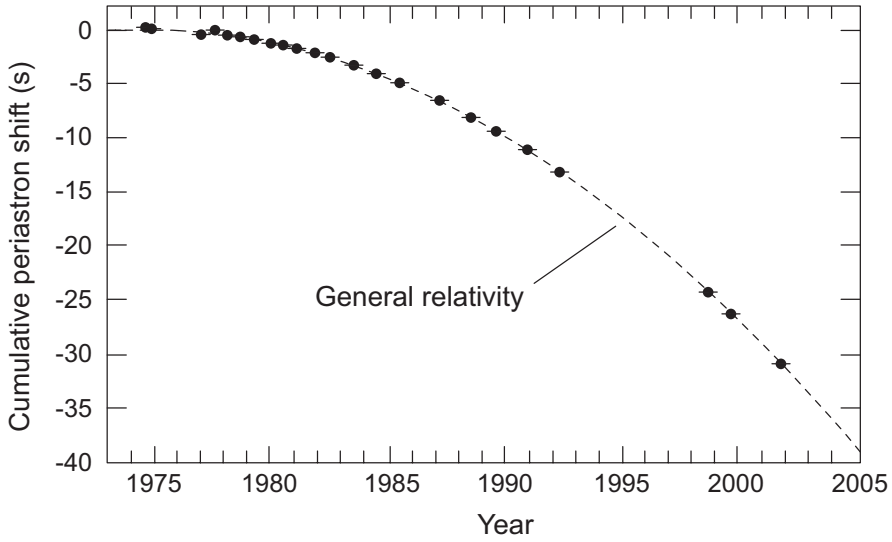


Figure 23.3: Observed periastron shift of the Binary Pulsar orbit because of gravitational wave emission. Data points are the measurements of Taylor and Weisberg and the dashed curve is the prediction from general relativity.

Fig. 23.3 illustrates for the Binary Pulsar

- the cumulative shift of the *periastron time* (time of closest approach) as a function of elapsed time,
- compared with a calculation assuming emission of gravitational waves to be responsible for this shift.
- The *quality of the data* (note error bars), and the *agreement with general relativity* are remarkable.

The Binary Pulsar provides *strong indirect evidence for the existence of gravitational waves*.

Chapter 24

Strong Sources of Gravitational Waves

Our discussion so far has been in terms of linearized gravity

- However, for gravitational wave sources such as
  - mergers involving some combination of neutron stars and black holes, or
  - the core collapse of a massive star

  strong curvature invalidates linear approximations.
- In strong gravity, reliable calculations are difficult and can only be produced through numerical simulations.
- Nevertheless, those simulations show that many features from the linear regime survive in strong gravity.

In particular, calculations suggest that many basic features of strong gravitational wave sources can be obtained by *dimensional analysis*.

24.1 A Survey of Candidate Sources

In this section we

- *summarize some likely strong sources* of gravitational waves and
- give some *results of numerical simulations*.

These fall into two categories:

- *Merger of binary compact objects* when their orbits decay because of gravitational wave emission.
  - Binary black hole mergers
  - Binary neutron star mergers
  - Merger of a black hole and a neutron star.
- *Core collapse* in massive stars.

Events in both categories may exhibit *rapid asymmetric motion of large compact masses*.

- This the essential condition for production of strong gravitational waves.
- The asymmetric motion of the mass in these events is expected to differ in characteristic ways.
- Hence the detailed gravitational wave pattern should carry information about the type of event that produced it.

24.1.1 Merger of a Neutron Star Binary

Gravitational waves from *merging neutron stars in relatively nearby galaxies* should be observable.

- The possibility of neutron stars merging might remote.
- However, once a neutron star binary is formed its
  - orbital motion radiates energy as gravitational waves,
  - the orbits must shrink by energy conservation,

  and *inevitably the two neutron stars must merge*.
- The most relevant question is the *timescale for merger*.
- One typically speaks of *merger within a Hubble time* [$\sim$ 14 billion years], meaning that
- the orbital separation decays fast enough for the merger to occur on a *timescale less than the age of the Universe*.
- Presently several observed binary systems are predicted to have a merger timescale less than the Hubble time.
- For example, the orbit of the Binary Pulsar
  - has shrunk to a separation at nearest approach of about 750,000 km today,
  - with a decay rate that should lead to merger in about 3×10^8 years.

Formation of the neutron star binary (more generally any binary involving neutron stars and/or black holes) *is not easy, however*.

- *Either* a binary, or multiple-star system
  - must form with two stars *massive enough to undergo core collapse* and produce two compact objects,
  - and the compact objects must *remain bound to each other* through both core collapse events,
- *or* the binary results from *gravitational capture* in a dense star cluster through
  - *tidal dissipation of kinetic energy* in a binary encounter, or
  - *a 3-body encounter* in which one object carries off enough energy to leave the other two bound

These are *improbable scenarios*.

- However, calculations and direct observation of compact-object binaries indicate that they are *not impossible*.
- Some estimates indicate that there is about one binary neutron star merger each day in the observable Universe.

> The probability of neutron star merger in any given region of space is very small, but *the Universe is a very big place*.

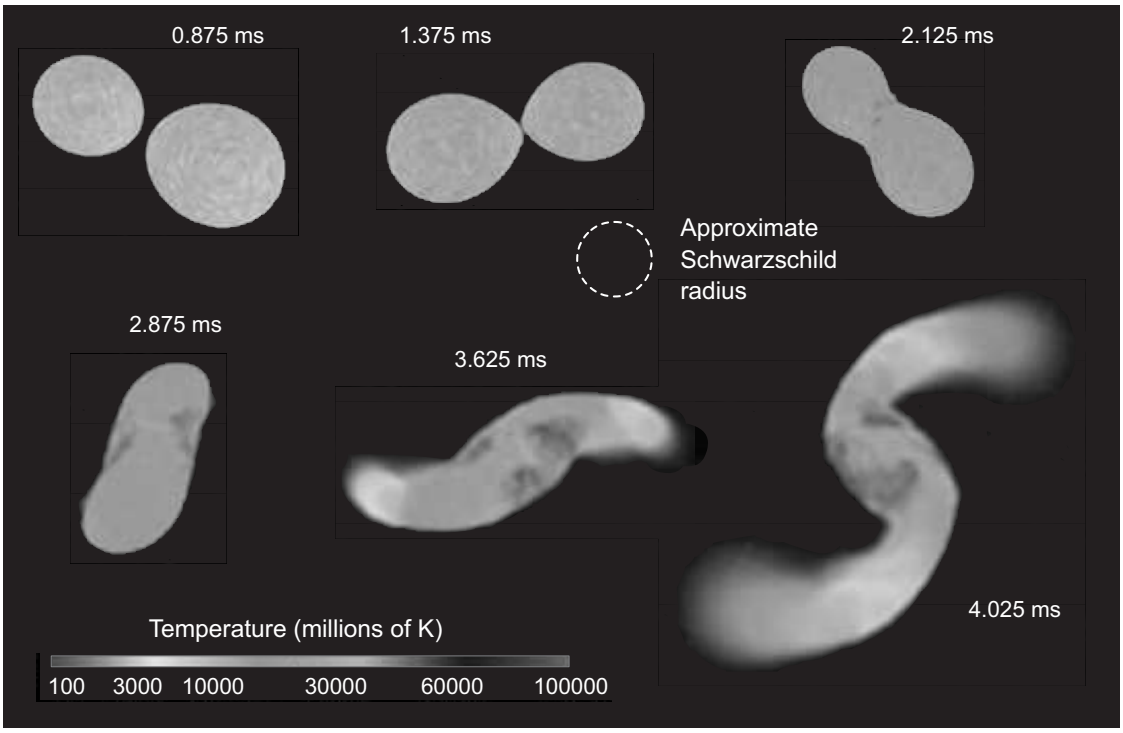
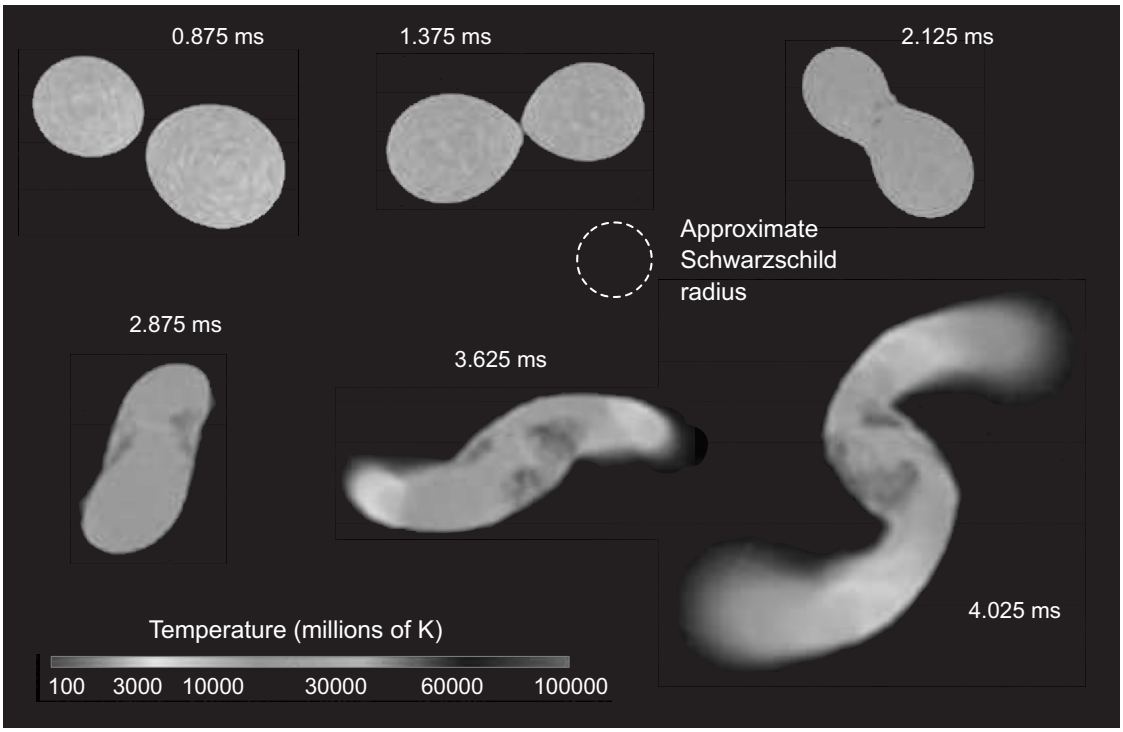


Figure 24.1: Simulation of the merger of two neutron stars. The elapsed time is about 3 ms and the approximate Schwarzschild radius for the combined system is indicated. The rapid motion of several solar masses of material with large quadrupole distortion and sufficient density to be compressed near the Schwarzschild radius indicates that this merger should be a strong source of gravitational waves (Source: S. Rosswog simulation).

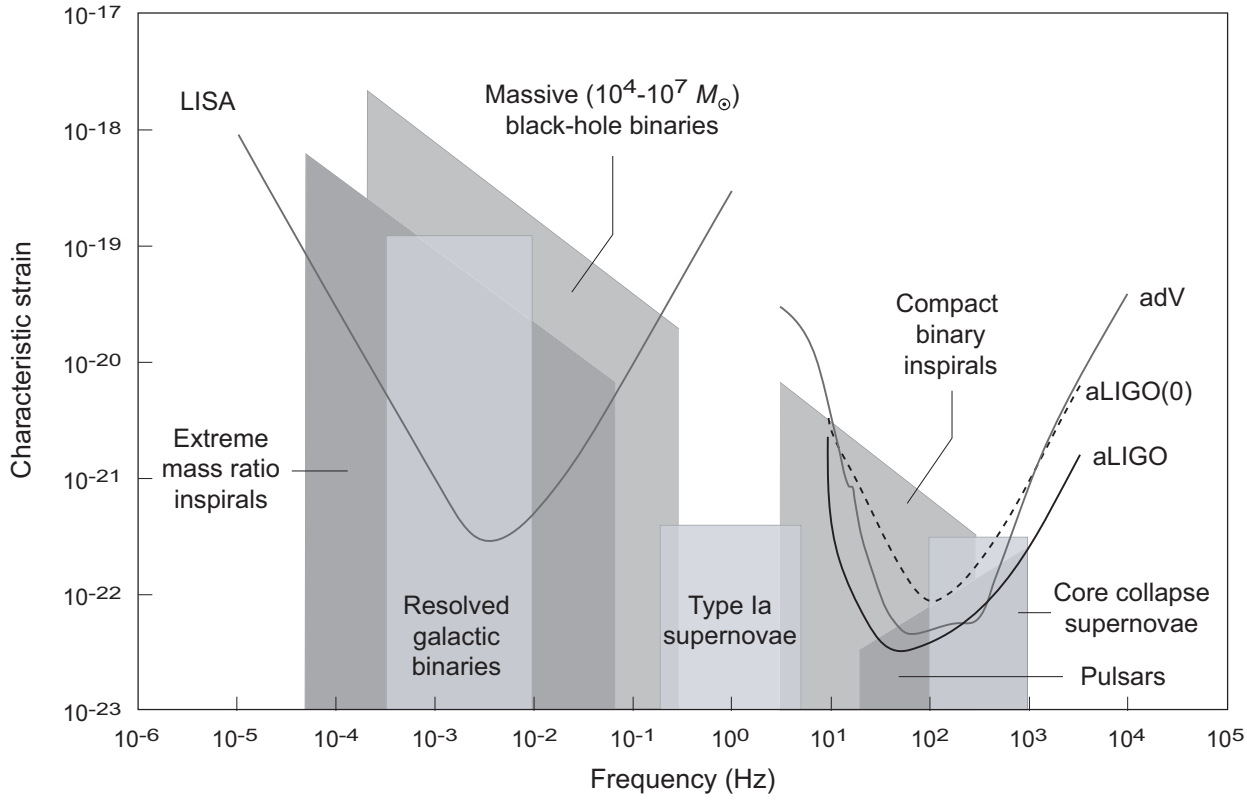
Figure 24.1 shows a numerical simulation of a merger.

- The orbit of the binary has decayed steadily as a result of gravitational wave emission, causing the stars to
- spiral together at a rapidly-increasing rate near the end.
- The sequence of images in Fig. 24.1 illustrates the merger over a period of milliseconds near when the surfaces of the neutron stars first touch.



- In this simulation,
  - The *large asymmetric mass distortion*,
  - the *high velocities* generated by
  - *revolution on millisecond timescales*, and
  - the *highly-compact mass distribution*,

  imply that mergers of neutron stars will be a *strong source of gravitational waves*.



- As the figure above illustrates,
  - frequencies for gravitational waves emitted from neutron star mergers (labeled *compact binary inspirals*)
  - are expected to be accessible to Earth-based interferometers like LIGO or Virgo.

Later we shall discuss the first observation of gravitational waves from neutron star merger.

The outcome of such neutron star mergers will depend on the total mass of the two neutron stars.

- Presently, our (relatively poor) understanding of neutron star equations of state suggests an upper limit $\sim 2-3\ M_\odot$ for the mass of a rapidly-spinning neutron star.
- If the total merger mass is less than this the merger will likely lead to a rapidly-spinning (because of the large initial orbital angular momentum of the binary) neutron star.
- If the total merger mass is more than $\sim 2-3\ M_\odot$, the likely outcome will be a *Kerr black hole with a large spin*.
- Such a merger of two neutron stars is also a prime candidate for producing a *short-period gamma ray burst*,
- which raises the possibility of *detecting a gamma-ray burst and gravitational waves in coincidence* from the same event.

Later we shall describe observation of the first such *multimessenger event*.

24.1.2 Stellar Black Hole Mergers

Merging stellar black holes produce strong gravitational waves.

- Such mergers have many similarities with merging neutron stars, and both are likely to leave behind
- a *rapidly spinning Kerr black hole*.
- There is *one essential difference*, however.
- Both Schwarzschild and Kerr black holes correspond to *vacuum solutions* of the Einstein equations.
- Thus, there is *only gravity—curved spacetime—and no matter* involved (neglecting accretion from other objects).
- All mass is concentrated at the black hole singularities, which are *hidden by the event horizons*.

Thus, *black hole mergers are simpler than neutron star mergers*, which involve matter interactions.

- The initial configuration is characterized only by
  - *spins and masses of the black holes*, and
  - the *binary orbital parameters*,

  as a consequence of the *no hair theorem*.
- The final configuration is even simpler: the *single Kerr black hole* is characterized *only by its mass and spin*.

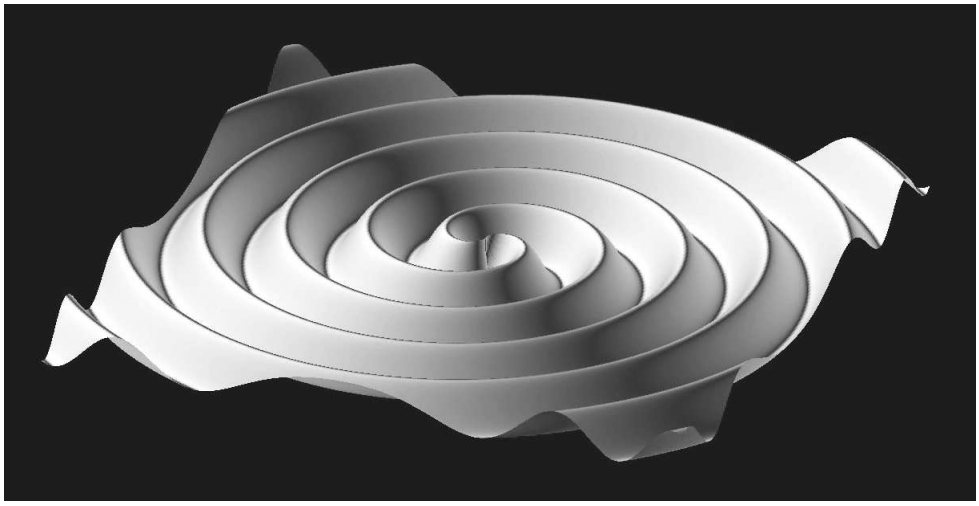


Figure 24.2: Snapshot of gravitational wave emission from inspiral of a binary black hole system.

Figure 24.2 illustrates *gravitational wave emission from an inspiraling black hole pair* based on numerical simulations of general relativity.

- The black holes revolving around their center of mass are at the center.
- The spirals trailing from them are the outwardly propagating gravitational waves
- that are being generated by the revolving binary system.

Later in this chapter, observational evidence for gravitational waves from the merger of stellar black holes will be discussed.

24.1.3 Merger of a Black Hole and a Neutron Star

It is possible to have a binary system in which one object is a neutron star and one a stellar black hole.

- As for other compact binaries, the *orbit will decay by gravitational wave emission*,
- leading eventually to a merger.
- Equations of state and observations suggest an upper limit of around $2M_\odot$ for neutron stars,
- but there is in principle no limit on the mass of the black hole.
- The black hole component of the binary already exceeds the upper mass limit for a neutron star.
- Thus it is likely that the outcome of a merger between a black hole and a neutron star is a *rapidly-spinning Kerr black hole*.

As for the neutron star merger, the dense matter and associated equation of state for the neutron star likely makes this merger more complex than that of two black holes.

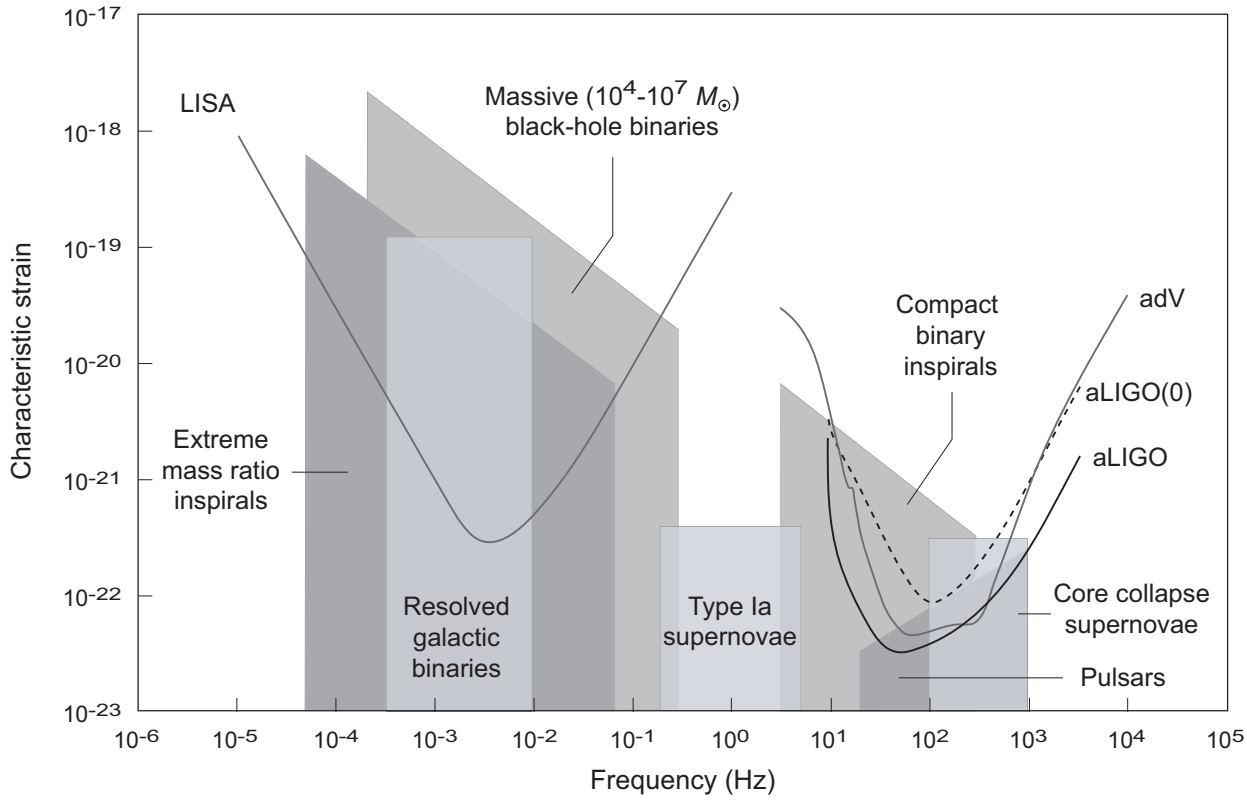
24.1.4 Core Collapse in Massive Stars

In the *core collapse of a massive star*

- the collapsing core has densities comparable to that of a neutron star and
- in the collapse
  - a solar mass or more of material may be set in motion
  - with velocities of order 10% of light velocity.
- If such a collapse were to proceed with spherical symmetry, no gravitational waves would be produced.
- However, numerical simulations of core collapse generally find
  - *large asymmetries* generated by
  - large-scale *supersonic convection*

  (which is in turn driven by *large entropy or concentration gradients* in the core).
- These asymmetries will produce *quadrupole distortions of mass* that *vary rapidly*.

Thus collapsing stellar cores are potential sources of detectable gravitational waves.



- As the figure above illustrates, the frequencies for gravitational waves emitted from core collapse events are expected to be accessible to Earth-based interferometers like LIGO or Virgo.

24.1.5 Merging Supermassive Black Holes

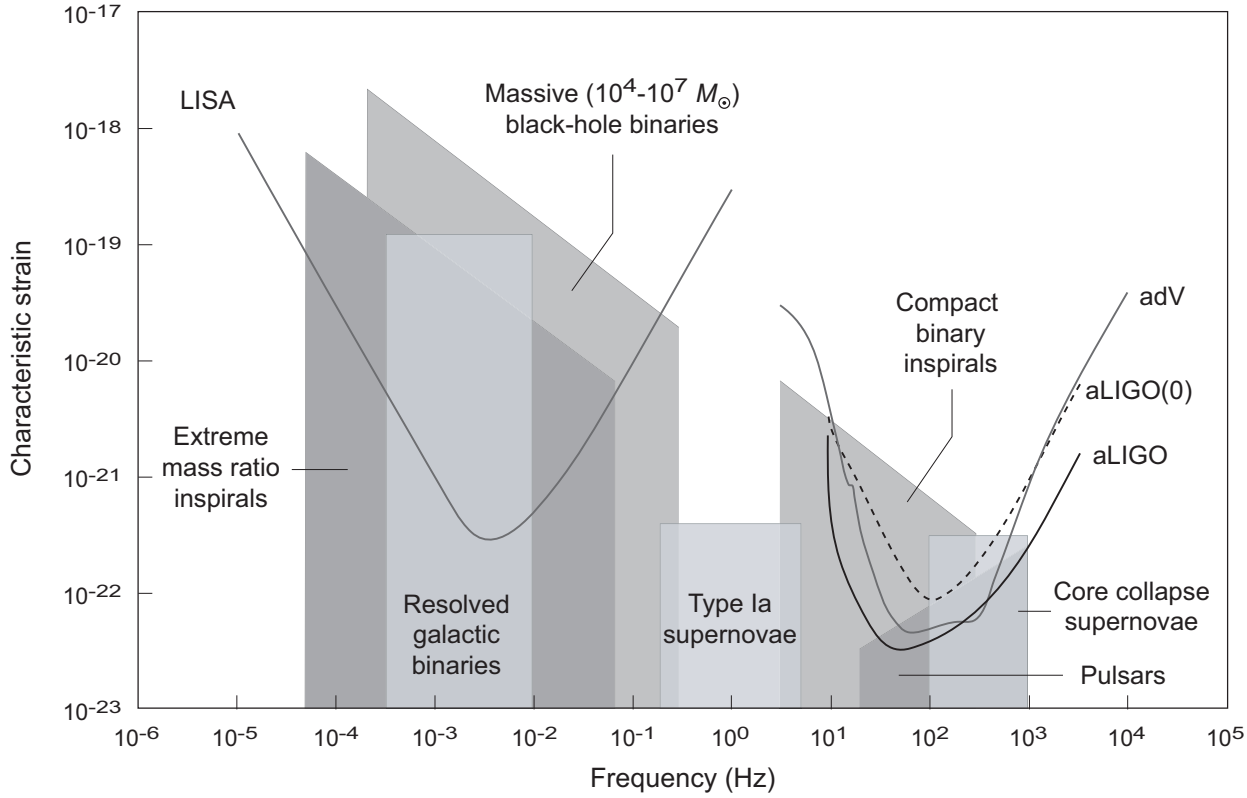
The cores of many—perhaps all—massive galaxies host *black holes containing* $10^6 M_\odot - 10^9 M_\odot$. For example,

- The fully-resolved orbits of individual stars at the center of the Milky Way indicate a $4 \times 10^6 M_\odot$ *black hole* and
- Analysis of velocity fields near the centers of giant elliptical galaxies often indicate that
  - billions of solar masses of non-luminous material
  - is packed into a region comparable in size to the Solar System.

  The simplest conclusion is that these are *supermassive black holes*.

There also is very strong observational evidence that *galaxy mergers have been common* in the past history of the Universe.

Therefore, it may be expected that the *merger of supermassive black holes* can occur as a consequence of *galaxy collisions*.



From the figure above,

- The frequency of gravitational waves from massive and supermassive black holes is *too low for observatories like LIGO or Virgo*.
- However the proposed *space-based LISA array* would be sensitive to such events with black hole masses less than about $10^7 M_\odot$.

Pulsar timing arrays may be able to detect gravitational waves from the merger of even more massive black holes.

Quantitative investigation of supermassive black hole mergers requires *large-scale computer simulations*.

- However dimensional analysis arguments
  - based on the *quadrupole power formula* and
  - supported by *numerical simulation*

  suggest that *peak luminosities are set by the scale* c^5/G,

$$L \simeq L_0 \frac{r_S^2}{R^2} \left(\frac{v}{c}\right)^6 \qquad L_0 \equiv \frac{c^5}{G} = 3.6 \times 10^{59}\,\mathrm{erg\ s^{-1}}$$

- For merger of supermassive black holes

$$L \propto L_0 \simeq \eta \frac{c^5}{G} \simeq \eta \times 10^{59}\,\mathrm{erg\ s^{-1}},$$

  with the factor η accounting for details of the merger and being of order 1% in typical cases.

- If such events occur, their luminosities would *surpass that of any other event in the Universe* for a period of days.

- To set some perspective, if $\eta \sim 0.01$ this luminosity is about
  - 10^6 times larger than *supernova or gamma ray burst photon luminosities*,
  - 10^{10} times larger than *quasar luminosities*, and
  - 10^{23} times larger than *solar luminosity*.

24.1.6 Sample Gravitational Waveforms

Some computed gravitational waveforms for events of the type described above are displayed in Figs. 24.3 (next slide).

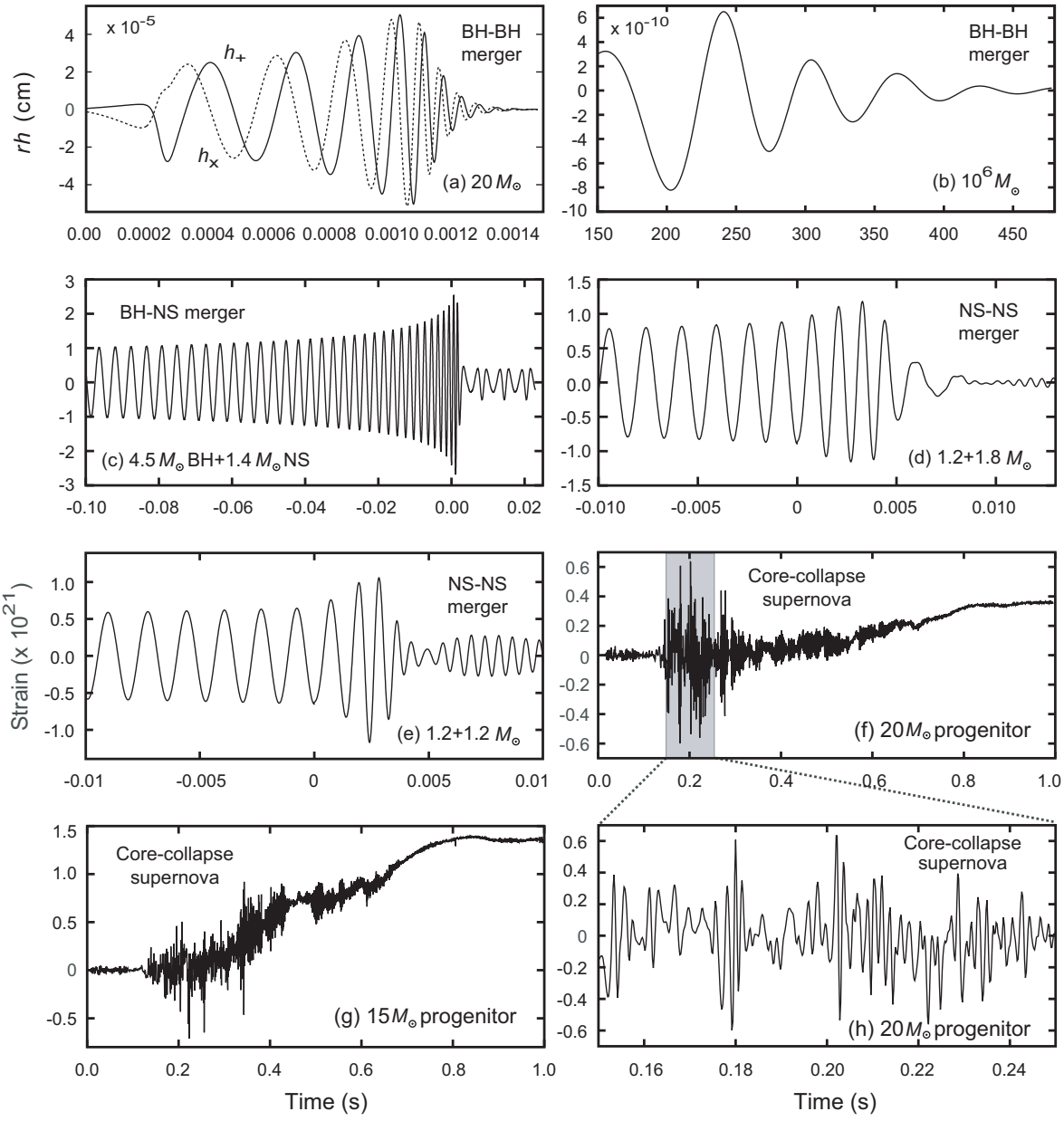


Figure 24.3: Some computed gravitational waveforms. All are h_+ polarization except for (a), where both h_+ and $h_\times$ are shown. In the bottom six panels strain is given in dimensionless units of 10^{-21} by assuming a distance to the source. In the top two panels rh is shown, where r is the distance to the source in cm. Further details may be found in the references for each case. (a)–(b) Equal-mass black hole mergers (BH–BH). (c) Black hole and neutron star merger (BH–NS) at 15 Mpc. (d)–(e) Neutron star mergers (NS–NS) at 15 Mpc. Case (d) is a $1.2\,M_\odot + 1.8\,M_\odot$ merger (all masses baryonic). Case (e) is a $1.2\,M_\odot + 1.2\,M_\odot$ merger. (f)–(h) Supernova at 15 kpc for two progenitor masses; time measured from bounce. Panel (h) displays the initial burst of panel (f) at higher resolution.

Different classes of events have *characteristic waveforms* and within classes the *waveforms are often sensitive to details*. For example,

- *Mergers* are characterized by the *chirp waveform,*
  - corresponding to amplitude and frequency rising rapidly near merger,
  - followed by low-amplitude ringdown,

  with details depending on the type of merger.
- *Supernova explosions* are characterized by a much more complex wave pattern reflecting detailed microphysics that varies with characteristics of the progenitor star.
- Hence, it may be expected that *waveform templates* like those displayed in the preceding figure may be used to *identify classes of gravitational wave events*.

With data from a sufficient number of events, the detailed waveforms might shed new light on the physics underlying each class of events.

24.2 Multimessenger Astronomy

Gravitational waves from neutron star mergers

- might determine the neutron star equation of state better than our present understanding, and
- gravitational waves from core collapse supernovae might constrain the supernova mechanism more rigorously than from present data.
- This would be particularly true if such events were observed in more than one way, such as
  - detection of both gravitational waves and neutrinos emitted from a core collapse supernova, or
  - detection of gravitational waves from a neutron star merger also observed as a short-period γ-ray burst.
- The simultaneous detection of multiple signals from the same event is termed *multimessenger astronomy*.
- While technically demanding, it has the potential to lead to much deeper understanding of objects like neutron stars and core collapse supernovae.

Later in this chapter we will describe an *observed multimessenger event* involving a gamma-ray burst from merging neutron stars in coincidence with gravitational waves.

24.3 The Gravitational Wave Event GW150914

On September 14, 2015, almost 100 years had passed with no direct detection of Einstein's predicted gravitational waves.

- The *LIGO interferometers* in Livingston, Louisiana and Hanford, Washington were online and taking data.
- At 09:50:45 UTC the Livingston detector observed a strong transient lasting ~ 200 ms.
- 7 ms later Hanford observed a similar transient.
- These transients were identified within 3 minutes by generic low-latency scans as *a likely gravitational wave*.
- The signal had the obvious character of a *compact merger event* (the *chirp pattern* described below).
- Low-latency data pipelines scanning focused almost immediately on a possible gravitational wave from the *coalescence of two black holes*.
- Several months of thorough analysis confirmed with *greater than* 5σ *confidence* that
- the transient GW150914 was indeed a *gravitational wave emitted from the merger of two black holes*.

Thus GW150914 confirmed Einstein's century-old prediction that fluctuations of the spacetime metric could propagate as gravitational waves.

24.3.1 Observed Waveforms

The observed waveforms in the Livingston and Hanford detectors for GW150914 are shown in Fig. 24.4 (next page)

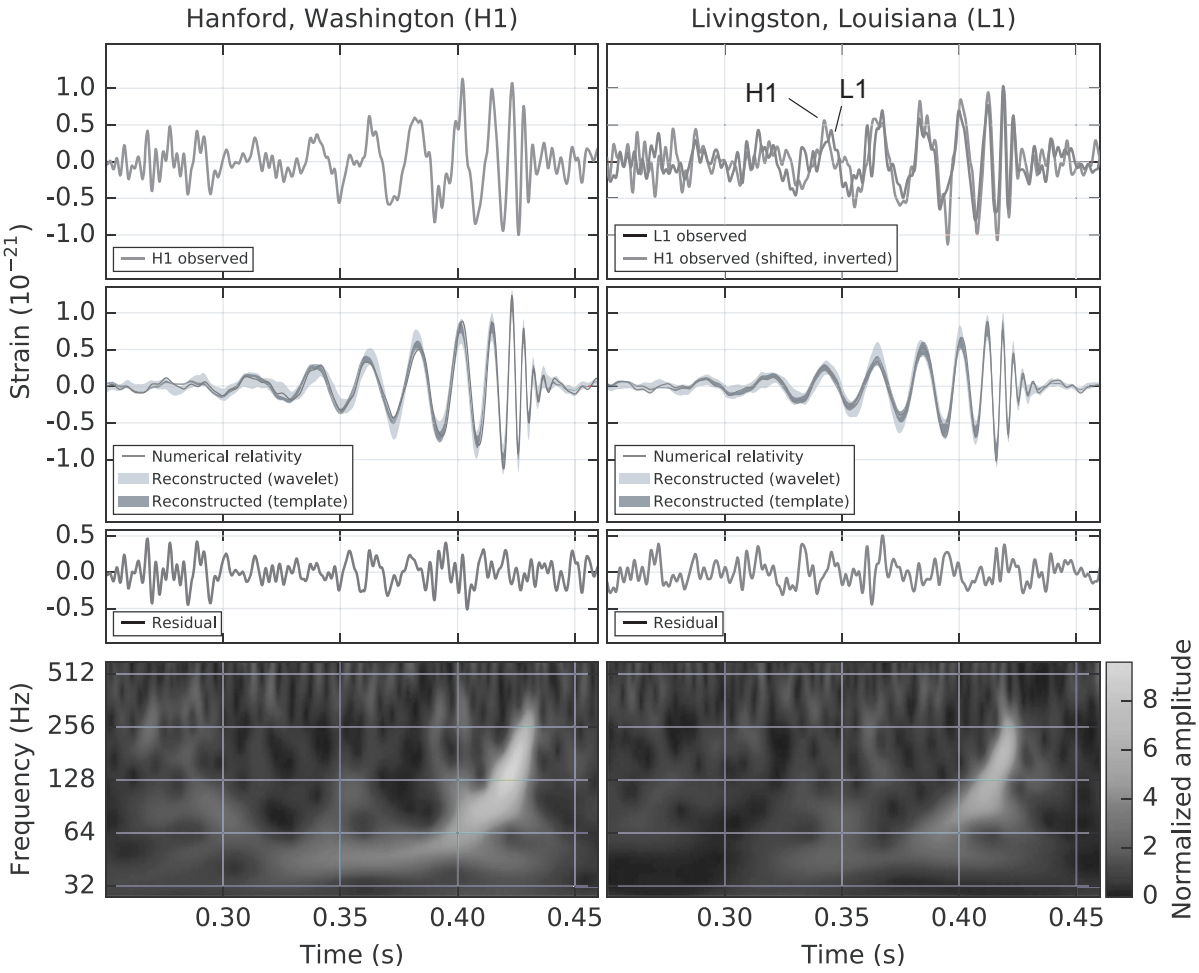
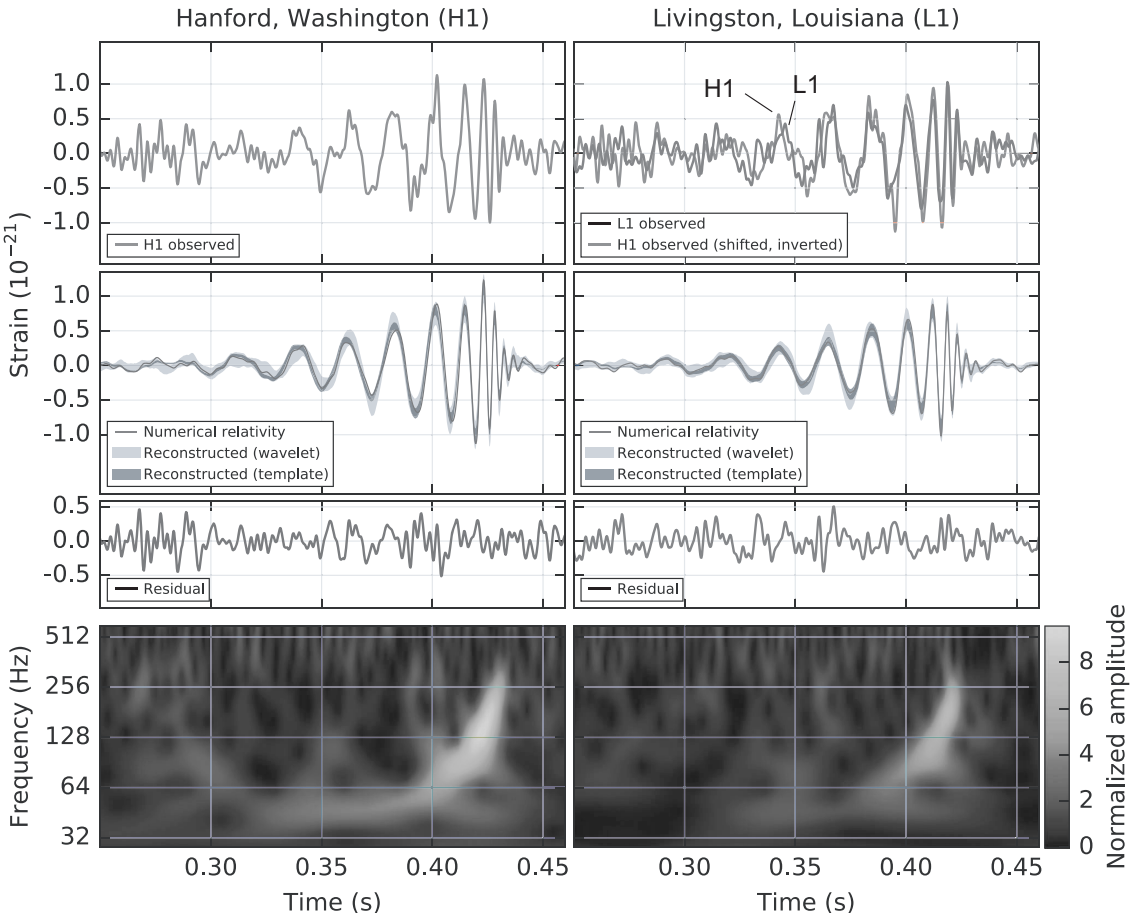
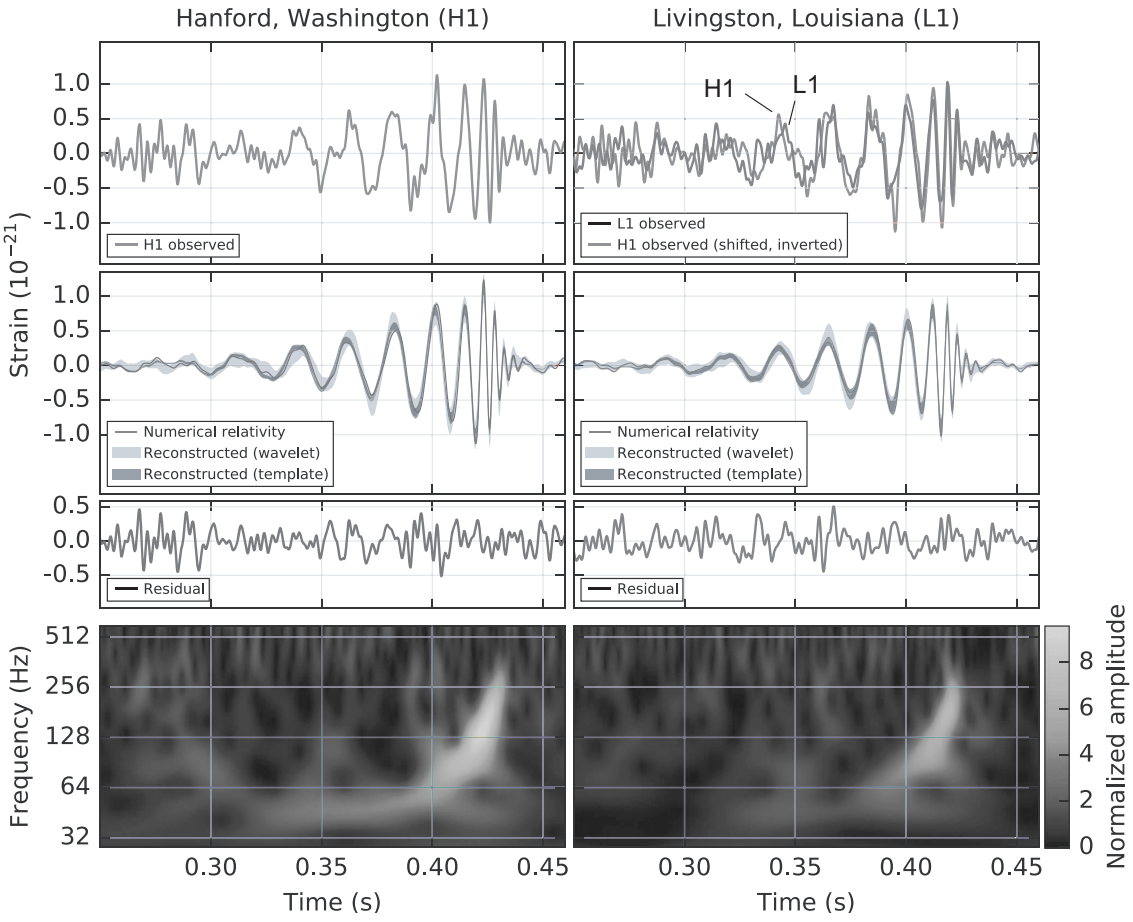


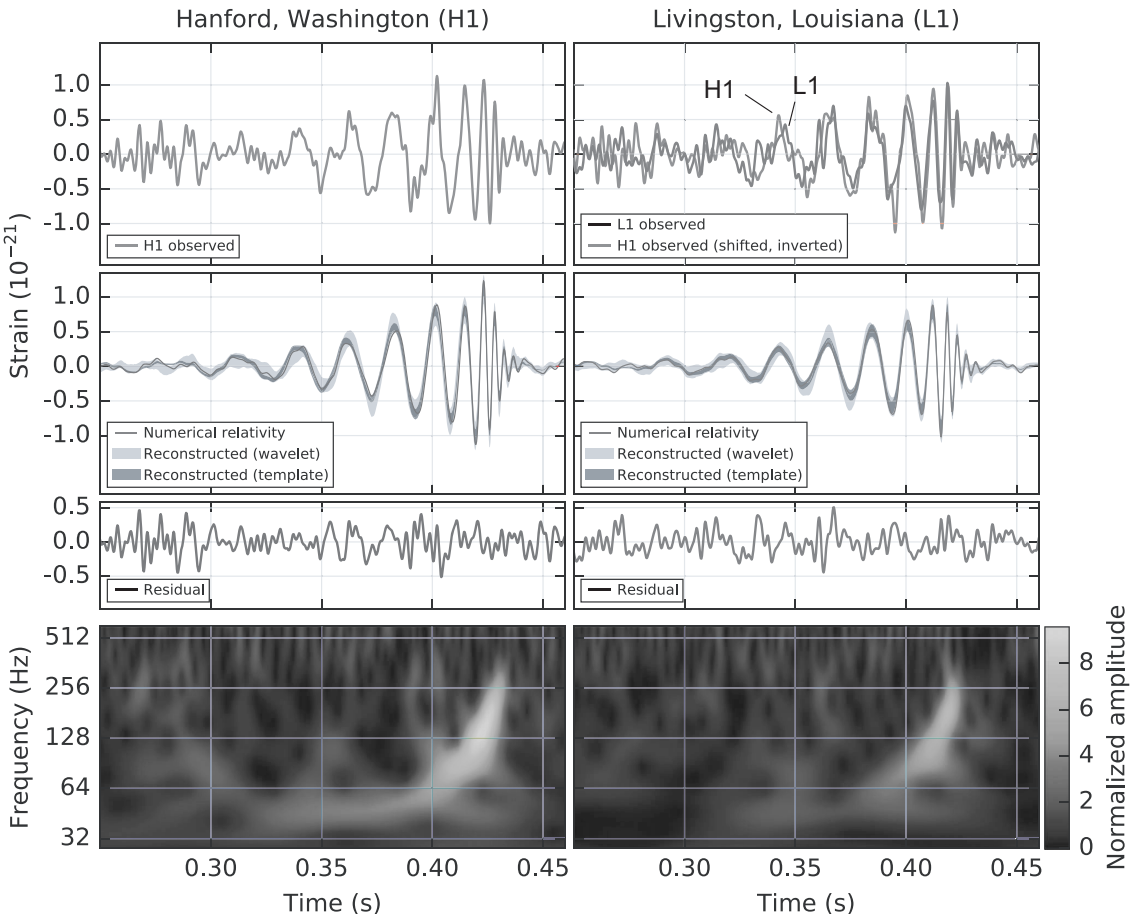
Figure 24.4: The gravitational wave event GW150914 observed by LIGO. The left panels correspond to data from the Hanford detector (H1) and the right panels to data from the Livingston detector (L1). All times are relative to 09:50:45 UTC on September 14, 2015. The top row is measured strain in units of 10^{-21}. In the top right panel the Hanford signal has been superposed on the Livingston signal as described in the text. The second row shows numerical relativity simulations of the waveform assuming a binary black hole merger event projected onto each detector in the 35-350 Hz band. Shaded areas indicate 90% credible regions for two independent waveform reconstructions. The third row shows the residuals after subtracting the numerical relativity waveform in the second row from the detector waveform in the first row. The fourth row shows strain versus frequency and time, with grayscale contours indicating strain amplitude. The rapidly-rising frequency pattern (chirp) is indicative of a binary merger event.



- The wave arrived first at the Livingston detector (L1) and then $6.9^{+0.5}_{-0.4}$ ms later at the Hanford detector (H1).
- In the top-right image the H1 wave has been superposed, shifted by 6.9 ms, and inverted to account for relative orientations of the two detectors (orientations relative to local north of L1 and H1 differ by 72°).
- This superposition and a *24:1 signal to noise ratio* leaves little doubt that the same wavefront, traveling at light-speed, passed first through Livingston and then Hanford.



- In the second row of the above figure, numerical relativity simulations of the waveform for merging black holes and wavelet reconstructions with and without an astrophysical black hole merger model are shown.
- The third row displays the result of subtracting the numerical relativity waveform in the second row from the observed waveform in the first row.
- The last row shows a *time-frequency representation of the data*, with the grayscale contours representing strain.



- The frequency–time plot in the bottom row indicates that over a period of ~ 0.2 seconds the signal swept upward in frequency from about 35 Hz to 250 Hz.
- This signal, rising in frequency and strain (*"the chirp"*), is indicative of the final rapid inspiral of a merger event, with a peak strain $\sim 1.0 \times 10^{-21}$.
- This changed the test mass separation by $\sim 4 \times 10^{-16}$ cm *(0.005 times the diameter of a proton)*.

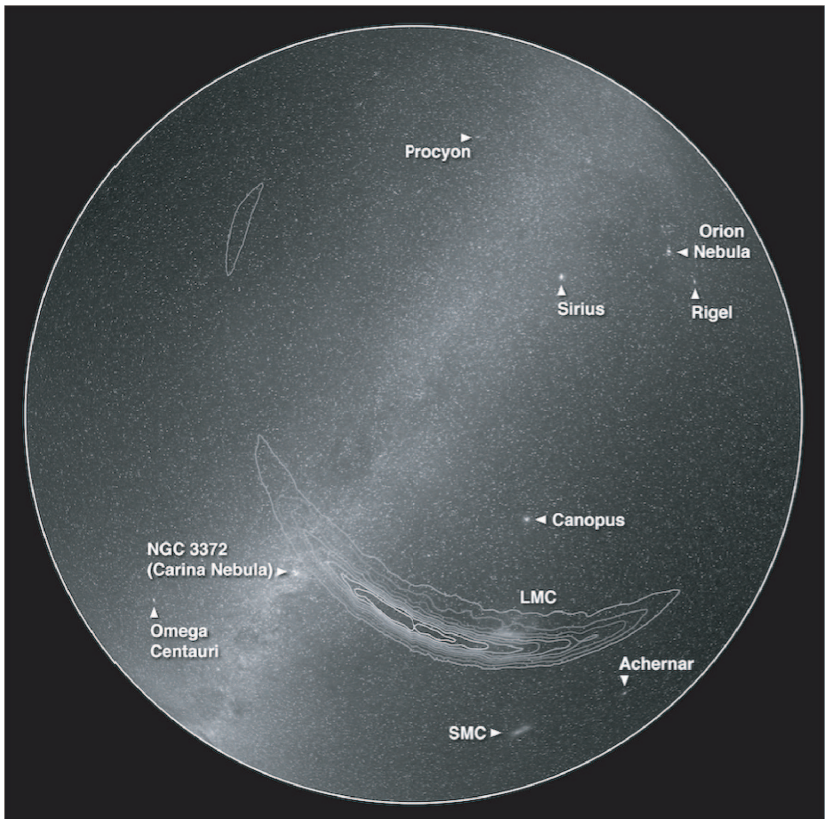
24.3.2 Source Localization

GR predicts that *gravitational waves travel at lightspeed*.

- The *light-travel time* from Livingston to Hanford is 10 ms.
- A valid coincidence between L1 and H1 must lie in a ± 10 ms window, depending on wavefront orientation.
- The actual difference in time for the coincidence then provides *primary directional information* for the source.
- *Amplitude and phase information* in the two detectors then can be used to further *refine the location*.
- The observed time difference was ~ 7 ms,
- with the gravitational wave passing through Livingston first.

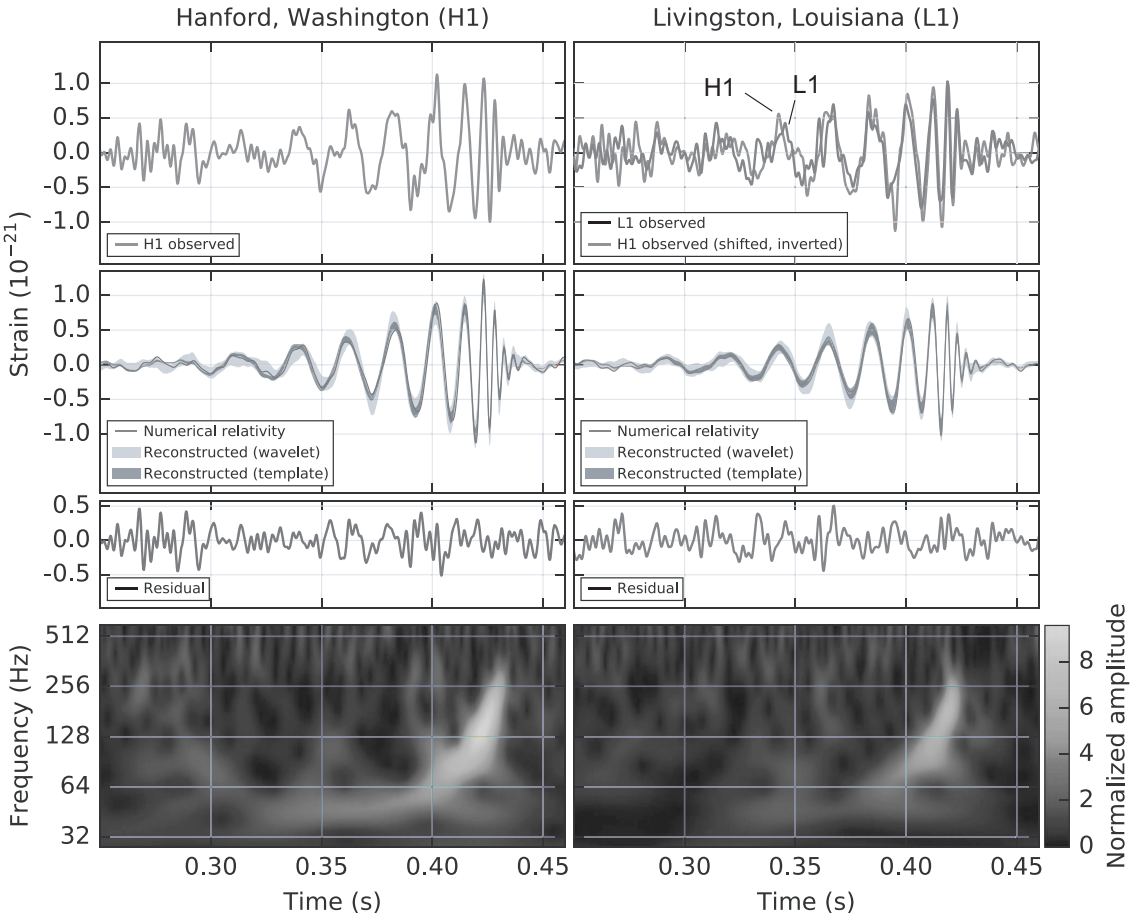
This indicates that the wave source was in the *southern-hemisphere sky*.

- At the time of observation no other gravitational wave observatories were acquiring data.
- Thus only the two LIGO detectors were available to determine source position, primarily through arrival time.



- Data from L1 and H1 permitted the source to be localized to an area of $\sim 600\,\mathrm{deg}^2$, as illustrated in the above figure.
- Further analysis improved this to $230\,\mathrm{deg}^2$.
- Follow-up observations at radio, optical, near-IR, X-ray, and gamma-ray wavelengths reported *no obvious electromagnetic counterpart* to the gravitational wave event.

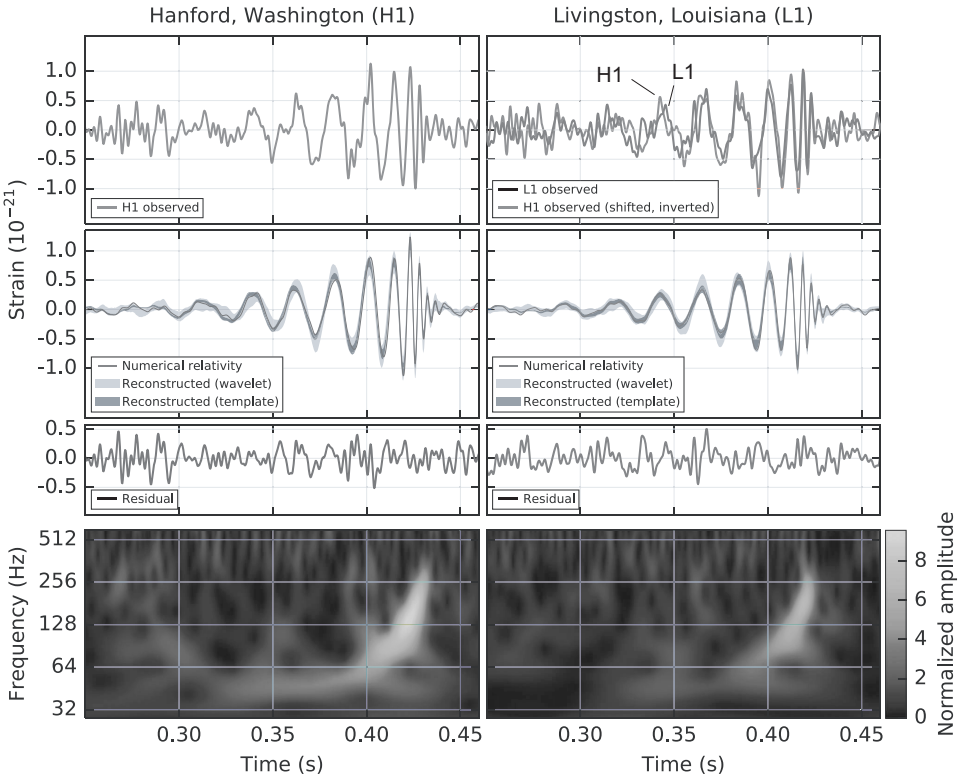
Thus GW150914 appears to have been *invisible to non-gravitational observatories*.



24.3.3 Comparisons with Candidate Events

The qualitative features of the data provide strong evidence that GW150914 corresponded to a binary merger event.

- Over 200 ms the frequency increased from 35 to 150 Hz, and the amplitude increased to maximum over ~ 8 cycles.
- This suggests the *inspiral of orbiting masses* m_1 *and* m_2 caused by gravitational wave emission.

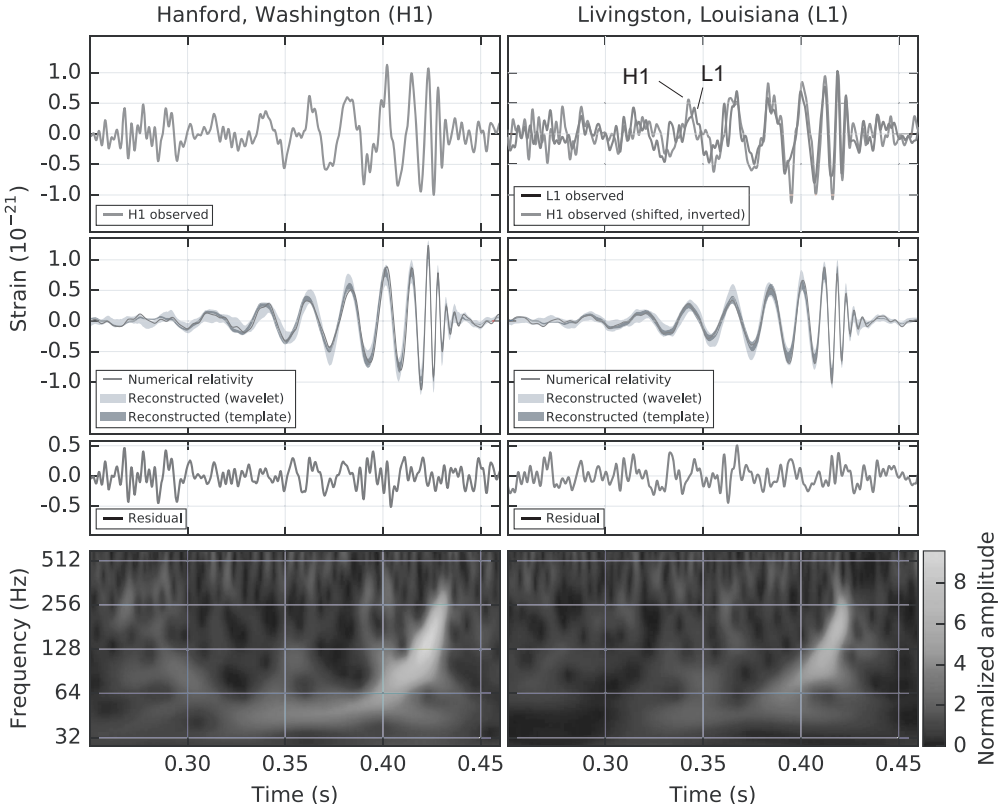


- To leading order in v/c, this evolution is characterized by a particular mass combination called the *chirp mass* $\mathscr{M}$,

$$\mathscr{M} = \mu^{3/5} M^{2/5} = \frac{(m_1 m_2)^{3/5}}{(m_1 + m_2)^{1/5}} = \frac{c^3}{G}\left(\frac{5}{96}\pi^{-8/3} f^{-11/3} \dot{f}\right)^{3/5},$$

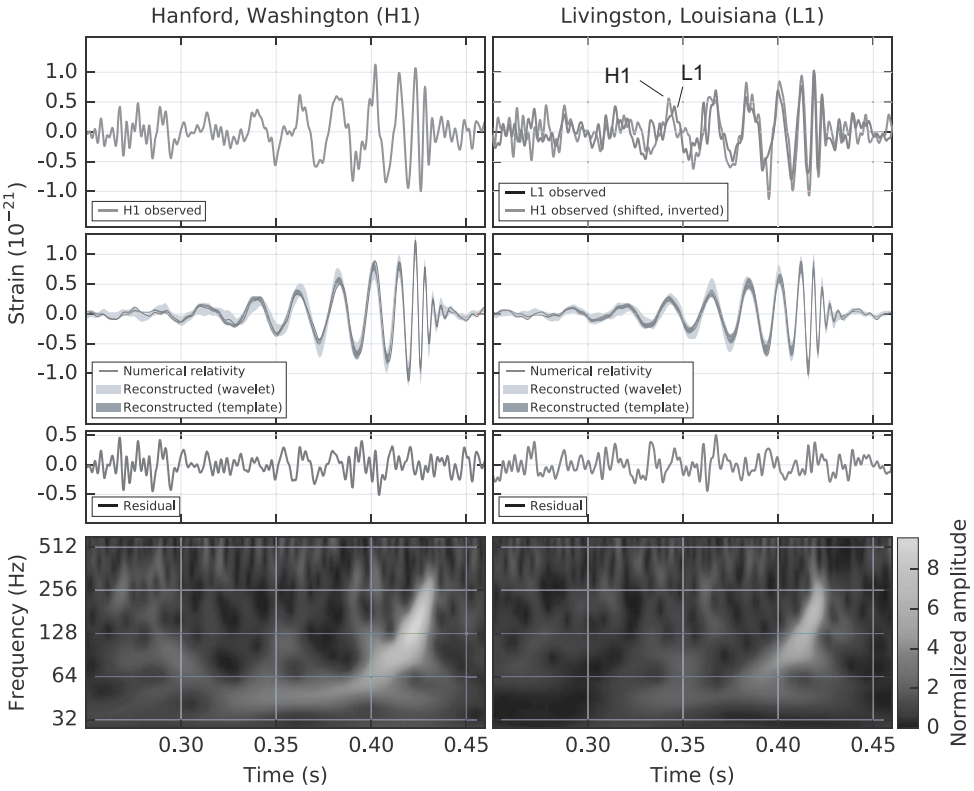
  - where $\mu = m_1 m_2/(m_1 + m_2)$ is the *reduced mass*,
  - $M = m_1 + m_2$ is the *total mass*,
  - f is the observed *frequency*, and $\dot{f}$ *its time derivative*.

- The last expression for $\mathscr{M}$ follows from retaining *lowest order in a post-Newtonian expansion*.



Thus the chirp mass can be deduced from the observed wave, and this in turn implies a relationship between m_1 and m_2.

- Estimating f and $\dot{f}$ from the data yields $\mathscr{M} \simeq 28\,M_\odot$.
- This implies a total mass $M = m_1 + m_2 \geq 70\,M_\odot$ in the detector rest frame (Problem).
- The sum of the Schwarzschild radii for the binary components is thus constrained to be $2GM/c^2 \geq 210\,\mathrm{km}$.
- This implies that to reach an orbital frequency of $75\,\mathrm{Hz}$ (half the maximum gravitational wave frequency)
- the objects had to be *very compact with small separation*.



- *Binaries involving white dwarfs are excluded* because they are neither compact enough nor massive enough.
- *Binary neutron stars are eliminated* because they are compact enough but not massive enough.
- *A neutron star, black hole binary is eliminated* because it would be required by the chirp mass to be very massive (Problem) and so would *merge at much lower frequency*.

Only *merger of 2 black holes with* $m_1 + m_2 \sim 60 - 70\ M_\odot$ is consistent with the data.

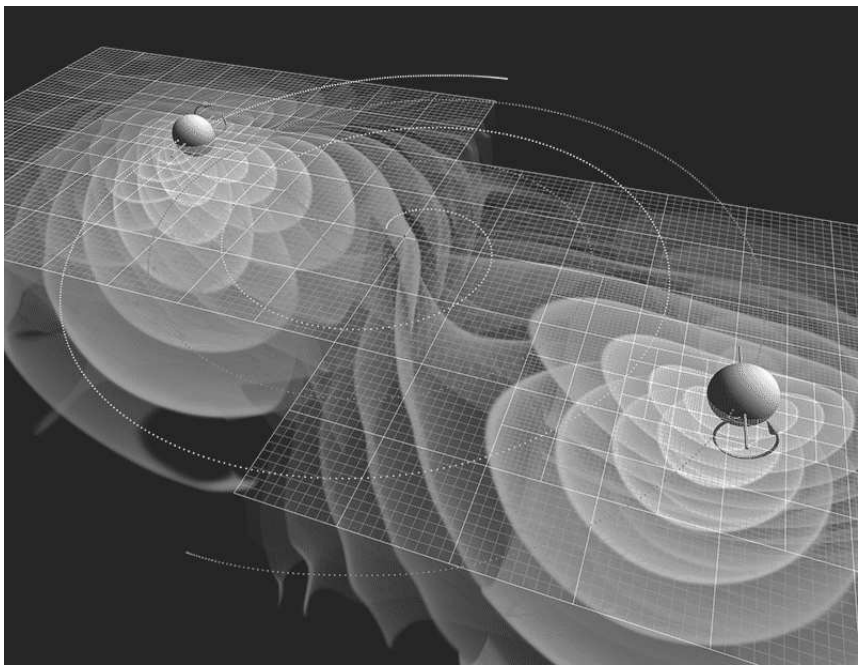
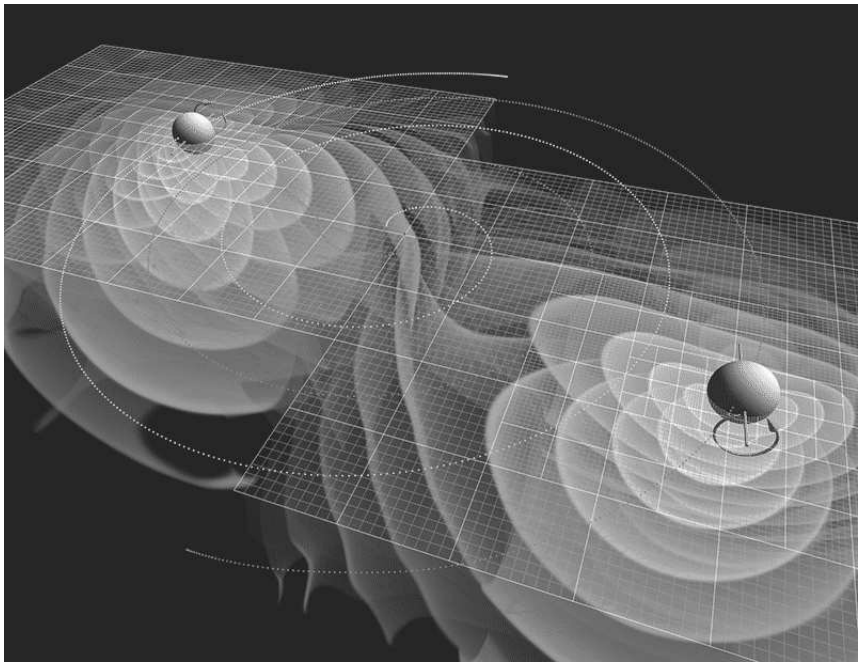


Figure 24.5: Numerical simulation of gravitational wave emission from merger of two black holes. Grayscale contours represent the amplitude of the gravitational radiation. Orbits of the black holes are indicated by dotted curves and their spins are indicated by the arrows.

24.3.4 Binary Black Hole Mergers

Simulation of such a merger is shown in Fig. 24.5. Comparisons with candidate templates indicated that

- GW150914 corresponded to the *merger of two* $\sim 30\ M_\odot$ *black holes* to form a $60\ M_\odot$ *Kerr black hole*.
- The final black hole had a spin $\sim 70\%$ of the maximum possible for a Kerr black hole.
- Approximately $\sim 3\ M_\odot c^2$ of gravitational wave energy was emitted in the merger.



- From the black hole merger model the peak gravitational wave luminosity was $\sim 3.6 \times 10^{56}\,\mathrm{erg\,s^{-1}}$.
- This equals the conversion of *~200 solar masses per second* into gravitational-wave energy.
- (100,000 times larger than the peak photon luminosity of a supernova)!
- From the strain amplitude a luminosity distance of $230 - 550\,\mathrm{Mpc}$ was inferred.
- This corresponds to a redshift $z = 0.05 - 0.12$.
- Assuming the standard cosmology, this redshift implies that the merger occurred $12.2 - 13.1\,Gyr$ *after the big bang*.

Table 24.1: Properties of three LIGO black-hole merger candidates

Quantity	GW150914	GW151226	LVT151012
Signal to noise ratio	23.7	13.0	9.7
False alarm rate (yr^{-1})	$< 6.0 \times 10^{-7}$	$< 6.0 \times 10^{-7}$	0.37
Significance	$> 5.3\sigma$	$> 5.3\sigma$	1.7σ
Primary black hole mass ($M_\odot$)	$36.2^{+5.2}_{-3.8}$	$14.2^{+8.3}_{-3.7}$	23^{+18}_{-6}
Secondary black hole mass ($M_\odot$)	$29.1^{+3.7}_{-4.4}$	$7.5^{+2.3}_{-2.3}$	13^{+4}_{-5}
Chirp mass ($M_\odot$)	$28.1^{+1.8}_{-1.5}$	$8.9^{+0.3}_{-0.3}$	$15.1^{+1.4}_{-1.1}$
Total mass ($M_\odot$)	$65.3^{+4.1}_{-3.4}$	$21.8^{+5.9}_{-1.7}$	37^{+13}_{-4}
Final black hole mass ($M_\odot$)	$62.3^{+3.7}_{-3.1}$	$20.8^{+6.1}_{-1.7}$	35^{+14}_{-4}
Final black hole spin	$0.68^{+0.05}_{-0.06}$	$0.74^{+0.06}_{-0.06}$	$0.66^{+0.09}_{-0.10}$
Radiated mass ($M_\odot$)	$3.0^{+0.5}_{-0.4}$	$1.0^{+0.1}_{-0.2}$	$1.5^{+0.3}_{-0.4}$
Peak luminosity ($\mathrm{erg\,s}^{-1}$)	$3.6^{+0.5}_{-0.4} \times 10^{56}$	$3.3^{+0.8}_{-1.6} \times 10^{56}$	$3.1^{+0.8}_{-1.8} \times 10^{56}$
Source redshift z	$0.09^{+0.03}_{-0.04}$	$0.09^{+0.03}_{-0.04}$	$0.20^{+0.09}_{-0.09}$
Luminosity distance (Mpc)	420^{+150}_{-180}	440^{+180}_{-190}	1000^{+500}_{-500}
Sky localization (deg^2)	230	850	1600

Masses in source frame. Multiply by $(1+z)$ for detector frame. Redshift assumes standard cosmology. Spin given in units of spin for an extremal Kerr black hole.

- Parameters for the black hole merger GW150914 determined from the waveform simulations compared with data are displayed in column two of Table 24.1.
- Also shown are data for two other events.
  - GW151226 was a confirmed black hole merger;
  - LVT151012 likely was also, but its confidence level of 1.7σ did not permit claiming it as such.

- The analysis is in principle able to determine the spins of the initial and final black holes,
- with the final spin generally more tightly constrained than the initial spins.
- Only the spin of the final Kerr black hole is shown in the table because in the present data analysis uncertainties for initial spins are of order 100%.

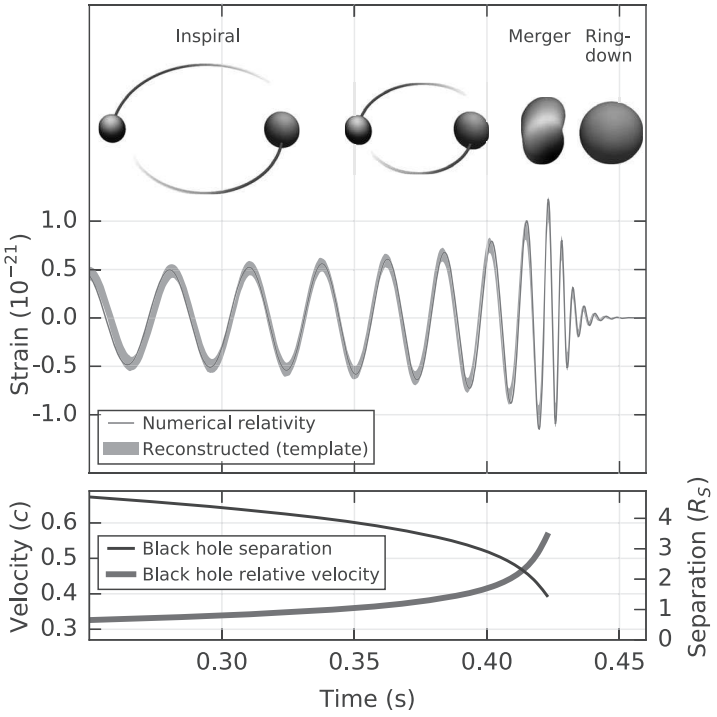


Figure 24.6: Estimated gravitational-wave strain projected onto detector H1 for GW150914. The inset image illustrates the black hole horizons as the two black holes coalesce, as deduced from numerical relativity calculations.

More GW150914 simulation details are given in Fig. 24.6.

- The inset image illustrates the event horizons of the two black holes merging into that of a single final black hole.
- The bottom panel shows black hole separation versus time in units of $r_S = 2GM/c^2$, and velocity in units of c, where
  - M is the total mass, the *relative velocity* is
$$\frac{v}{c} = \left(\frac{GM\pi f}{c^3}\right)^{1/3},$$
  - and f is the gravitational wave frequency.

In Fig. 24.7 (next page) a simulation of what the black holes might have looked like from up close during the merger is shown.

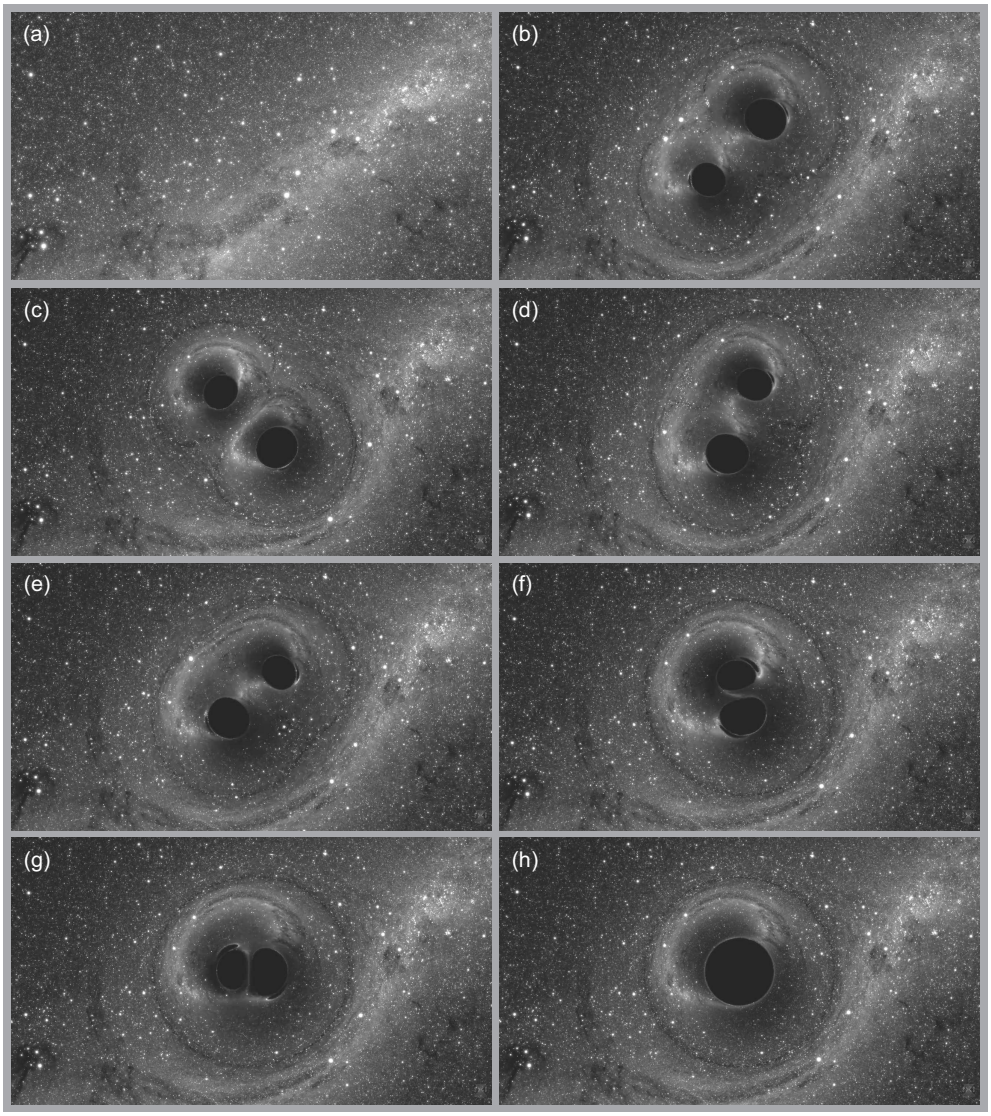
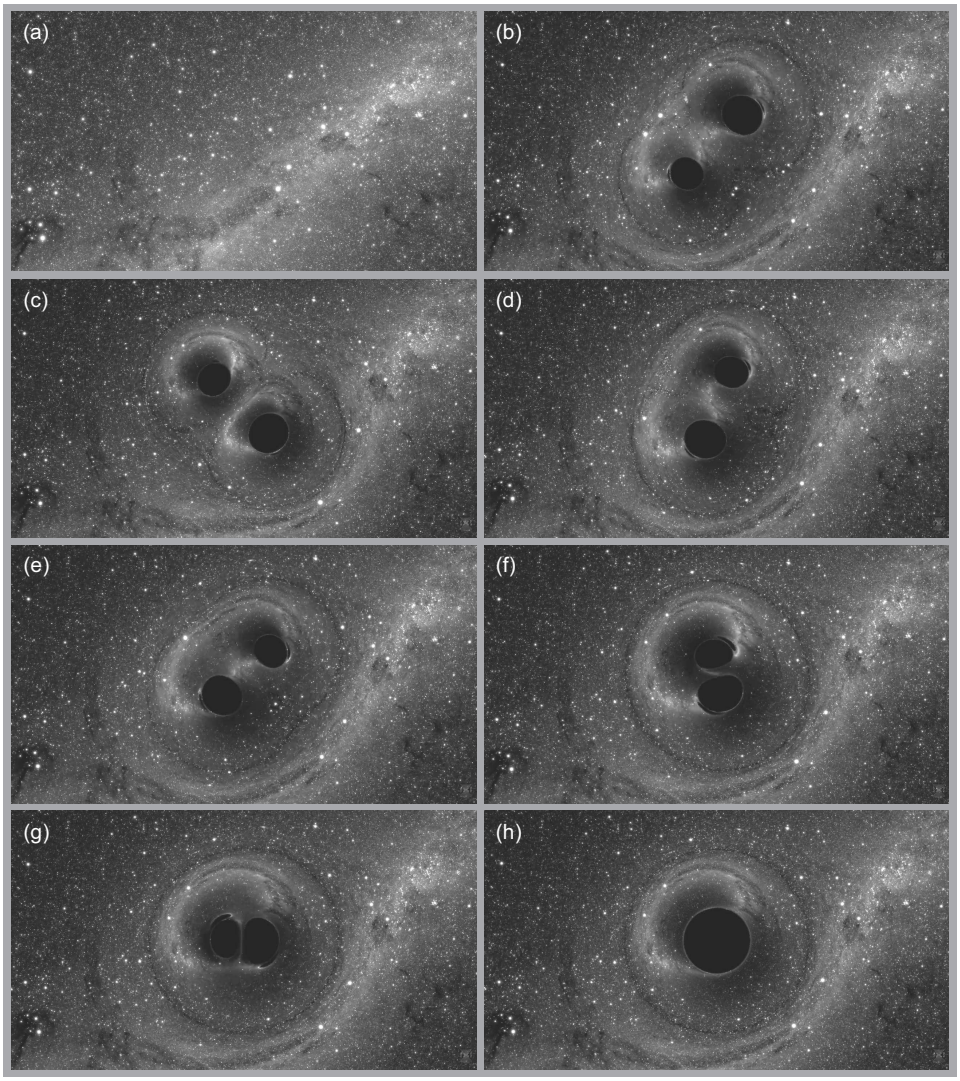


Figure 24.7: Computer simulation of the GW150914 merger. (a) The undistorted background field of stars in the absence of the black holes. (b)–(g) Successively later times in the merger sequence. (h) The final Kerr black hole. Notice the strong gravitational lensing effects near the black holes. The background stars are fixed in position as in (a) for each panel but the gravitational lensing completely distorts their apparent positions. The ring around the black holes is an *Einstein ring* caused by strong focusing of light from stars behind the black holes.



- The dark, well-defined shapes are the shadows of the black hole event horizons as they block all light from behind.
- The flattened dark features around them and distorted star fields are strong gravitational lensing effects.

24.3.5 Matched Filtering

Detection of gravitational waves often relies on *matched filtering.*

- The method originated in applications such as radar and 2D image processing.
- Matched filtering correlates
  - a known (template) signal with a measured signal
  - to ascertain the presence of the template in the measured signal.
- It is considered to be the *optimal linear filter* for maximizing signal to noise in the presence of added random noise.

For gravitational waves, libraries of template waveforms computed theoretically are used to analyze the detected signal for a match.

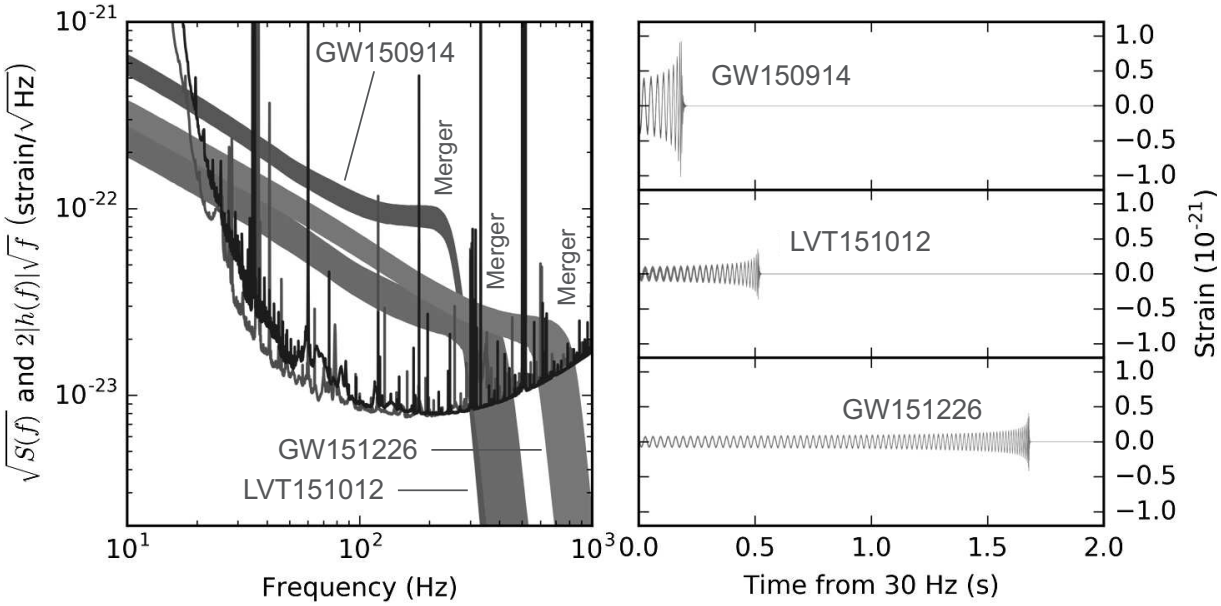


Figure 24.8: (*Left*) Amplitude spectral density of strain noise for the L1 (black spectrum) and H1 (gray spectrum) detectors, and recovered signals of GW150914, GW151226, and LVT151012 (bands). (*Right*) Time evolution of the recovered strain signals from a reconstruction..

Strain signal and noise levels are displayed in the left panel of Fig. 24.8 for the first three LIGO gravitational wave candidates.

- The amplitude of GW150914 is large and at merger it lies well above the detector noise spectrum.
- Hence a signal to noise ratio of ~ 24 for GW150914.
- For GW151226 and LVT151012 the amplitudes are much closer to the noise at merger, leading to lower values of signal to noise (13.0 and 9.7, respectively).

In these latter events, *matched filtering was essential* to extract the waveform from the noise.

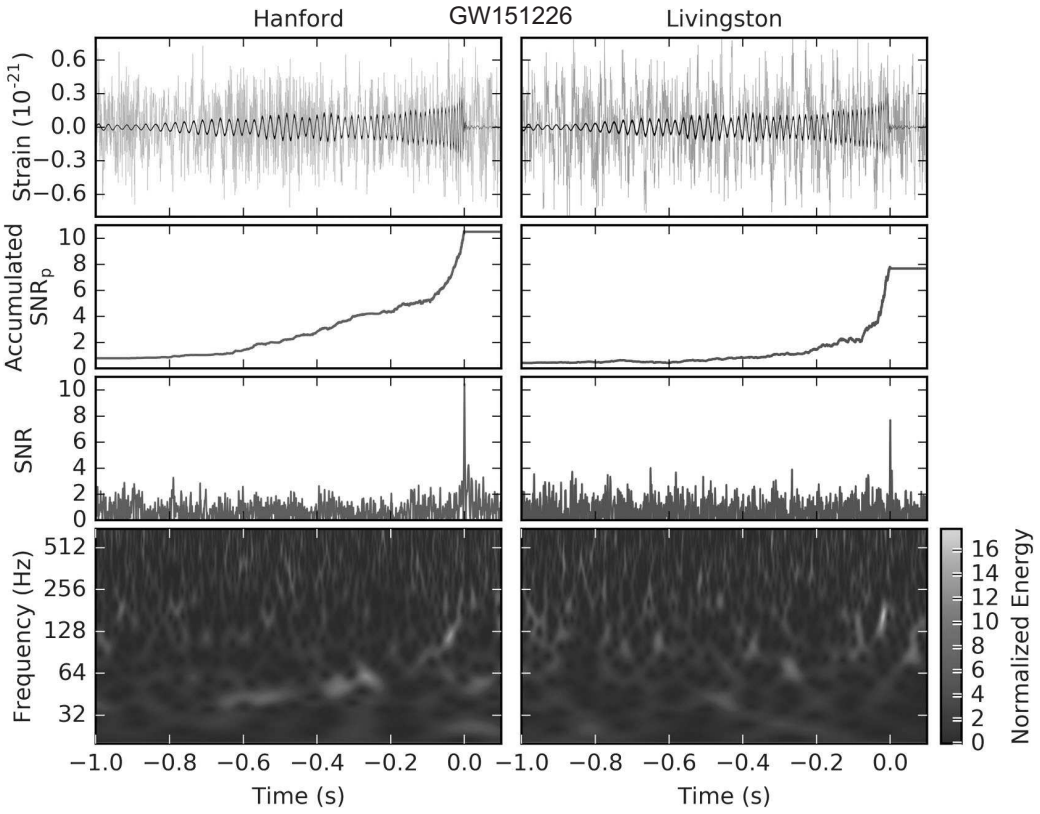


Figure 24.9: Top: strain (gray) and template wave (black); Next two panels: signal to noise ratio (SNR); Bottom: energy vs. frequency and time.

Consider the discovery of GW151226 (Fig. 24.9):

- *Top panel:* Strain data filtered to reject known noise in gray; best-match template with same filters in black.
- The signal to noise ratio is in general poor, as indicated in the next two panels, and in the frequency–time plot shown in the bottom panel almost no structure can be discerned.

Matched filtering was *essential to even identify GW151226* as a gravitational wave.

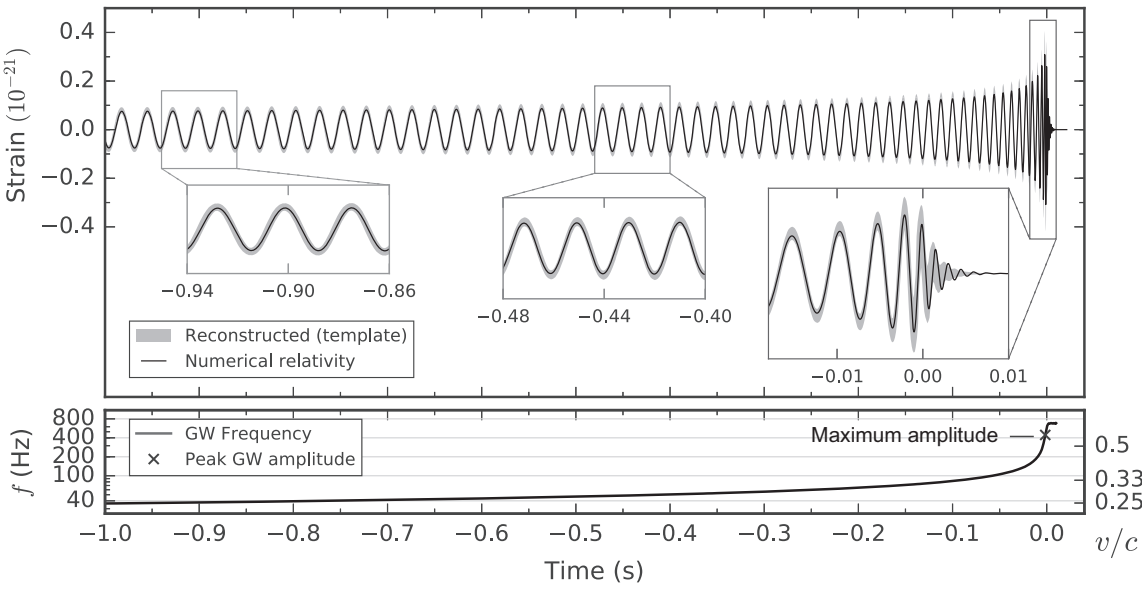
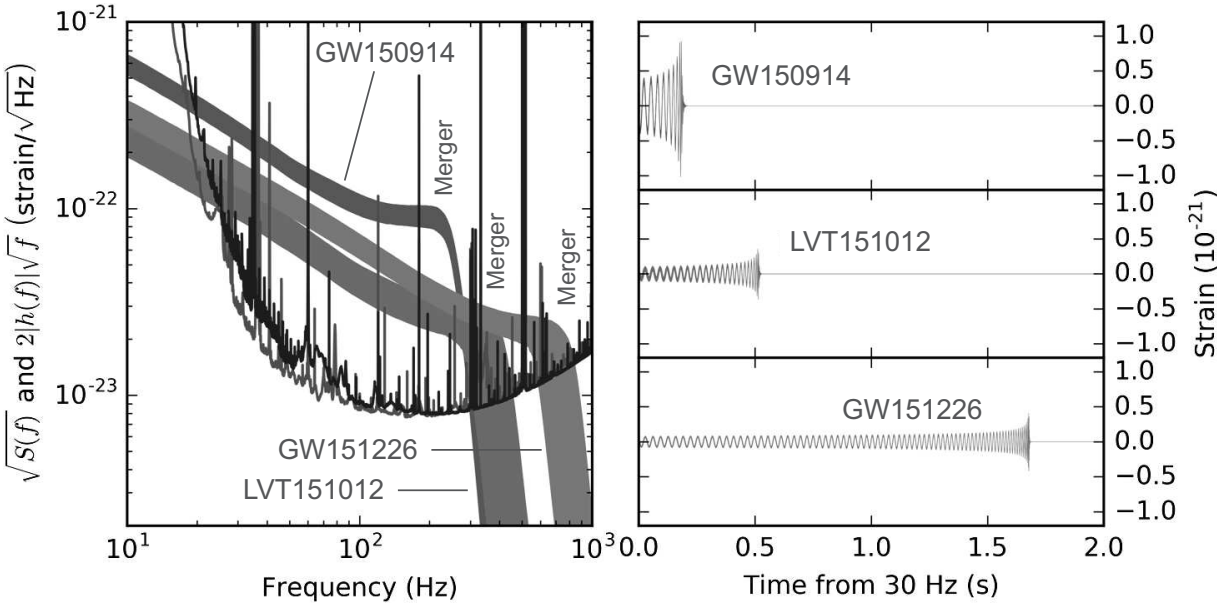


Figure 24.10: Estimated gravitational wave strain for GW151226 projected onto the LIGO Livingston detector (full bandwidth, without filtering). *Top:* The reconstructed wave (90% credible region) is shown in gray and a numerical relativity simulation in black. *Bottom:* Gravitational wave frequency (left axis) computed from numerical relativity waveform. The cross marks the maximum amplitude, approximately coincident with merger of the black holes. The right axis gives an effective relative velocity v/c that can be related to f during the inspiral using post-Newtonian approximations.

The reconstructed wave for GW151226 is shown in Fig. 24.10.

- GW151226 has been confirmed to be a gravitational wave at the 5.3σ confidence level.
- Analysis indicates that this event corresponded to
  - merger of $14M_\odot$ and $8M_\odot$ black holes
  - at a distance of about 440 Mpc.

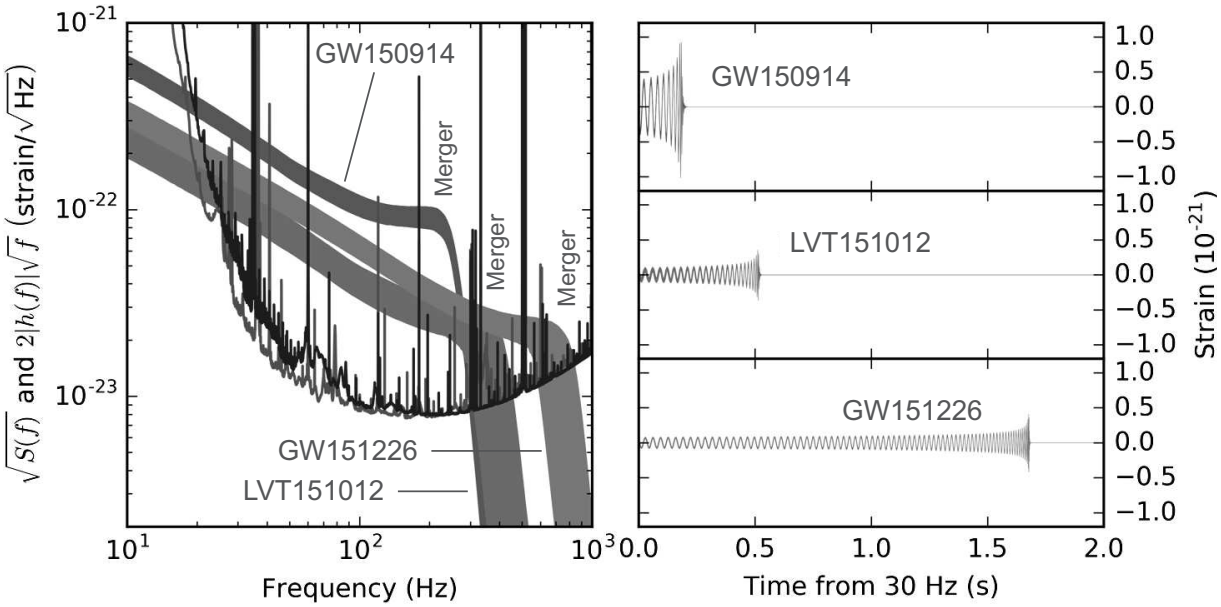


24.3.6 Binary Masses and Inspiral Cycles

The right side of the above figure indicates that

- many more oscillations were detected in GW151226 over a longer period ($\sim$ 55 cycles over almost 2 seconds)
- than for GW150914 ($\sim$ 8 cycles over 0.2 seconds).
- This is a consequence of the *lower masses for the merger in* GW151226 relative to GW150914,
- which means that the GW151226 wave spent more time in the detector sensitive band above about 30 Hz.

The 55 cycles observed for GW151226 extended the effective reach of the instrument by a factor proportional to $\sqrt{55}$ for that event.



24.3.7 Binary Masses and Inspiral Cycles

The dependence of gravitational wave frequency on black hole masses in mergers has some important consequences:

1. The lower-mass merger GW151226 was more difficult to detect than GW150914 because of smaller strain, but
   - because GW151226 was observed over many cycles, *matched filtering was particularly effective*, and
   - lower-mass GW151226 was *more sensitive to some inspiral parameters* than higher-mass GW150914.
2. Mergers more massive than for GW150914 will be *increasingly difficult for LIGO to study* because fewer cycles will come above the detector cutoff at ~ 30 Hz.

24.3.8 LIGO–Virgo Triple Coincidences

A *network of multiple interferometers* can determine the location of a gravitational wave source by triangulation using

- *time differences*,
- *phase differences*, and
- *amplitude ratios* for arrival of the gravitational wave.

When *advanced Virgo came online in 2017*, simultaneous detection of a gravitational wave by the *two LIGO detectors* and *Virgo* became possible.

- The *first such triple coincidence* was realized in the detection of GW170814.
- GW170814 corresponded to the *merger of* $30.5\,M_\odot$ *and* $25.3\,M_\odot$ *black holes* to produce
- a *final* $53.2\,M_\odot$ *black hole* with *spin* $\sim 70\%$ *of maximal* for a Kerr black hole.
- The merger occurred at a *luminosity distance* of 540 Mpc (*redshift* $z \sim 0.11$).
- It released a total of $\sim 2.7 M_\odot c^2$ *of gravitational-wave energy*, and
- reached a *peak luminosity of* $\sim 3.7 \times 10^{56}\,\mathrm{erg\,s^{-1}}$.

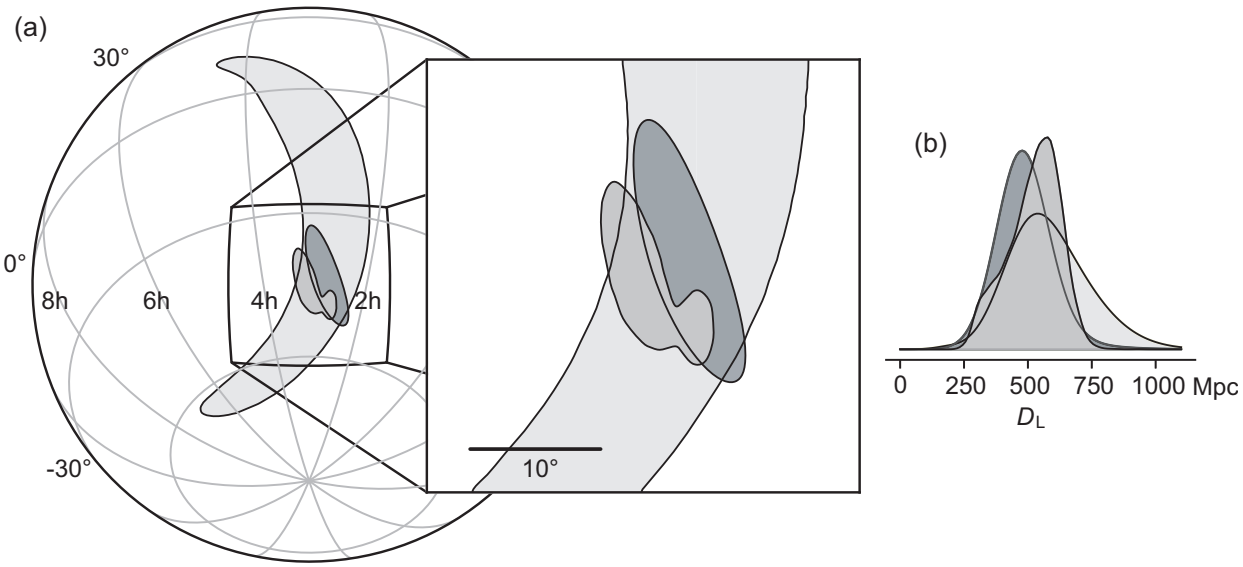
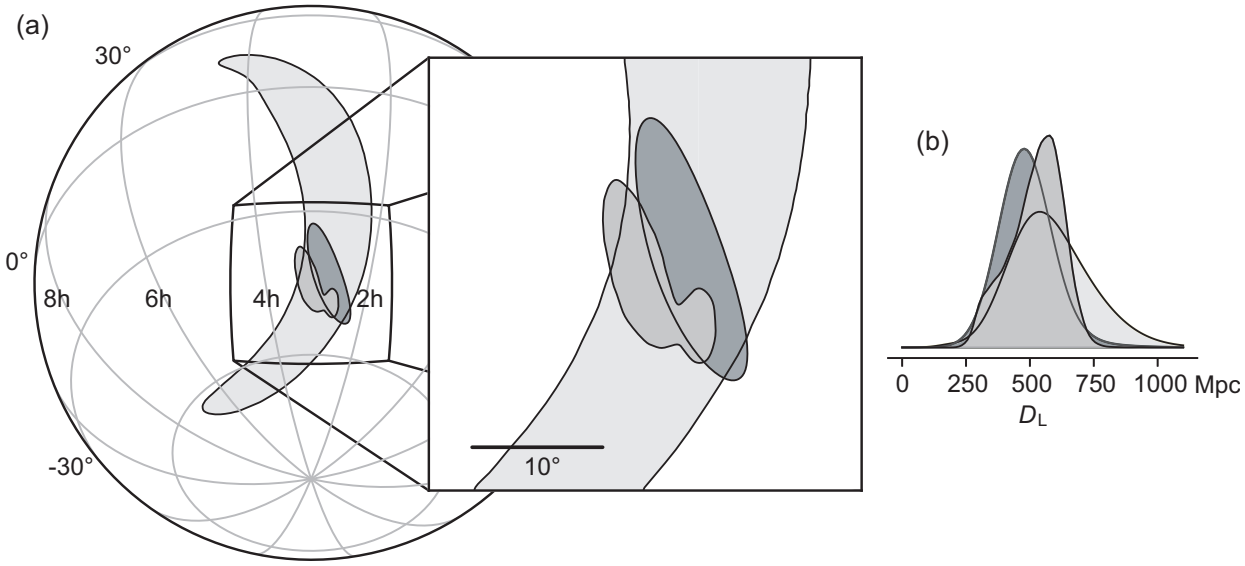


Figure 24.11: Localization of GW170814 by a LIGO–Virgo triple coincidence. Lightest gray indicates LIGO rapid localization, darkest gray represents initial LIGO–Virgo coincidence, and intermediate gray represents the triple coincidence after full analysis. All contours are 90% confidence level. (a) Position on the celestial sphere. (b) Luminosity distance.

The increased localization both in

- *angular position* and
- *luminosity distance*

afforded by the *LIGO–Virgo* triple coincidence for GW170814 is illustrated in Fig. 24.11.



- The *angular localization* was improved from
  - an uncertainty box of $1160\,\mathrm{deg}^2$ in the LIGO rapid response alone to
  - $60\,\mathrm{deg}^2$ in the the LIGO–Virgo triple coincidence.
- This also permitted the *luminosity distance* to be sharpened from
  - $570^{+300}_{-230}\,\mathrm{Mpc}$ in the initial LIGO rapid response to
  - $540^{+130}_{-210}\,\mathrm{Mpc}$ in the final triple-coincidence analysis.

The overall effect was to *decrease by an order of magnitude the event volume* relative to LIGO analysis alone.

24.4 Testing General Relativity in Strong Gravity

General relativity has passed all tests but earlier tests have all been in the regime of relatively *weak gravity.*

- The most interesting predictions and most stringent tests of GR are in the *dynamical, highly nonlinear, strong-field regime* such as the merger of black holes, where
  - gravitational fields are very large,
  - space is strongly and dynamically curved, and
  - velocities approach the speed of light.
- All prior evidence for black holes was indirect, typically depending on electromagnetic radiation emitted far from the event horizon (from relatively weak gravity).

  The best strong-gravity tests before GW150914 were in high-gravity binary pulsar systems. But these still probe *relatively weak gravity*.

- The signature for GW150914 near peak strain and in ring-down bears imprints of the *strong field near the horizons* of the merging and final black holes.

  GW150914 is perhaps *the most direct evidence yet for existence of black holes*, and of whether their properties are consistent with GR.

Analysis of the first gravitational waves detected by LIGO indicate *no obvious breakdown of GR*.

- To give one example, general relativity predicts that *gravitational waves should travel at the speed of light*.
- This implies that the putative graviton mediating the gravitational force is *massless*.
- Hence any deviations of gravitational wave speed from c may be parameterized in terms of a *finite graviton mass*.
- If the graviton had a finite mass it would *travel at* $v < c$ and *there would be a dispersion of the wave* (lower frequencies traveling slower than higher frequencies).
  - The LIGO/Virgo analysis of GW170104 found *no evidence for such an effect* and placed an upper limit for the *mass of the graviton* of

$$m_g \leq 7.7 \times 10^{-23}\,\mathrm{eV}/c^2.$$

  - Thus GW170104 provided strong evidence that *the speed of gravity is indeed equal to the speed of light*.
  - An even more stringent limit on the speed of gravity set by merging neutron stars will be discussed below.

Thus initial data suggest that the basics of GR in strong fields are correct, but *even more stringent tests should be forthcoming* as more gravitational waves are detected.

24.5 A New Window on the Universe

The most enduring contribution of the gravitational waves detected in Advanced LIGO's and Virgo's first observing runs may be that

> It opens a *new window on the Universe* for strong-gravity events.

For example,

- the detection of GW150914,
- the enormous radiated power (a peak luminosity of more than $10^{56}\,\mathrm{erg\,s^{-1}}$) inferred from the data, and that
- no reproducible electromagnetic observations signalled the event despite a concerted effort, indicate that

> There could be many events occurring in our Universe that
>
> - *emit gravitational waves of incredible power* but that
> - leave *little or no electromagnetic footprint*.

24.6 Gravitational Waves from Neutron Star Mergers

On August 17, 2017, LIGO–Virgo detected GW170817.

- This gravitational wave had a *very different signature* relative to previous black hole merger events.
- The amplitude and frequency built slowly with *more than* 3000 cycles *recorded over* $\sim$ 100 seconds before peak.

This new kind of GW would soon be interpreted as originating in *merger of 2 neutron stars*, but the show wasn't over yet!

- Approximately *1.7 seconds after peak GW strain* both the
  - Fermi Gamma-ray Space Telescope *(Fermi)* and the
  - International Gamma-Ray Astrophysics Laboratory *(INTEGRAL)*

  observed a *gamma-ray burst of two seconds duration* in the same part of the sky as the GW source.
- Within hours various observatories discovered a *new point source* in the irregular/elliptical galaxy NGC 4993, *lying within the position error box* for the gravitational wave.
- In the ensuing weeks a multitude of observatories studied the *transient afterglow in* NGC 4993 (named officially AT 2017gfo) intensively at various wavelengths.

Thus was the discipline of *multimessenger astronomy* born.

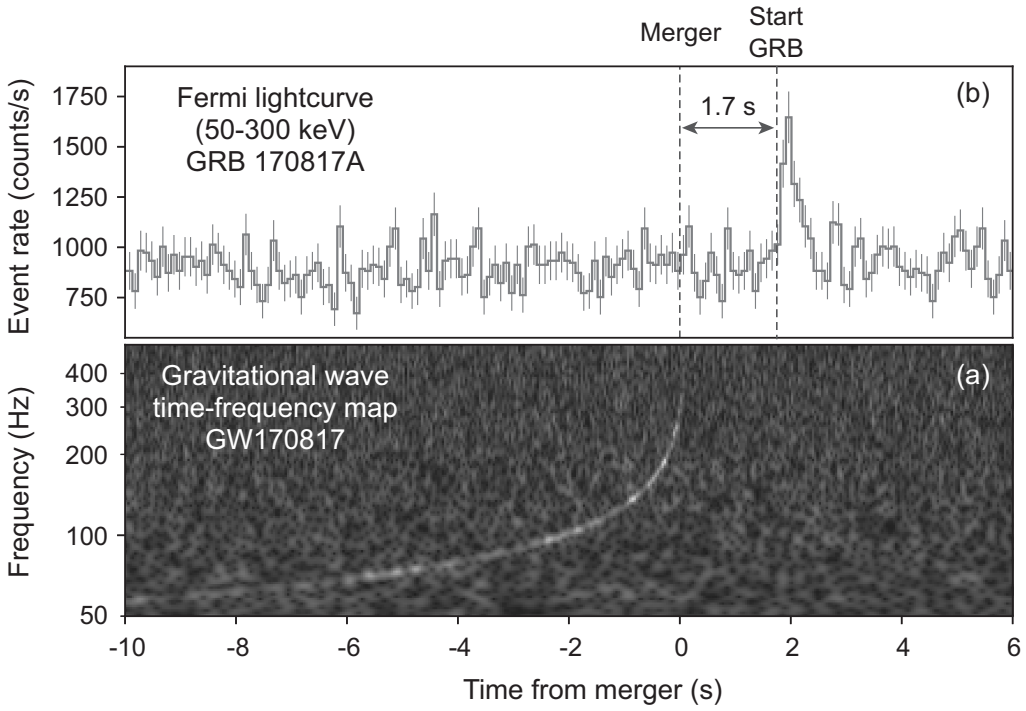


Figure 24.12: (a) Gravitational wave GW170817 (LIGO) and (b) gamma-ray burst GRB 170817A (Fermi satellite). The source was at a luminosity distance of 40 Mpc (130 Mly) and the gravitational wave and gamma-ray burst arrived at Earth separated by only 1.7 seconds.

The gravitational wave GW170817

- was *identified by matched filtering* against post-Newtonian waveform models and
- corresponded to the loudest gravitational wave signal observed to that date, with a *signal to noise ratio of* 32.4.

The coincidence of the gravitational wave and the gamma-ray burst is illustrated in Fig. 24.12.

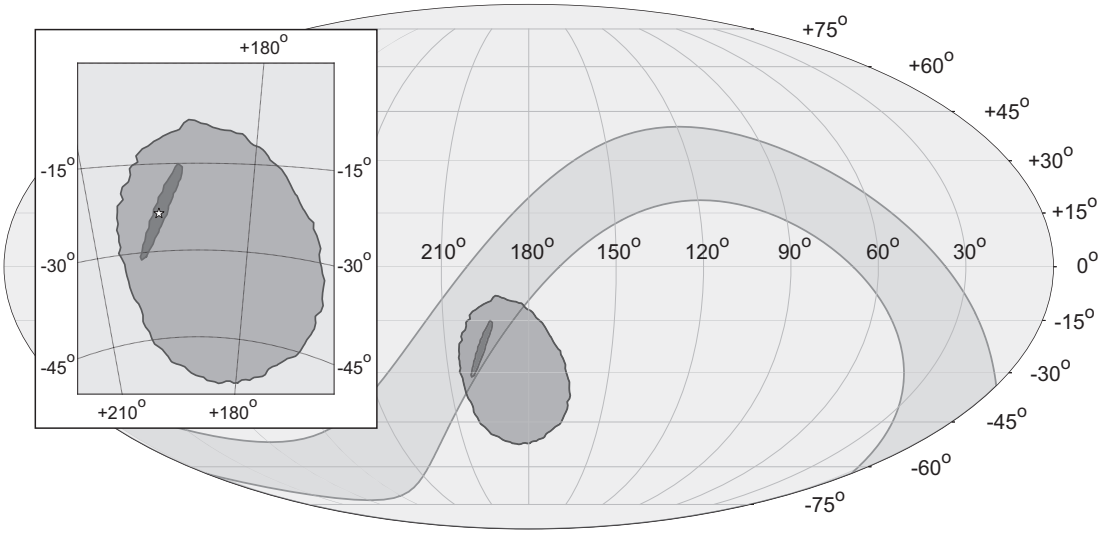
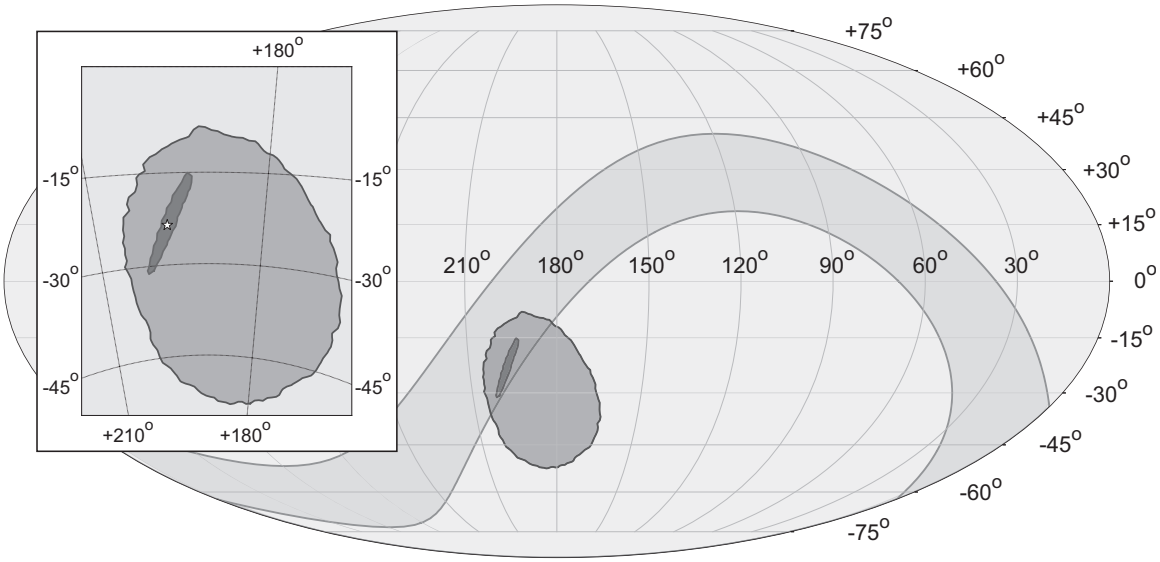


Figure 24.13: Localization of gravitational wave GW170817 and gamma-ray burst GRB 170817A. The 90% contour for LIGO–Virgo localization is shown in the darkest gray. The 90% localization for the gamma-ray burst is shown in intermediate gray. The 90% annulus from triangulation using the difference in GRB arrival time for Fermi and INTEGRAL is the lighter gray band. The zoomed inset shows the location of the transient AT 2017gfo (small white star) that was observed at various wavelengths.

Sky localization of GW170817 is illustrated in Fig. 24.13.

- The final combined LIGO–Virgo sky position localization corresponded to an uncertainty area of $28\,\mathrm{deg}^2$.
- The total mass determined for the binary was between $2.73\,M_\odot$ and $3.29\,M_\odot$, and the two individual masses
- were between $0.86\,M_\odot$ and $2.26\,M_\odot$.

These masses and the waveform indicate that the *compact objects that merged were neutron stars*.



24.6.1 New Discoveries Associated with GW170817

The location of the *afterglow is indicated by the small white star* in the error box of the figure above.

- The *luminosity distance* was 40^{+8}_{-14} Mpc, which is
- consistent with the known distance to the host galaxy.

The multimessenger nature of GW170817 proved to be a *treasure trove of discoveries* having fundamental importance in

- astrophysics,
- the physics of dense matter,
- gravitation, and
- cosmology.

Viability of Multimessenger Astronomy:

The event confirmed that

- gravitational wave detectors could see and distinguish events that did not correspond to merger of two black holes.
- It also demonstrated for the first time that electromagnetic signals could be detected in coincidence with a confirmed gravitational wave event, and
- demonstrated sufficient source localization that the event could be observed at many different wavelengths.

All told, more than 70 facilities observed the event at

- optical,
- radio,
- X-ray,
- gamma-ray,
- infrared, and
- ultraviolet

electromagnetic wavelengths.

Mechanism for Short-Period GRBs:

- The interpretation of the event as the merger of binary neutron stars and
- the coincident (short-period) gamma-ray burst

provided the first conclusive evidence for the hypothesis that short-period gamma-ray bursts are produced in the merger of neutron stars.

- The gamma-ray burst was relatively weak, suggesting that the gamma-ray burst beam was not pointed directly at Earth.
- Confirmation of this hypothesis came two weeks after the initial event when radio waves and X-rays characteristic of a gamma-ray burst were detected.

This evidence taken together represents the first definite association of a gamma-ray burst with a progenitor.

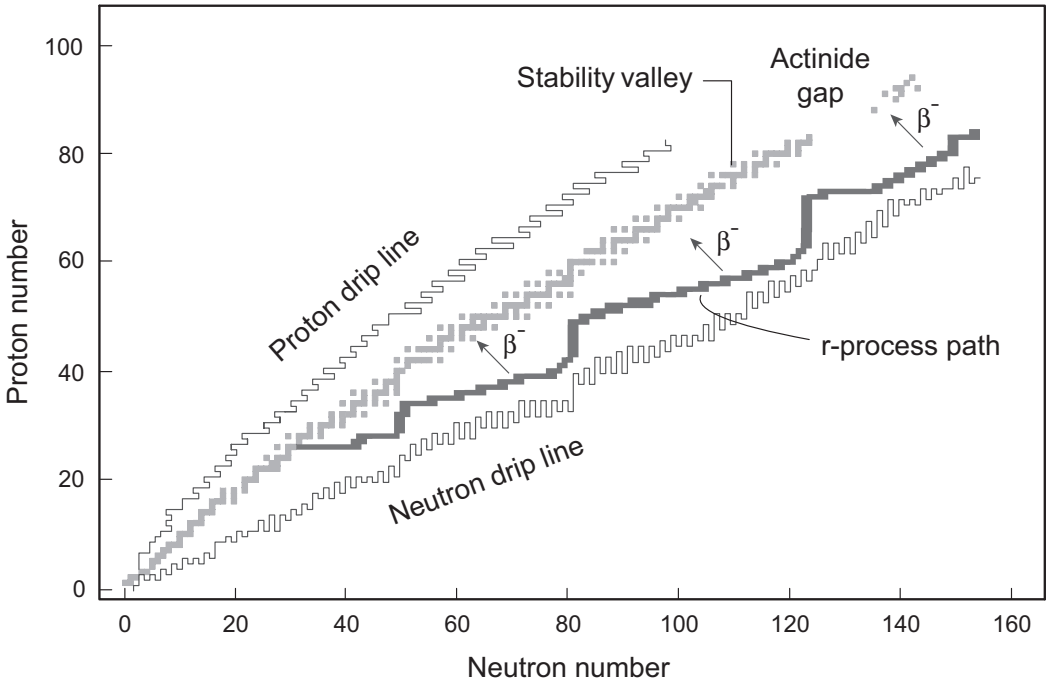
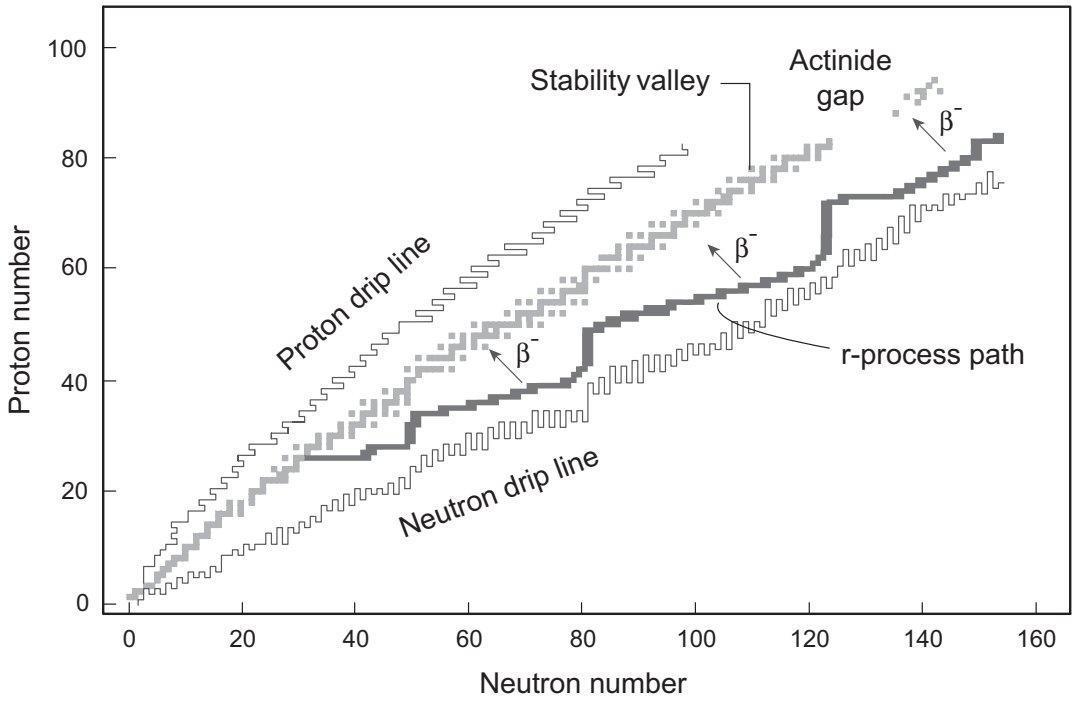


Figure 24.14: Theoretical path for the r-process. Nuclei produced along the r-process path will undergo rapid β^- decay back toward the stability valley, thus producing most of the neutron-rich and some of the β-stable isotopes, as well as all the actinide nuclei found in nature. (The β-stable isotopes beyond iron but below the actinide gap can be produced also in the slow neutron capture or s-process in red giant stars.) The two drip lines denote the boundaries beyond which a nucleus becomes unstable against spontaneous emission of neutrons or protons, respectively.

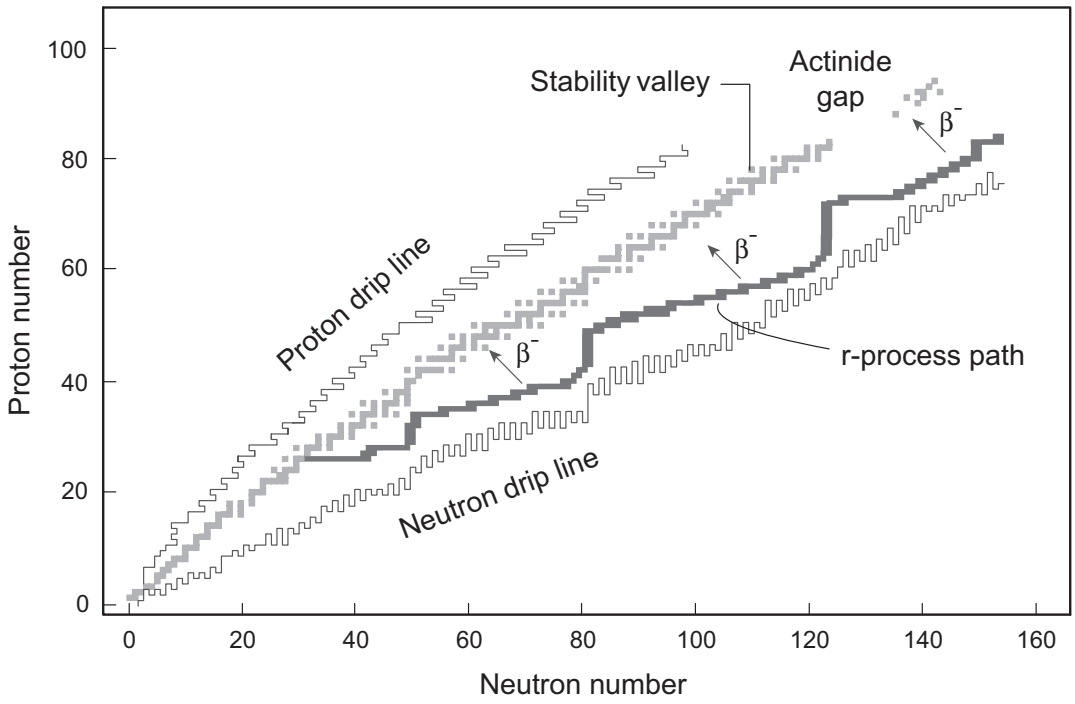
Site of the r-Process:

The signature of heavy-element production in the event

- demonstrated that *neutron star mergers are one (perhaps the dominant) source* of the
- *rapid neutron capture or r-process* thought to make many of the heavy elements (see Fig. 24.14).



- Now we have a quantitative way to investigate the relative importance of the *two primary candidate sites for the r-process:*
  - core collapse supernovae, and
  - neutron star mergers.
- Already it is clear that the dominant attitude of not very long ago that the r-process was associated mostly with core-collapse supernovae *is probably not correct*.



One common theme for understanding the origin of r-process nuclei is to ask whether they were produced

- *in a few rare events* (neutron star mergers occur maybe only once every million years in a large galaxy), or
- *in many much more common events* (core collapse supernovae occur about once every 50 years in a large galaxy).

Some evidence had been accumulating that at least some r-process nuclei were produced in rare events.

The neutron star merger leading to GW170817 gives direct evidence for *significant production of r-process nuclei in a single rare event*.

Observation of a Kilonova:

The expanding radioactive debris in GW170817 was observed at UV, optical, and IR wavelengths.

- This gave the first direct evidence for the *kilonova* (also termed a *macronova*)
- that is predicted to occur following such mergers as a result of *radioactive heating by newly-synthesized r-process nuclei*.
- The direct nucleosynthesis of r-process species likely ceases after a second or two,
- but most initially-synthesized isotopes would be *highly radioactive* and
- the cloud of debris can be *kept warm* ($10^3 - 10^4$ K) *by radioactive decay for as long as weeks*.

That the *gamma-ray burst was emitted off-axis* relative to Earth may have been essential in allowing the kilonova to be observed.

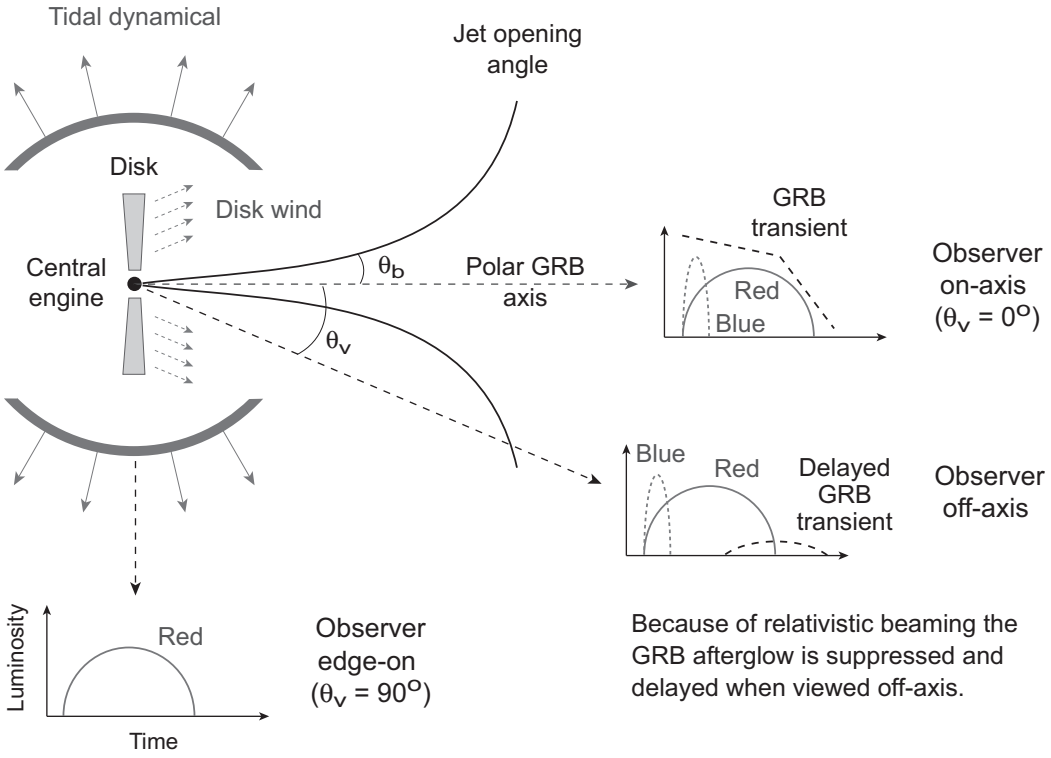
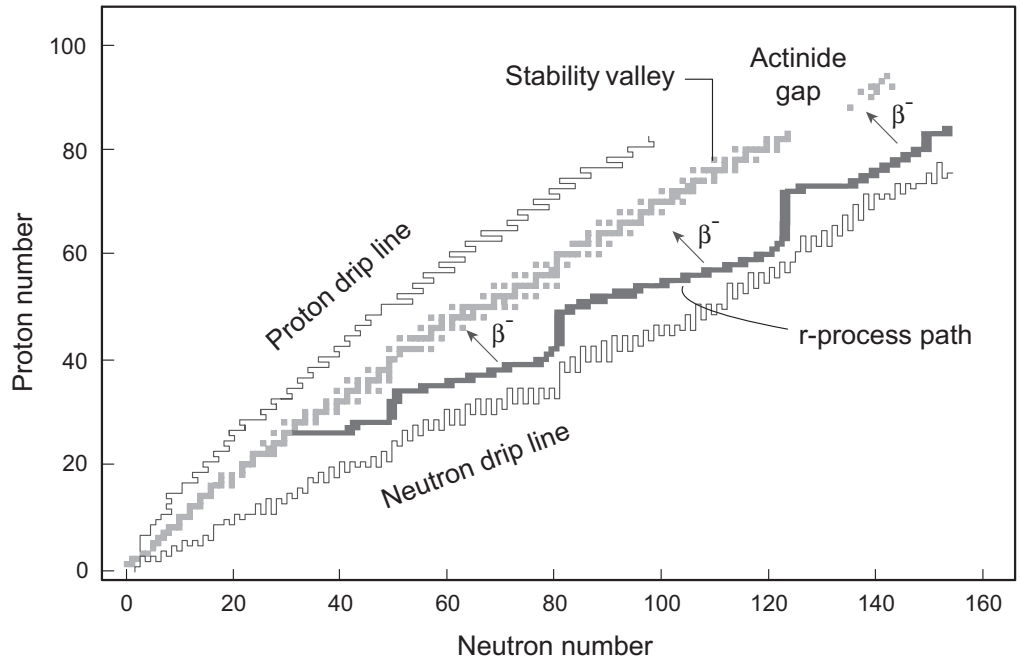


Figure 24.15: Geometry of GW170817 afterglows. Neutron-rich ejected matter labeled "Tidal dynamical" emits a kilonova peaking in the IR (solid arrows and solid curves labeled "Red" in the time–luminosity diagrams) associated with production of heavy r-process nuclei and high opacity (the *red kilonova*). Additional mass is emitted by winds along the polar axis (dotted arrows and dotted curves labeled "Blue") that is processed by neutrinos emitted from the hot central engine, giving matter less rich in neutrons and a kilonova peaking in the optical that is associated with production of light r-process nuclides and lower opacity (the *blue kilonova*).

As illustrated in Fig. 24.15, if the GRB is seen nearly on-axis, the GRB afterglow masks the kilonova.

- The usual GRB afterglow is indicated by dashed curves.
- It dominates on-axis but viewed off-axis it appears as a low-luminosity component delayed by days or weeks

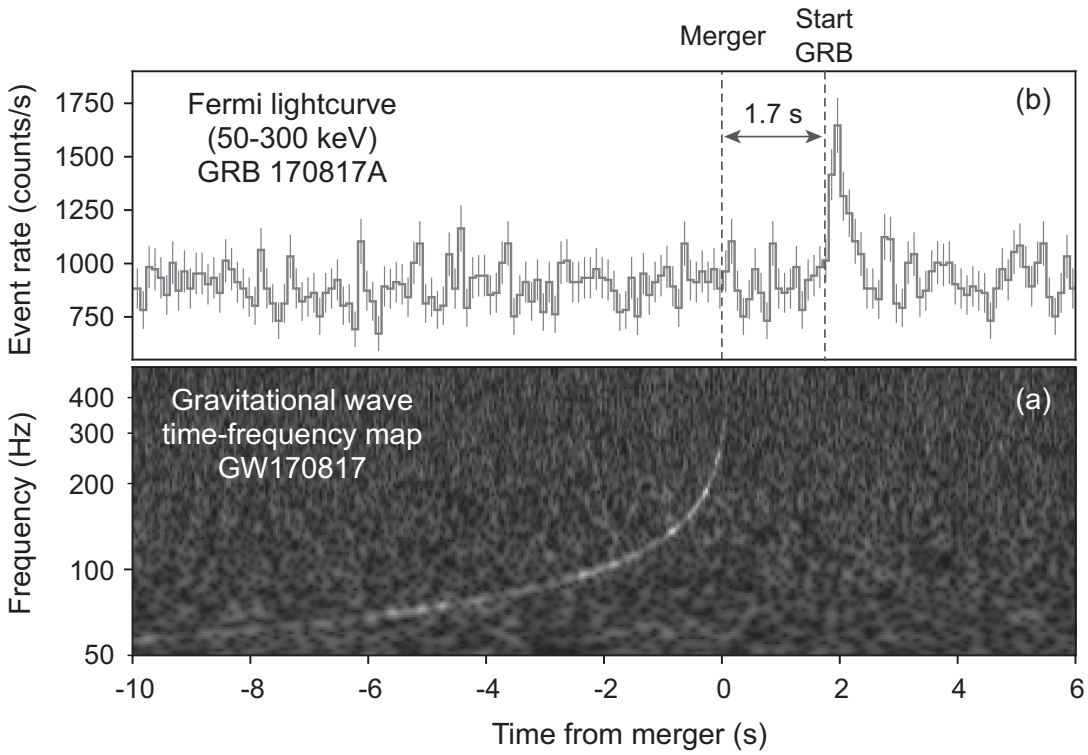


Nuclei Far from Stability:

The r-process runs far to the right (neutron-rich) side of the β-stability valley in the chart of the isotopes shown above

- Little definitive information exists here because the isotopes cannot be made in traditional accelerators.
- Kilonova lightcurves are a statistical mix of contributions from many neutron-rich nuclei with *no sharp lines*
- because of the *high velocities* ($v \sim 0.3c$) *for the ejecta*.
- However, they carry information about the *average decay rates and other general properties* of these largely unknown r-process nuclei.

This can give nuclear structure constraints far from stability.



The Speed of Gravity:

The GRB arrived within 1.7 seconds of the gravitational wave from a distance of 40 Mpc (figure above).

- This established conclusively that the difference between
  - the *speed of gravity* and
  - the *speed of light*

  is no larger than 3 parts in 10^{15}.

Thus it took 1.7 seconds of observation to *eliminate from contention* alternatives to general relativity for which gravity does not propagate at c.

Neutron-Star Equation of State:

The multimessenger nature of GW170817 indicates that neutron star mergers will

- provide an opportunity to make much more *precise statements about the neutron-star equation of state*.
- For example, the merger wave signature is *sensitive to the tidal deformability* of the neutron star matter near merger.
- This is of fundamental importance for *understanding dense matter* because
- prior observations have been unable to constrain candidate equations of state sufficiently to understand (for example)
  - the *maximum mass of a neutron star* and
  - the *minimum mass of a black hole*

  to better than an *uncertainty of about a solar mass*.

Demographics of Neutron-Star Binaries:

The observation of GW170817 provides quantitative information about the probability that neutron star binaries form in orbits that can lead to *merger in a Hubble time* .

- This probability has been *rather uncertain* to this point.
- The rate currently inferred corresponds to 0.8×10^{-5} mergers per year in a galaxy the size of the Milky way.

An accurate determination of the merger rate has implications for

- our *understanding of stellar evolution*,
- the *site of the r-process*, and
- the expected *rate of gravitational wave detection* from such events.

Determination of the Hubble Constant:

The multimessenger nature of the GW170817 event provides an *independent way to determine the Hubble constant* H_0.

- This can be accomplished by comparing the
  - distance inferred from the gravitational wave with
  - the redshift of the electromagnetic signal.
- Presently, different methods of determining the Hubble constant yield a value typically in the range

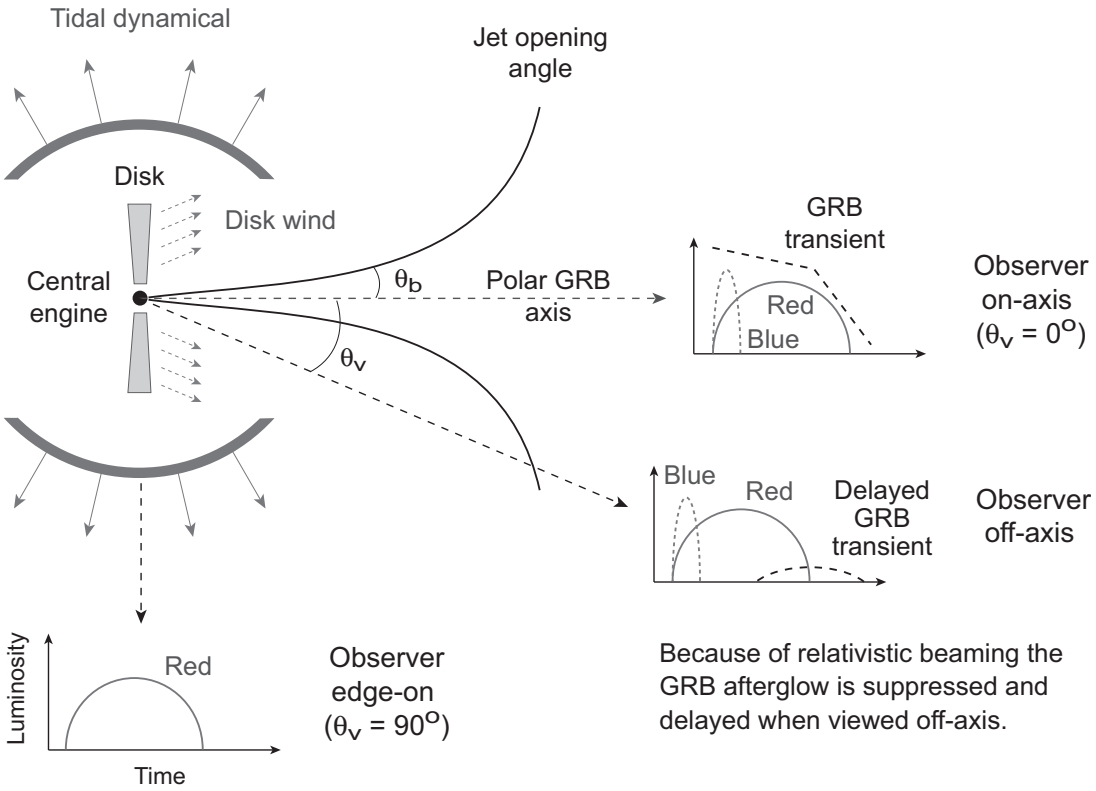
$$H_0 \sim 67 - 73\,\mathrm{km\,s^{-1}Mpc^{-1}},$$

  - with *analyses of the CMB* giving *values nearer the lower end* of this range and
  - traditional *"distance-ladder" methods* like Cepheid variables giving *values nearer the higher end*.
- Analysis of GW170817 suggests a value in the middle with large uncertainties,

$$H_0 \sim 70^{+12}_{-8}\,\mathrm{km\,s^{-1}Mpc^{-1}}.$$

- We may expect an accumulation of such multimessenger events to yield a *precise, independent determination of* H_0.

For example, 100 independent GW detections with host galaxy identified as in GW170817 could *determine* H_0 *with an uncertainty of 5%*.



Off-Axis Gamma-Ray Bursts:

The initial observation of the kilonova followed days later by observation of X-ray and radio emission

- provides strong corroborating evidence for the *beamed nature of gamma-ray bursts* and
- represents the first clear detection of a *weak, off-axis GRB* and its slowing in the interstellar medium.

Systematic studies of such events should greatly enrich our understanding of gamma-ray bursts.

We have insufficient space to elaborate on all these topics, but let' look at one in more detail: the *kilonova* powered by the *production of radioactive r-process nuclei*.

24.6.2 The Kilonova

Simulations of neutron star mergers identify two mechanisms for mass ejection:

1. Matter may be
   - *expelled dynamically by tidal forces on millisecond timescales* during the merger itself, and
   - as surfaces come into contact *shock heating at the surfaces* may squeeze matter into the polar regions.
2. On a longer (~ 1 s) timescale matter in an accretion disk around the merged objects can be *blown away by winds*.

As ejected matter decompresses, heavy elements are made by rapid neutron capture.

- If the matter is highly neutron-rich, repeated neutron captures form the *heavy r-process nuclei* ($58 \leq Z \leq 90$);
- if the ejecta is less neutron-rich, *light r-process nuclei* ($28 \leq Z \leq 58$) are synthesized.
- Matter ejected in the tidal tails is cold and very neutron rich, and tends to form *heavy r-process nuclei*.
- The disk winds and ejecta squeezed into the polar regions
  - are *irradiated by neutrinos* from the central region, which *converts some neutrons to protons*.
  - This *favors the light r-process*.

The *photon opacity of the r-process ejecta* may play a leading role in the observable characteristics of kilonova events.

- The *photon opacity* is generated largely by transitions between bound atomic states (*bound–bound transitions*).
- For light r-process nuclei the *valence electrons typically fill atomic d shells*.
- In contrast a substantial fraction of heavy r-process species produced by simulations (often 1–10% by mass) are *lanthanides* ($58 \leq Z \leq 71$).
- For lanthanides the *valence electrons fill the f shells*.
- These have
  - *densely-spaced energy levels* and
  - an *order of magnitude more line transitions*

  than for the d shells in light r-process species.
- As a consequence,

  The *opacity of heavy r-process nuclei*

  - is *roughly a factor of 10 larger* than the opacities for light r-process species, and
  - they have correspondingly *long photon diffusion times*.

Hence the *cloud of light r-process species*

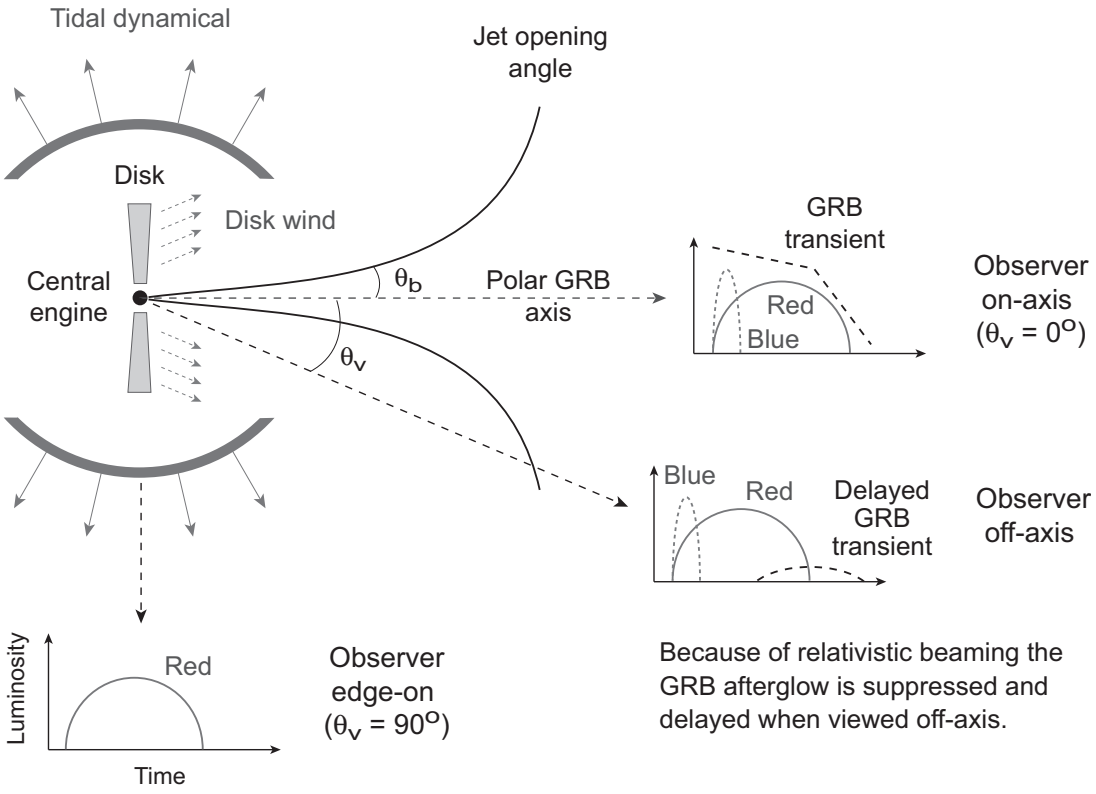
- is considerably *less opaque*
- has *shorter diffusion times*, and
- *tends to radiate in the optical* and fade over a matter of days.

In contrast, the *cloud of heavy r-process species*

- radiates *in the IR*
- *for as long as weeks* because of the
  - *high opacity* and
  - long *photon diffusion times*.

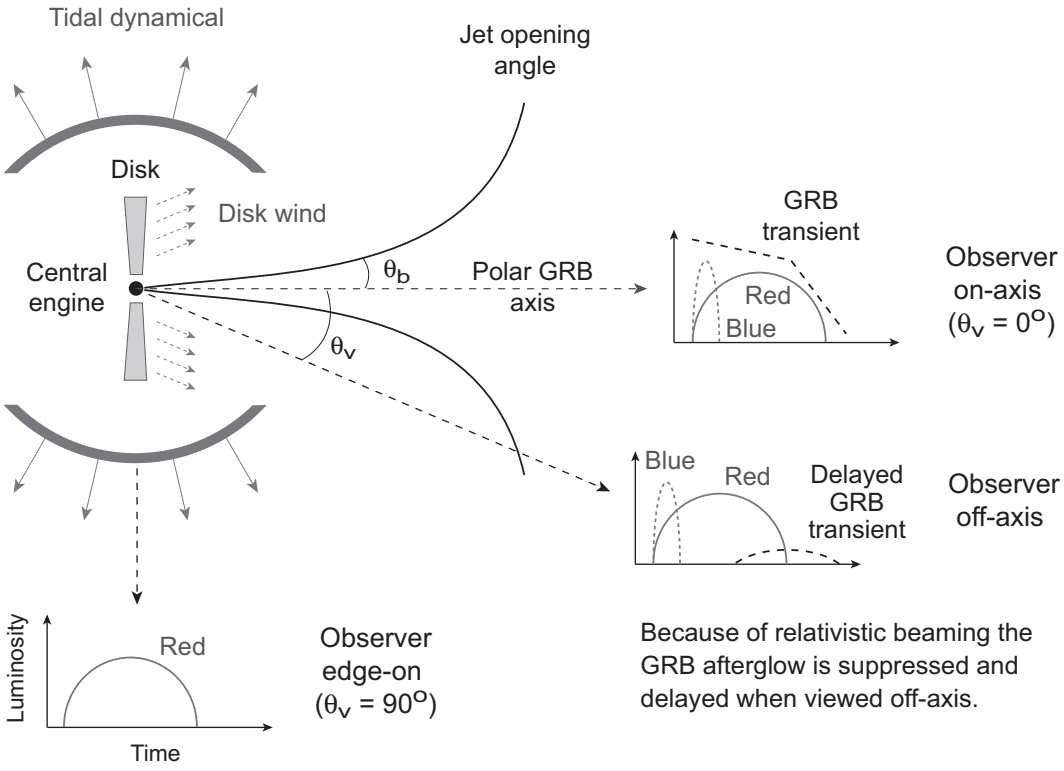
This accounts for *observed characteristics of the transient* AT 2017gfo, which *differed essentially* from normal astrophysical transients:

- It *brightened and faded quickly in the optical*, but
- quickly-growing *IR emission* was strong for many days, and
- only after a period of days or weeks did *X-ray and RF signals* begin to emerge.



The preceding considerations suggest a general picture of the *geometry of* GW170817 that is illustrated in the figure above.

- The kilonova transient AT 2017gfo that followed the gravitational wave GW170817 and associated gamma-ray burst GRB 170817A had *two distinct components*.
- First, *tidal dynamical ejection* flung out on ms timescales very neutron-rich matter at high velocities $v \sim 0.3c$.
  - This matter underwent *extensive neutron capture* to produce *heavy r-process species*.
  - It had *extremely high opacity* because of the lanthanide content.



- Second, *winds ejected matter from the disk region* on a timescale of seconds.
  - This matter was subject to
    - *shock heating* and to
    - *irradiation by neutrinos* from the hot center.
  - Both tended to *decrease the neutron to proton ratio*.
- Nucleosynthesis in this *less neutron-rich matter*
  - was likely to produce *light r-process matter of lower opacity*, since
  - there *weren't enough neutrons* to produce lanthanides and other heavy r-process nuclei.

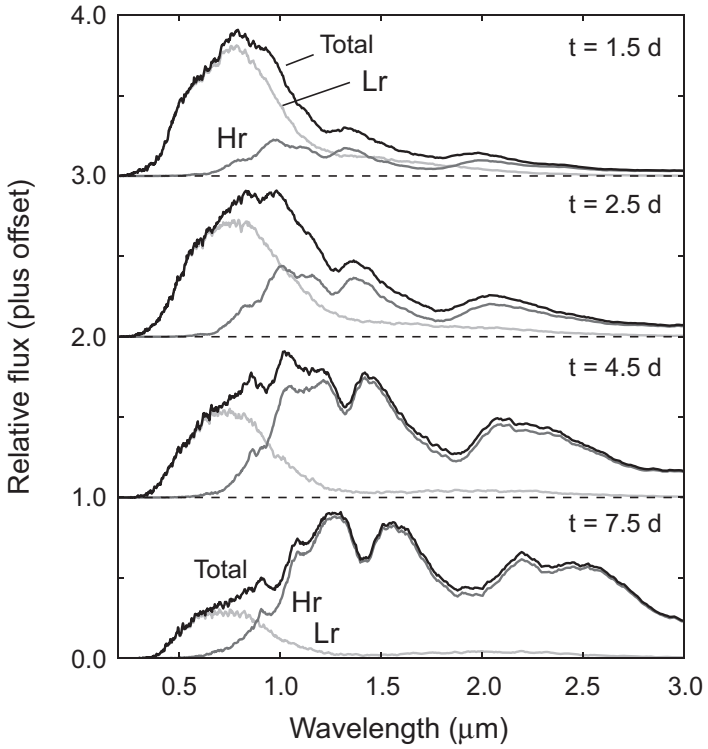
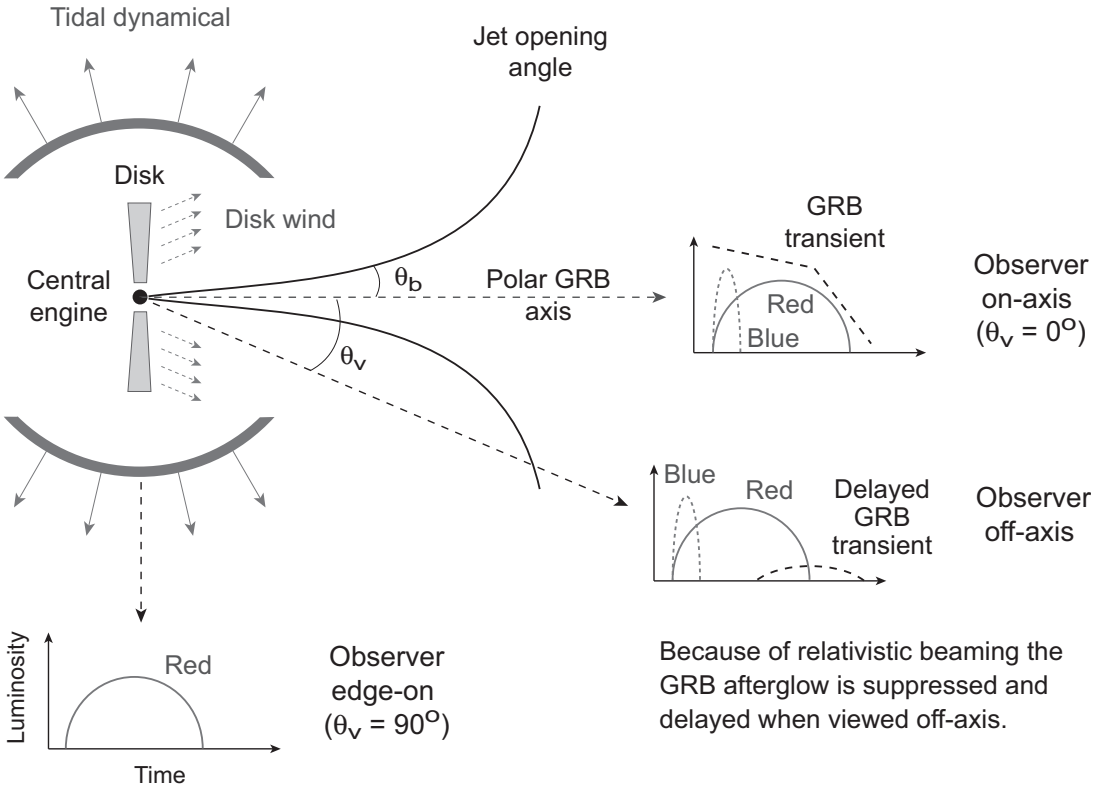


Figure 24.16: Evolution of components of the GW170817 kilonova. Flux is a sum of two spatially-separated components: dominantly-optical emission from light r-process isotopes ("blue kilonova", labeled Lr) and dominantly-IR emission from heavy r-process isotopes ("red kilonova", labeled Hr).

This picture is supported by model simulations of Fig. 24.16.

- These simulations exhibit clearly early emergence and rapid decay of the *optical component* associated with the light r-process (*blue kilonova, labeled Lr*).
- This is followed by the longer-lived *IR component* associated with the heavy-r process (*red kilonova, labeled Hr*),
- which grows within days to dominate the lightcurve.



- The general features of this composite model *resemble the observed time evolution of* AT 2017gfo.
- The *red and blue kilonova components* were visible only because
  - the *GRB was seen off-axis*, which
  - suppressed the GRB afterglow by *relativistic beaming* so that
  - the underlying kilonova was visible (see figure above).

24.7 Gravitational Waves and Stellar Evolution

The interpretation of GW150914 as a black hole merger poses some *challenging questions* for theories of stellar evolution.

> For example, how does a binary composed of two $30M_\odot$ black holes even form?

Presumably either

- A binary formed with two stars of large mass and *survived successive core collapses* for each star, or
- The black holes *formed independently* in a dense cluster and then were *captured by gravity* into a binary orbit.

Neither is easy to realize without untested assumptions.

> Thus, future gravitational waves detected from
>
> - merger of two black holes,
> - merger of a black hole and neutron star,
> - merger of two neutron stars, and
> - core collapse supernovae
>
> will illuminate—and pose challenges to—our understanding of *stellar evolution for massive stars*.

Comprehensive simulations indicate that the binary black holes responsible for the gravitational waves observed thus far

- could have formed in isolated binary star evolution,
- provided that they formed in regions having low concentration of elements heavier than helium (regions of *low metallicity*).

24.7.1 Metallicity and Stellar Mass

Major factors limiting stellar masses are *(1) cooling of gas clouds* collapsing to stars and *(2) stability with respect to mass ejection* caused by radiation pressure.

- Radiation pressure *grows rapidly with star mass* since
  - it *varies as* T^4 and
  - more massive stars are *hotter*.
- *Photon opacity* has a major influence on how massive a star can become, and
- *opacity is greatly increased by the presence of metals* (elements with atomic number greater than two) because
  - they *produce many electrons* when ionized and
  - *photons scatter strongly* from free electrons.
- Generally, *low metal content* (low *metallicity*) favors the birth of *more massive stars:*
  - *Metals aid in cooling gas clouds* collapsing to stars,
  - so *low metallicity implies higher temperatures* and
  - these higher temperatures favor collapse of more massive gas clouds (this is called the *Jeans criterion*).
  - Then *reduced wind mass loss because of low opacity* keeps these nascent stars from expelling mass,
  - so they can *grow even more massive by accretion*.

The big bang *produced almost no metals*.

- Thus metallicities and hence opacities in the first generation of stars (called *Population III* or *Pop III*) were very low.
- Simulations indicate that typical Pop III stars could have had masses of *hundreds to thousands of solar masses*.
- These could have produced black holes of comparable mass when their cores collapsed at the end of stellar evolution.

Subsequent generations of stars formed from material enriched in metals produced in supernova explosions of earlier stars.

- They had *increasingly larger metal content*.
- In today's more metal-rich Universe *few stars grow more massive than* 50–100 $M_\odot$.

24.7.2 A Possible Evolutionary Scenario for GW150914

Figure 24.17 (following page) illustrates a possible evolutionary scenario for the production of GW150914.

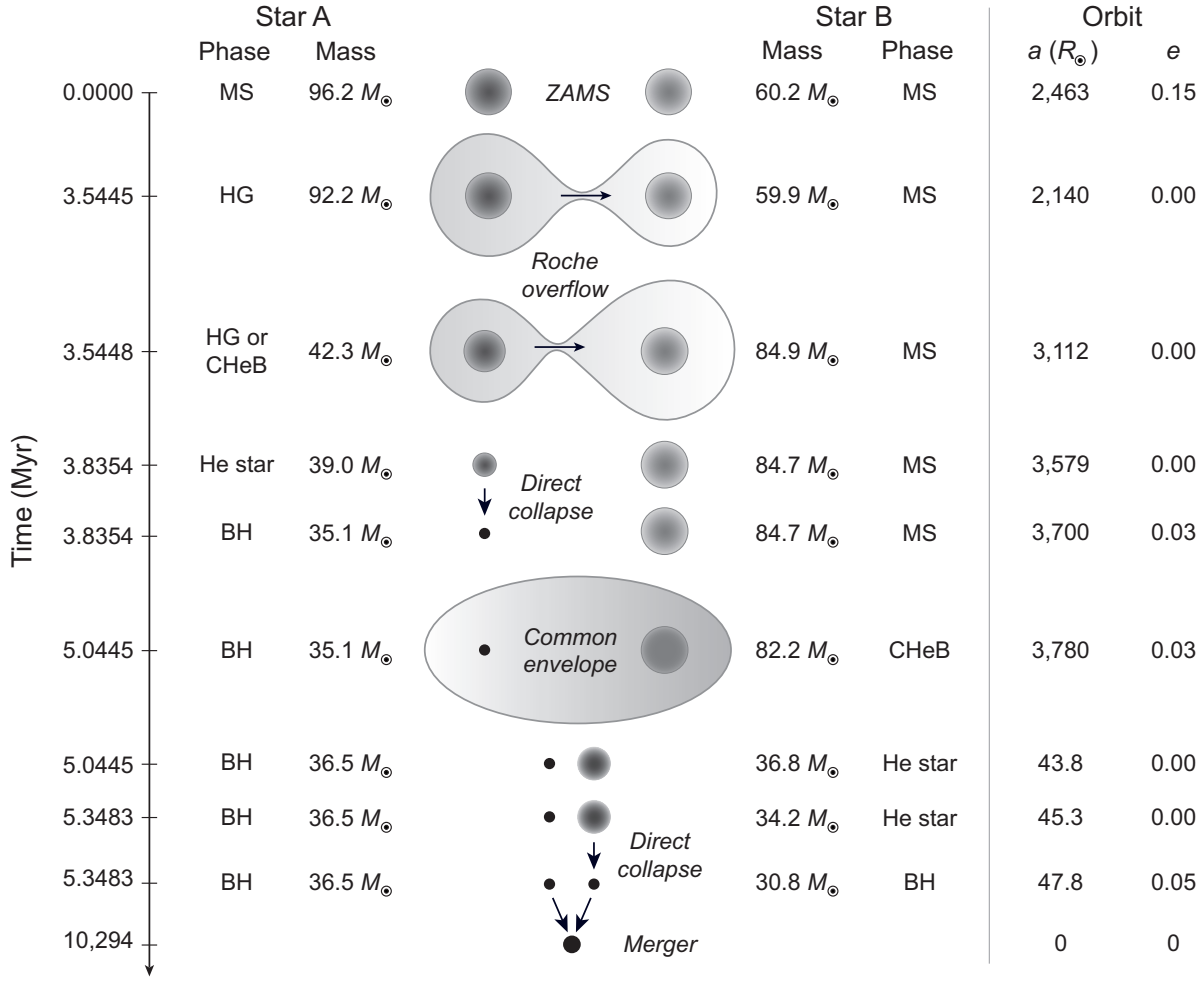
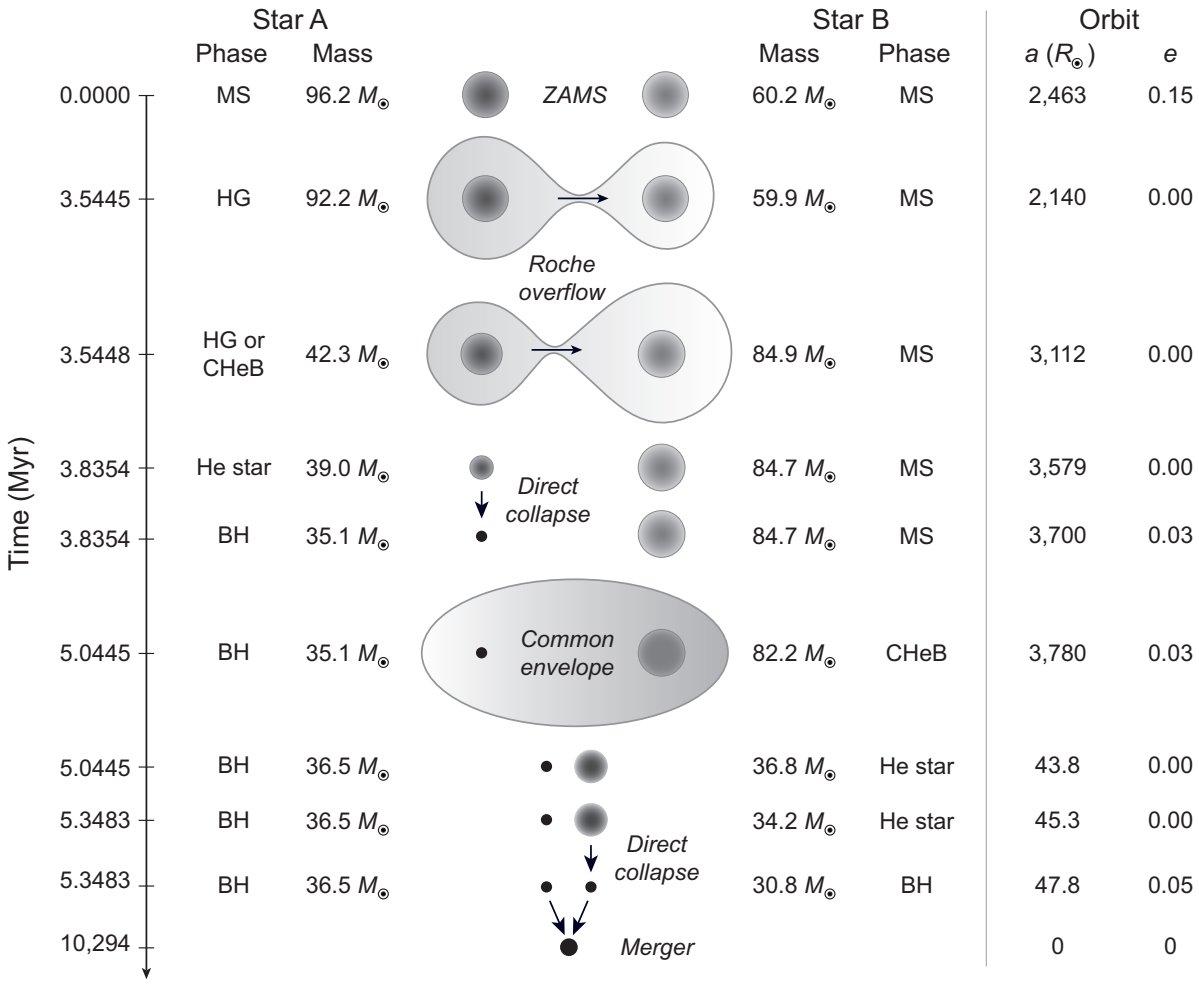
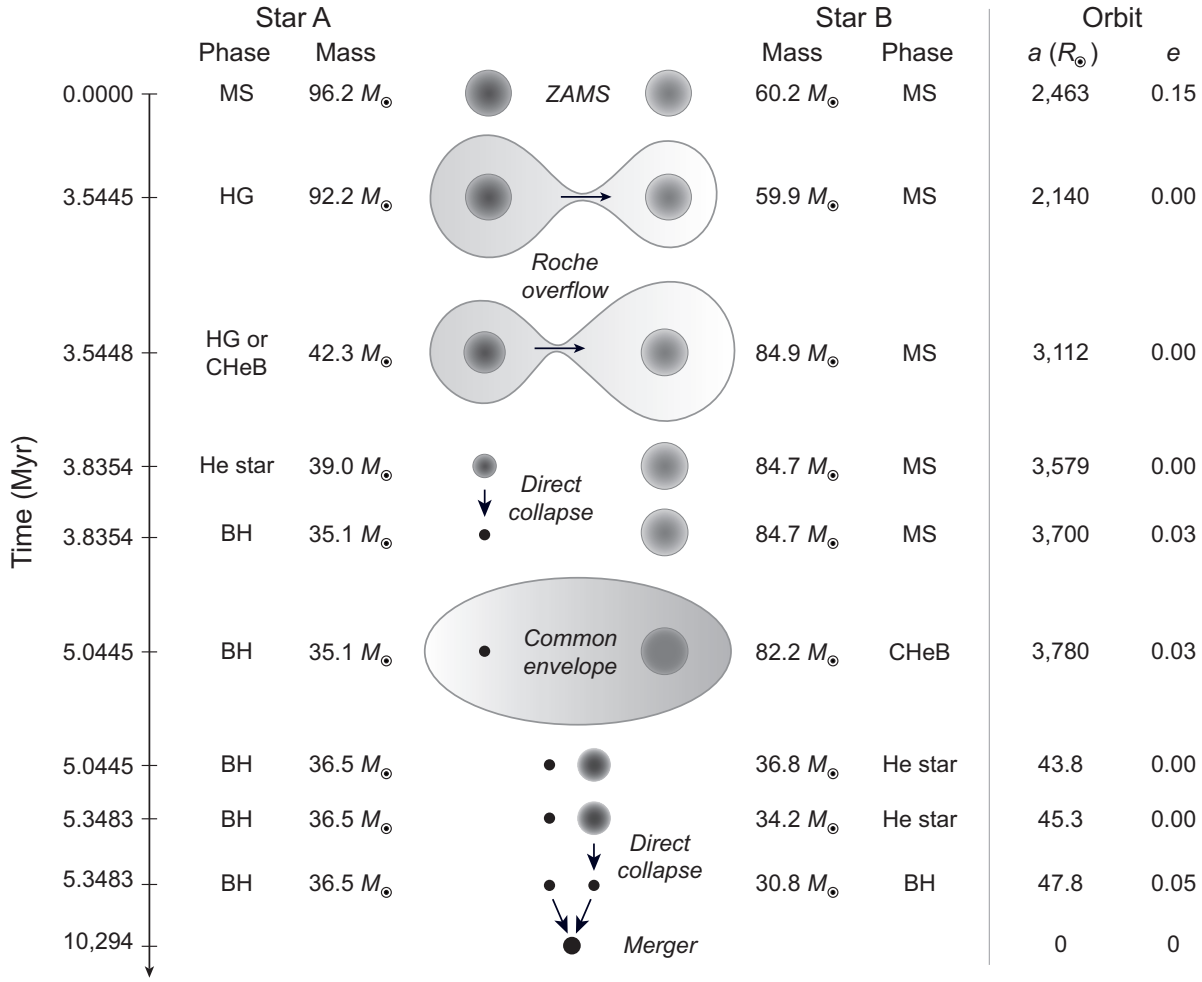


Figure 24.17: A scenario for evolution of the massive black hole binary leading to GW150914. ZAMS means zero age main sequence (the time when the star first enters the main sequence), MS means main sequence, HG means a star evolving through the Hertzsprung gap (the evolutionary region between the main sequence and the red giant branch), CHeB means core helium burning, a He star is a star exhibiting strong He and weak H lines (indicating loss of much of its outer envelope), and BH indicates a black hole. Time is measured from formation of the binary and the scale is *highly nonlinear*. The separation of the pair is a and the eccentricity of the orbit is e.



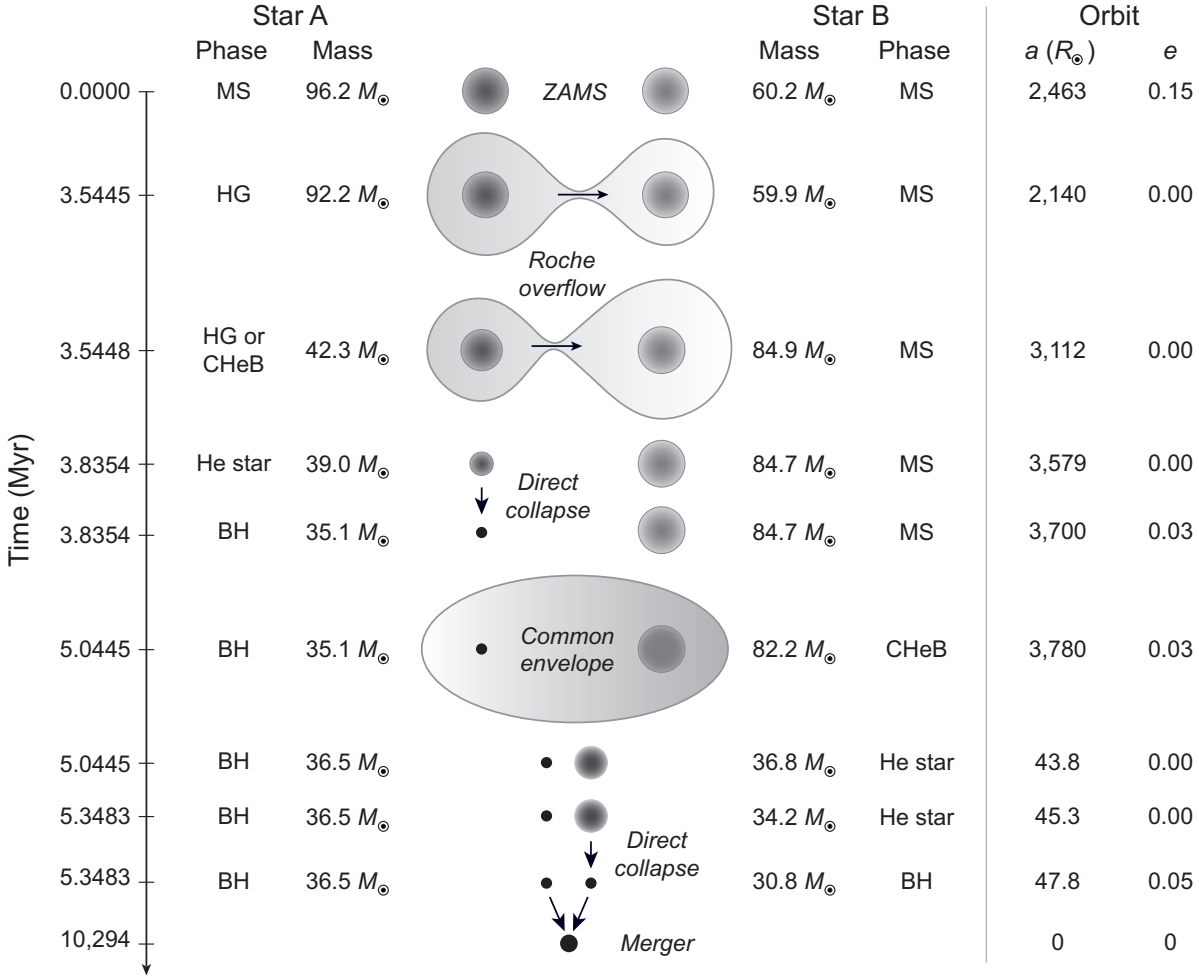
A massive binary formed about 2 billion years after the big bang (redshift $z \sim 3.2$), with

- initial masses of $96.2\,M_\odot$ (star A) and $60.2\,M_\odot$ (star B),
- a metal fraction Z that was 0.03 times that of the Sun,
- an average separation $a \sim 2500\,R_\odot$, and
- an orbital eccentricity $e = 0.15$.



Star A evolved quickly, expanded, and *transferred more than half of its mass to star B* by Roche lobe overflow, as star A evolved through the Hertzsprung gap to core helium burning.

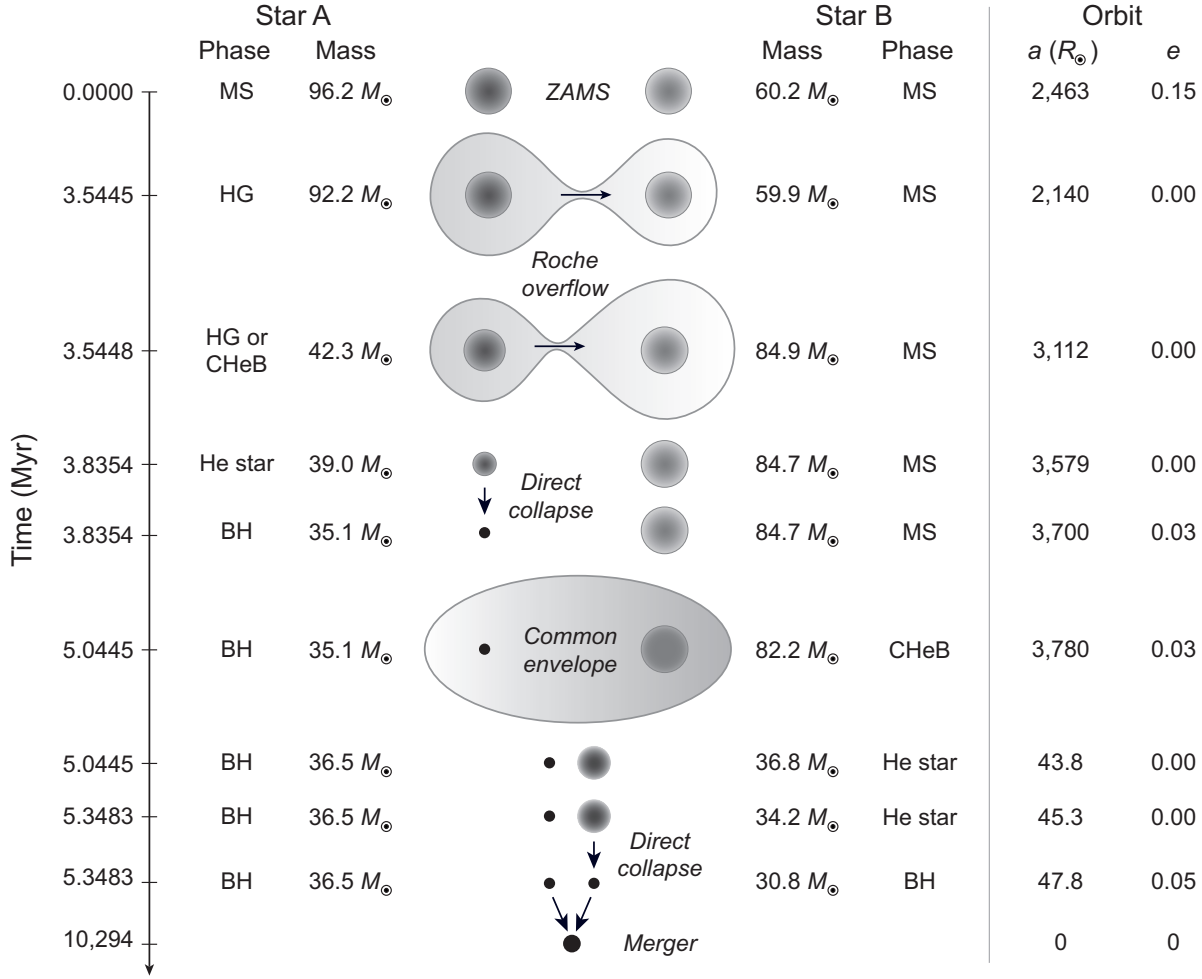
- Star A then *collapsed directly to a* $35.1\,M_\odot$ *black hole*, with no ejection of a supernova remnant, but with *10% of the mass carried off by neutrinos* during the collapse.
- When the first black hole had formed, star B had accreted to $84.7\,M_\odot$ and evolved to core helium burning.



The expansion of star B initiated a *common envelope (CE) phase* with the black hole that formed from star A.

- During the CE phase the average separation of the binary components was reduced from $a \sim 3800\,R_\odot$ to $a \sim 45\,R_\odot$.
- At the end of the CE phase the mass of the black hole formed from star A was $36.5\,M_\odot$ and star B was now a helium star of mass $36.8\,M_\odot$.

Star B then *collapsed directly to a black hole*.



- This left a *binary black hole system* with masses of $36.5\,M_\odot$ and $30.8\,M_\odot$, respectively, and
- orbital separation $a = 47.8\,R_\odot$.
- This system then *spiraled together through gravitational wave emission* for 10.3 billion years,
- and merged about 1.1 billion years ago ($z \sim 0.09$) to produce GW150914.

The simulations described above paint a compelling picture but they entail large uncertainties because of

- *assumptions such as the value of metallicity*, and because
- *accretion* and (in particular) *common envelope evolution* are the *poorly understood aspects of binary evolution*.

Tests of these assumptions and increasingly strong constraints on models of massive binary star evolution may be expected as gravitational wave astronomy matures.

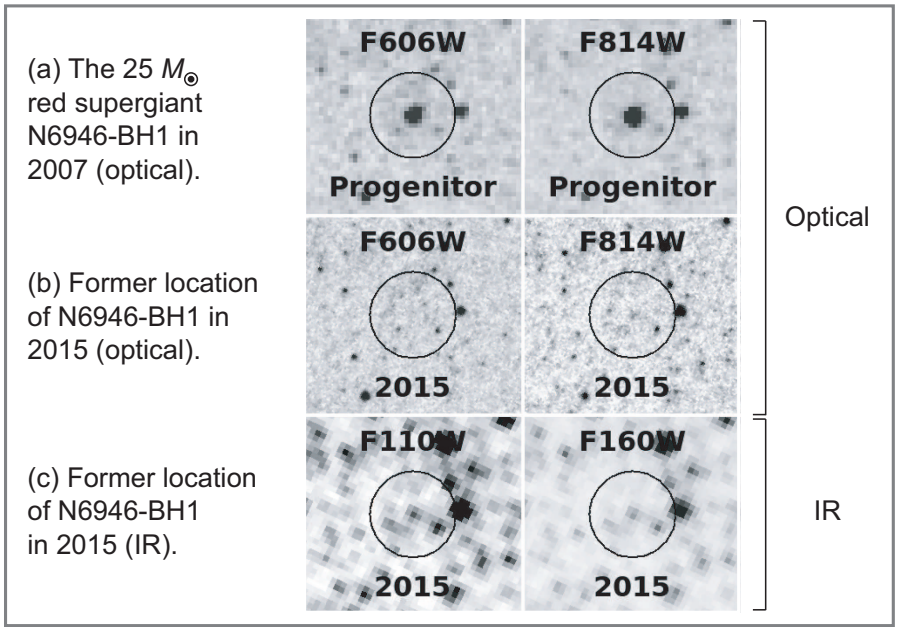
One crucial feature of the mechanism outlined above is *direct collapse of massive stars* in a binary to black holes,

- *without ejecting a supernova remnant* and
- without giving a strong *natal kick* to the black hole that is formed, so that it remains in the binary.

The following page gives direct observational evidence that such *failed supernovae* may occur in nature.

24.7.3 Direct Collapse of Massive Stars to Black Holes

A *failed supernova candidate* is shown in the following figure.



- In 2007 the $25\,M_\odot$ *red supergiant* N6946-BH1 was bright in optical images [center of circles in panel (a)].
- In 2009 this star underwent a *weak optical outburst* and then *faded rapidly* over a matter of months.
- In 2015 N6946-BH1 had *disappeared in the optical* [panel (b)] but a *faint IR signal remained* [panel (c)].

These results suggest *birth of a black hole in a failed supernova*, with *IR emission from weak accretion* on the black hole.

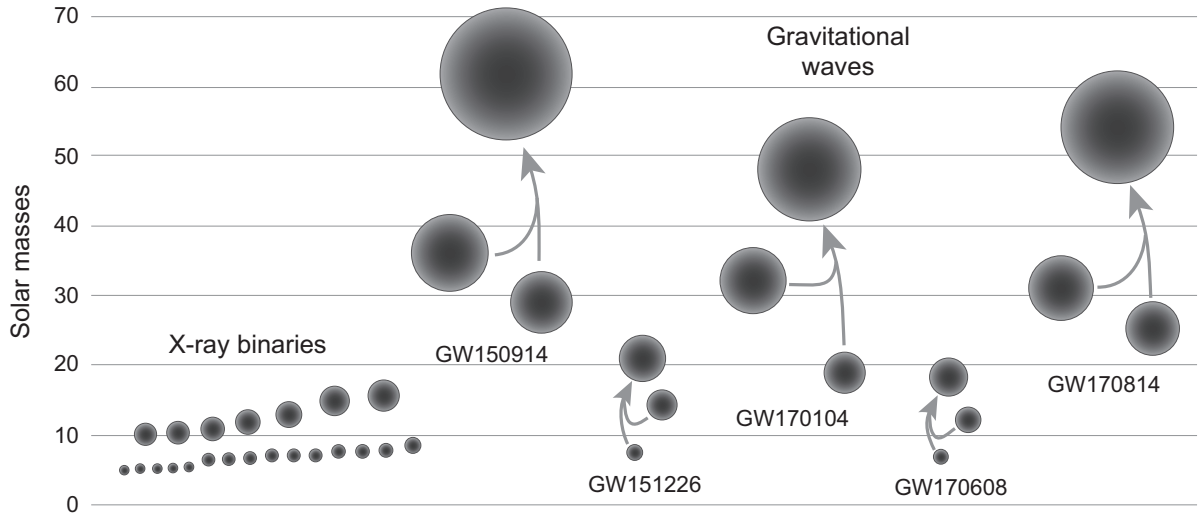


Figure 24.18: A summary of black hole masses determined from X-ray binary and gravitational wave data as of November, 2017. Arrows indicate black hole mergers.

24.7.4 Measured Stellar Black Hole Masses

Black holes are one conjectured endpoint for stellar evolution. We have discussed *two reliable methods* of identifying stellar black holes and determining their masses:

- *Mass-function analysis* of X-ray binary systems, and
- *Analysis of gravitational waves* from binary black hole mergers discussed in this chapter.

Fig. 24.18 summarizes masses for 37 black holes determined from these two types of analysis.

24.7.5 Are Stellar and Supermassive Black Holes Related?

Is there is a connection between the formation of stellar-mass black holes and of supermassive black holes? *Two extremes have been discussed*.

- The *light seeds model,* in which supermassive black holes formed by successive merger of black holes created by core collapse of stars.
- The *heavy seeds model,* in which $10,000 - 100,000 M_\odot$ seeds for supermassive black holes may have formed by
  - *direct collapse of gas clouds*, or possibly from
  - rapid merger of *black holes from massive first-generation (Pop III) stars*.

As of 2019, the *jury is still out* on this issue.

> Thus, it is possible that evolution of massive stars ending in the creation of stellar black holes also has implications for supermassive black holes.

Part V

General Relativity and Beyond

Chapter 25

Tests of General Relativity

This final Part concerns *the future of gravitational theory*. We begin with an analysis of how well GR has stood the test.

- This chapter gathers in one place a list of some *observational and experimental tests of general relativity*. It is relatively brief because
  - the list of observables for which general relativity predicts meaningful differences with Newtonian gravity is itself rather brief, and
  - many of these tests already have been discussed in preceding chapters.
- Our primary concern will be *comparison of general relativity with observation and experiment*.
- However, it is important also to *compare the predictions of GR with alternative theories of gravity*.

25.1 Alternative Theories of Gravity

Many alternatives to general relativity as the modern theory of gravity have been proposed. Two motivations have assumed increasing importance in recent years:

- The *incompatibility of general relativity with quantum mechanics*, suggesting that general relativity is not complete.
- The *distaste of some for the mysterious nature of dark matter and dark energy* in the standard cosmology,
- which has motivated a view that they may be *fictions masking a failure of current gravitational theory*.

One of the best-known example of an alternative to general relativity is the *scalar–tensor theory of Brans and Dicke*.

Alternatives and Observations:

Most alternatives to general relativity are *ruled out quickly by observations*.

- For example, they may fail to predict the correct deflection of light in a gravitational field.
- However, some alternatives are not obviously falsified by current observations.
- Typically in these still-viable theories
  - the gravitational force *derives from the spacetime metric*, as for general relativity, but
  - they differ in that *additional fields are added*.
  - For example, a long-range scalar (spin-0) field may supplement the spin-2 tensor field of GR.

For such theories to be consistent with the observed properties of gravity *the relationship of the additional fields to matter is one-way*:

- Matter may create the fields and thereby make contribution to the curvature of the spacetime metric, but
- The new fields don't act back on the matter because *matter is assumed to respond only to the metric*.

Tests with Gravitational Waves:

One distinction between general relativity and alternatives may be in *gravitational wave emission*.

- For example, in scalar–tensor theories *scalar and dipole gravitational waves* may be emitted in addition to the *quadrupole waves* predicted by general relativity.
- This could modify the rate of gravitational wave emission late in the inspiral for merging neutron-star or neutron star, black hole binaries in an observable way.

Various alternatives to general relativity require also that gravitational waves travel at a speed different from that of light.

> The finding from GW170817 that the speed of gravity can differ by no more than *3 parts in* 10^{15} from that of light would seem to pose serious problems for such theories.

It is convenient to divide the tests of general relativity into four broad categories:

1. the *classical tests,*
2. the *modern tests,*
3. the *strong-field tests,* and
4. the *cosmological tests.*

As part of this discussion, we shall introduce also the *parameterized post-Newtonian (PPN)* formalism, which is a *standard framework* used to compare the predictions of GR

- *with data* and
- *with other theories* of gravity.

25.2 The Classical Tests of General Relativity

The tests proposed originally by Einstein are sometimes termed the *classical tests* of general relativity.

- *Precession of the perihelion for Mercury* was confirmed already at the publication of the theory. Much larger precession effects have since been measured for systems such as
  - The *Binary Pulsar* (4.2° per year) and
  - The *Double Pulsar* (17° per year),

  in agreement with the amount predicted by general relativity.
- The *bending of light in the Sun's gravitational field* was confirmed in 1919 during Solar eclipse observations.
  - These measurements have since been repeated more precisely and
  - extended to a variety of gravitational lensing phenomena described further below,

  with results consistent with the theory of general relativity.

- The *redshift of light in a gravitational field,* or equivalently *gravitational time dilation*
  - was suggested initially in the *spectra of white dwarfs*.
  - It was confirmed later in *direct experiments on Earth*.

  It also is *confirmed implicitly* every time the *Global Positioning System (GPS)* is used for location and navigation (Problem).

  The existence of a gravitational redshift is a necessary condition for general relativity to be valid.
  - However, this is sometimes viewed as a *less-stringent condition than others* because it
  - follows from the *equivalence principle alone*.

- The *existence of gravitational waves* was confirmed
  - *indirectly* by the Binary Pulsar and
  - *directly* by LIGO.

  It is fair to conclude that the *classical tests of general relativity have all been passed* with a rather high level of confidence.

25.3 The Modern Tests of General Relativity

There are various newer tests that we may term the *modern tests of general relativity*. In many cases the feasibility, accuracy, and precision of this class of tests has been aided by

- rapid *advances in detection and measurement* technology, and by
- the advent of *routine spaceflight* to place observational and experimental platforms in space.

The analyses of these modern tests of general relativity often systematize results in terms of the *parameterized post-newtonian (PPN) formalism,* which we now introduce.

25.3.1 The PPN Formalism

In comparing general relativity with other theories of gravity and with observation it is useful to have a method that

- Expresses the *nonlinear Einstein equations* in terms of the *lowest-order deviations from Newtonian gravity*.
- Such approaches may be applied to any theory of gravity
  - that *derives from a spacetime metric* and
  - embodies the *strong equivalence principle*.
- However, they are *most useful in weak gravity* since they are based on expansions.
- GR and most viable alternative theories of gravity fit this prescription.
- Under conditions of *weak gravity*
  - the differences between the predictions of general relativity and Newtonian gravity can be expressed in terms of *expansions in powers of a small parameter*.
  - The relevant small parameter is v/c, where v is a *characteristic velocity for the system*.

This *post-Newtonian expansion* (it adds corrections that go beyond Newtonian gravity) reduces to Newtonian gravity in the limit $c \to \infty$.

We now illustrate how the PPN formalism works in more detail.

- The *most general static, spherically-symmetric metric* can be expressed in the form

$$ds^2 = -A(r)(cdt)^2 + B(r)dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\varphi^2.$$

- In the limit of flat spacetime this becomes the *Minkowski metric* expressed in spherical coordinates,

$$ds^2 = -(cdt)^2 + dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\varphi^2.$$

- Thus $A(r)$ and $B(r)$ parameterize the deviation of the metric from flat spacetime.
- This suggests that for weak gravity the departure from a flat metric can be described by an *expansion in an appropriate small parameter*.

Any viable relativistic gravitational theory should agree with the results of Newtonian gravity in the weak-field limit.

- Comparing with the general metric, agreement with Newtonian gravity for weak fields requires to lowest order (Problem)

$$A(r) = 1 - \frac{2GM}{rc^2} + \ldots \qquad B(r) = 1 + \ldots,$$

  where the dots indicate higher-order corrections.

- The expansions

$$A(r) = 1 - \frac{2GM}{rc^2} + \ldots \qquad B(r) = 1 + \ldots,$$

suggest that the *small dimensionless parameter* appropriate for an expansion of the metric is GM/rc^2.

- This is not surprising, since it was shown earlier that GM/rc^2 is a *natural measure of the strength of gravity*.
- Expanding $A(r)$ and $B(r)$ in powers of GM/rc^2 gives

$$A(r) = 1 - \frac{2GM}{rc^2} + 2(\beta - \gamma)\left(\frac{GM}{rc^2}\right)^2 + \ldots$$

$$B(r) = 1 + 2\gamma\left(\frac{GM}{rc^2}\right) + \ldots,$$

where the particular form of the expansion coefficients in terms of the parameters β and γ is conventional.

  - The lowest-order corrections to the flat-space spherical metric

$$A(r) = 1 - \frac{2GM}{rc^2} + \ldots \qquad B(r) = 1 + \ldots,$$

may be termed the *Newtonian corrections.*

  - Hence the higher-order terms (shown in red above) are called the *post-Newtonian corrections.*
  - Their size is characterized by the *post-Newtonian parameters* β and γ in this example.

More generally, the *parameterized post-Newtonian (PPN) formalism*

- isolates the differences between general relativity and experiments or other theories of gravity
- in the values of a set of parameters associated with coefficients of terms in the post-Newtonian expansion of the metric.
- There are ~ 10 *such parameters in typical applications*, each tested by particular experiments or observations.
- For example, in the *PPN formulation of Will*
  - the parameter γ is associated with *light deflection and time delay*, and
  - the parameter β is associated with *orbital perihelion shift*.
  - *General relativity predicts* $\beta = \gamma = 1$, whereas
  - a different metric theory of gravity might predict values different from unity.

Thus, modern experimental tests of gravity are often reported in terms of *measured differences of PPN parameters like* β and γ from their values expected in general relativity.

Example: Frequency shifts of radio signals to and from the Cassini spacecraft as it passed near the Sun on its way to Saturn in 2002

- were used to *constrain the PPN parameter* γ to the range

$$\gamma = 1 + (2.1 \pm 2.3) \times 10^{-5}.$$

- Using the *precise Cassini value of* γ coupled with *lunar rangefinding data* permitted the constraint

$$\beta = 1 + (1.2 \pm 1.1) \times 10^{-4}.$$

to be placed on *the PPN parameter* β.

25.3.2 Results of Modern Tests

We now summarize results of some *modern tests of GR*.

1. *Shapiro time delay* is a test in the same spirit as the classical tests and has been confirmed for observations in the Solar System and beyond.

   - For example, both the time delay and the deflection of light in a gravitational field are proportional to a PPN factor $\gamma+1$, where $\gamma=1$ for general relativity.
   - Radio ranging to the *Viking spacecraft on Mars*
     - verified the light delay prediction of GR for light travel near the Sun to an accuracy of 0.1%
     - and found that $\gamma = 1 \pm 0.002$.
   - As already noted, frequency shifts of radio signals to and from the *Cassini spacecraft* found

     $$\gamma = 1 + (2.1 \pm 2.3) \times 10^{-5}.$$

2. *Gravitational lensing*, which is the generalization of the deflection of light in the Sun's gravitational field to a range of lensing phenomena, *has been confirmed in many deep-field observations*.

3. *Frame dragging and geodetic effects*, which have been *confirmed by Gravity Probe B measurements* of gyroscopic precession for a satellite in polar Earth orbit.

4. *Modern equivalence principle tests* in several categories:

   - High-precision *torsion balance experiments:*
     - The original Eötvös experiments established the *equivalence of inertial and gravitational mass* with a sensitivity of one part in $\sim 10^9$.
     - *Modern variations* have achieved 1 part in $\sim 10^{13}$.
   - The *strong equivalence principle* requires self-gravitating bound objects to follow the same paths in gravitational fields.
     - This *Nordtvedt effect* has been tested precisely by *lunar-ranging experiments* that
     - monitor the distance between Earth and reflectors on the Moon to centimeter accuracy.
     - They confirm that the Earth and Moon *fall toward the Sun at the same rate to* 1 part in $\sim 10^{13}$.
   - Another consequence of the *strong equivalence principle* is that G *should not vary with time or place*.
     - *Lunar ranging measurements* place an upper limit

$$\frac{\dot{G}}{G} = (4 \pm 9) \times 10^{-13}\,\mathrm{yr}^{-1}.$$

     - The precision with which $\dot{G}/G$ is measured is growing *quadratically with the elapsed time* for lunar-ranging measurements.

We may conclude that the *modern tests of general relativity have provided strong confirmation of the theory* within the limits of their (often quite high) precision.

25.4 Strong-Field Tests of General Relativity

Tests of general relativity described above were all carried out in *relatively weak gravity,* where *nonlinearities are small*.

- Until recently, it wasn't easy to test GR in the *strong-gravity, highly nonlinear regime* because
- all gravitational fields in the Solar System are *weak*.
- This has changed now, with *tests in much stronger fields:*

1. Tests at intermediate field strength using *binary and triplet star systems containing pulsars*.
2. Tests at intermediate field strength by *tracking stars near the Milky Way central black hole*.
3. Tests at high field strengths using direct detection of *gravitational waves from compact mergers*.

1. *Pulsars in binary and triplet systems* containing compact objects in tight orbits generally

- *sample stronger gravitational fields,* and
- *can be measured precisely* because of *pulsar timing*.

Tests with the *Binary Pulsar*, the *Double Pulsar*, and *PSR J0348+0432*

- support *validity of GR in stronger fields*, and
- confirm indirectly *emission of gravitational waves* at the rate predicted by GR.

2. Orbits of stars around the *supermassive black hole at Sgr A** provide tests of general relativity for

- *gravitational source masses* $\sim 10^6$ times larger
- and *gravitational fields* ~ 100 times stronger

than for tests of general relativity by binary pulsars.

3. *Direct observation of gravitational waves* now promises a whole range of general relativity tests in the

- *strong gravity regime* (binary neutron star merger and core collapse supernova), and the
- *very strong gravity regime* (binary black hole merger).

In particular, analysis of gravitational waves observed so far suggests that

- *black holes exist,*
- which confirms a fundamental strong-gravity prediction of GR, and that
- the extreme gravity near their event horizons is *correctly described by general relativity*.

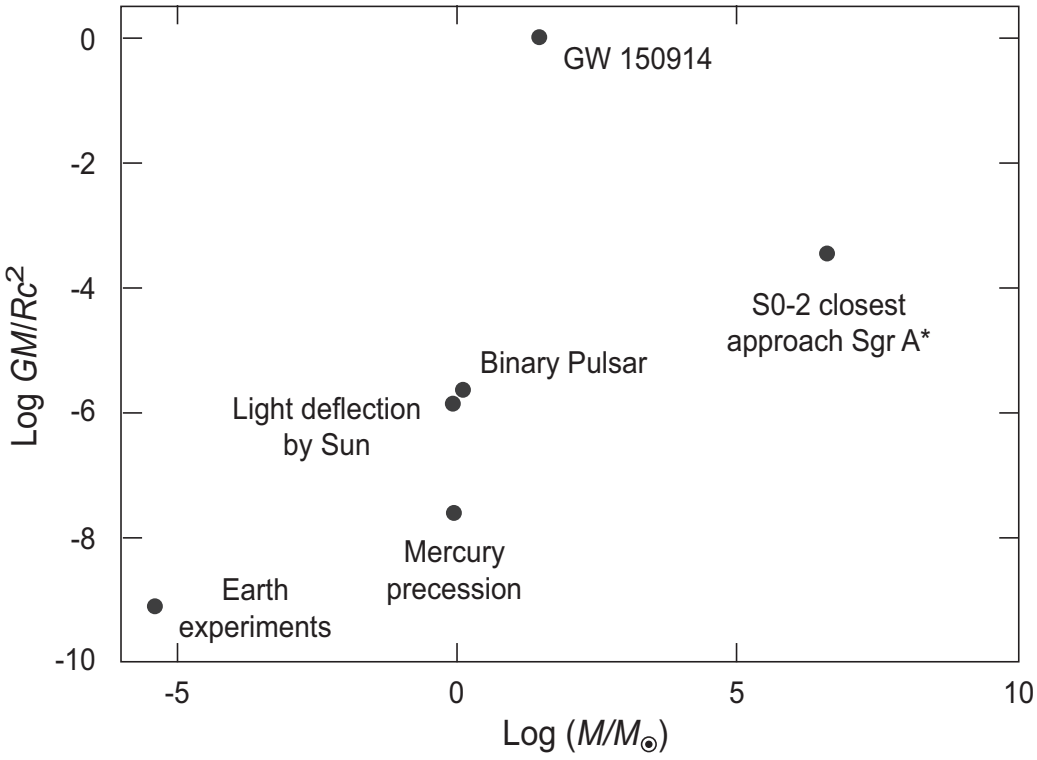


Figure 25.1: Field strength versus mass for some tests of general relativity. Gravitational waves from merging stellar-size black holes and the orbits of test stars around the $4 \times 10^6 M_\odot$ black hole at the center of the Milky Way test gravity in fields that are 10^2 to 10^6 times stronger than that for all previous tests.

The *characteristic field strength* in units of $\varepsilon = GM/Rc^2$ versus *source mass* for some experimental and observational tests of general relativity are summarized in Fig. 25.1.

Example: A typical *modified form of gravity* assumes that

- gravity is *described by a metric tensor,* as for GR, but
- an *additional field* alters the interaction over some range.

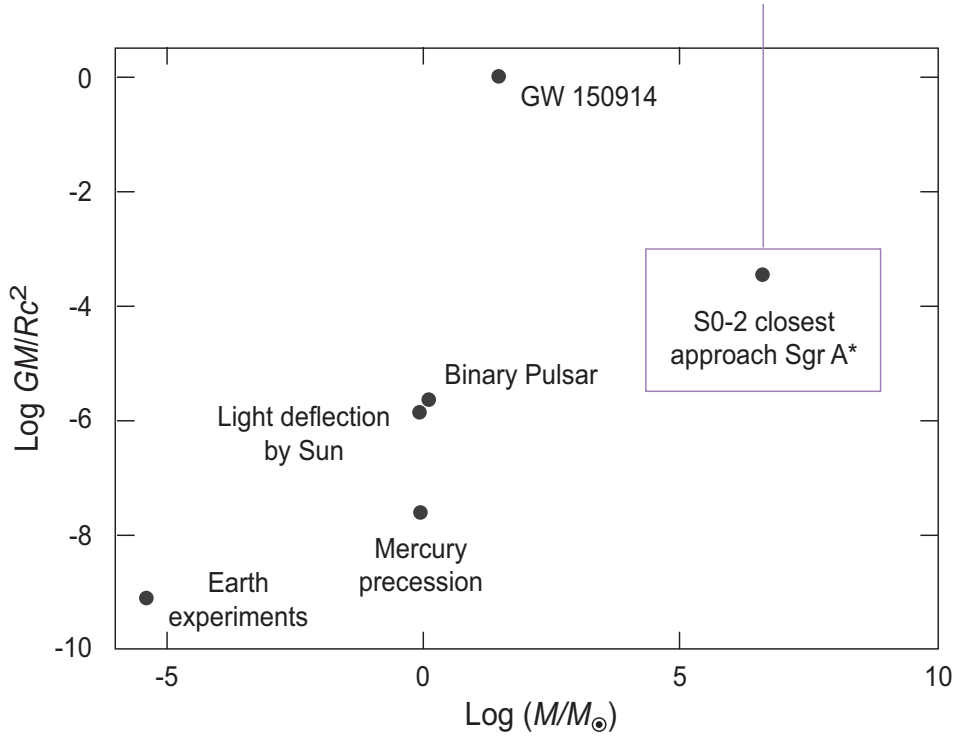
The *modified gravitational potential* is often written as

$$U = \frac{GM}{r}\left(1 + \alpha e^{-r/\lambda}\right),$$

where α and λ parameterize the additional interaction.

> Star orbits around the black hole at Sgr A* place an *upper limit* $|\alpha| < 0.016$ for $\lambda = 150\,\text{AU}$.

These Sgr A* observations *probe GR in an untested regime:*



We may expect that increasingly-stringent tests of general relativity in stronger gravity will accumulate through

- further monitoring of multiple star systems containing pulsars,
- tracking the stars orbiting the supermassive black hole at the center of the Galaxy, and
- the detection of gravitational waves from various binary merger and supernova events.

25.5 Cosmological Tests of General Relativity

In principle cosmological observations test general relativity.

- Indeed, some would argue that the observed expansion of the Universe is a confirmation of general relativity.
- It is indeed true that general relativity is *consistent* with a remarkable variety of cosmological data for the expanding Universe.
- However, the lack of fundamental understanding for the
  - source of *dark matter* and
  - source of *dark energy*

  undermines attempts to confirm general relativity independently from cosmological observations alone.

To take one extreme example:

- The mainstream has accepted that dark matter and dark energy are necessary components of a Universe that is correctly described by general relativity.
- However, a small minority argues that effects attributed to the presence of “dark matter” and “dark energy" are not evidence of unseen matter and energy at all.
- Rather (in this view) they are a failure of current gravitational theory to describe the interactions of normal matter and energy.

Recall that in the history of astronomy an apparent failure of gravitational predictions typically has been ascribed to *one of two alternatives:*

1. Gravitational theory is correct but there is unknown mass affecting the system, or
2. Gravitational theory is failing to describe the interactions of known masses and must be replaced.

Historically, each alternative has been right at different times:

1. The first in the discovery of Neptune through its perturbations on of Uranus using Newtonian gravity with an assumed *unknown mass* outside the orbit of Uranus, and
2. the second by Einstein's demonstration that the *excess precession of Mercury's perihelion* was not caused by an unknown mass in Newtonian gravity but instead indicated that *a better theory of gravity was required*.

Evidence *favors strongly the mainstream view* that general relativity is the correct classical theory of gravity, and thus that *dark matter and dark energy are real*.

- However, until there is a deeper understanding of the *source of dark matter and dark energy* the consistent point of view must be that
- *Cosmological observations are consistent with general relativity but are not yet an independent test of it*.
- Stated slightly differently, dark matter and dark energy are key components of the new cosmology and at present the *only evidence for their existence is gravitational*.

Hence it is clear that the new cosmology

- Is in impressive agreement with the aggregate of cosmological data, but
- It cannot be viewed as a rigorous *independent test* of general relativity as the correct classical gravitational theory

until more is understood about the intrinsic nature of dark matter and dark energy.

Chapter 26

Beyond Standard Models

Preceding chapters have described gravitational physics within the context of *"standard models"* of particular disciplines:

- General relativity as the standard model of classical gravity.
- The gauge theory of telectroweak and strong interactions as the Standard Model of elementary particle physics.

These theories have been *remarkably correct*, but there is evidence that they are *incomplete*. For example,

- That the principles upon which general relativity and quantum mechanics rest are *logically incompatible*, and
- the *black hole quantum information paradox* point toward the need for a new *theory of quantum gravity*, without defining what that theory should be.
- Evidence that neutrinos undergo flavor oscillations indicates that the *Standard Model of particle physics is incomplete*, but does not indicate clearly how to extend it.

In this chapter some of the still-developing ideas for extending our current understanding of gravity and its interactions with elementary particles will we introduced.

26.1 Supersymmetry and Dark Matter

In preceding chapters we have examined a number of results suggesting that most of the matter in the Universe is not seen by any standard probe except gravity.

- This unseen mass is termed *dark matter,* implying that it does not couple significantly to non-gravitational probes (electromagnetism, the weak interactions, or the strong interactions).
- A dilute gas of dark matter obeys an equation of state similar to that of normal matter, but we do not know what dark matter is.
- Likewise, we have seen empirical evidence that the Universe contains a mysterious *dark energy* that permeates all of space and is causing the expansion of the Universe to accelerate.
- The dark matter appears at this point to be distinct from the dark energy. For example, dark energy seems to require an equation of state that is fundamentally different from that of any known particle or of dark matter (negative pressure and antigravity effects).

Potential explanations of the dark matter in particular generally invoke a property of the Universe that is conjectured theoretically but for which this is not yet any evidence called *supersymmetry*.

26.1.1 Fermions and Bosons

In quantum mechanics, all elementary particles can be divided into two classes, *bosons* and *fermions,* having *fundamentally different statistical properties*.

- Fermions have *half-integer spin* and obey *Fermi–Dirac statistics.*
  - Fermion wavefunctions are *completely antisymmetric* under exchange of two identical fermions.
  - The most notable implication of fermionic statistics is the *Pauli exclusion principle:* no two identical fermions can occupy the same quantum state.
- Bosons have *integer spin* and obey *Bose–Einstein statistics.*
  - Boson wavefunctions are *completely symmetric* under exchange of two identical bosons.
  - The most notable implication of bosonic statistics is *boson condensation:* Many identical bosons can (indeed, often prefer to) *occupy the same quantum state*.

> We do not have an explanation of *why* fermions and bosons have these properties, but they do, and *all elementary particles discovered so far can be classified as either bosons or fermions.*

26.1.2 Normal Symmetries

Normal symmetries in quantum field theory relate *bosons to bosons* or *fermions to fermions*, but *not bosons to fermions*.

- For example, the (approximate)[†] symmetry called *isotopic spin* is important in nuclear physics and particle physics.
  - One implication of isotopic spin symmetry is that there is a relationship between protons and neutrons such that,
  - in a certain sense the neutron and the proton are really just *different manifestations of the same particle*.
- Particles that are related in this way by an isotopic spin symmetry are termed *isotopic spin multiplets*.

[†]Experts will note that isotopic spin symmetry

- is not exact,
- is not fundamental, and that
- neutrons and protons are composite and not fundamental particles.

Nevertheless, it is useful to introduce normal symmetries in a simple way using isotopic spin with nucleons considered as elementary particles (which they effectively are at low energy).

A symmetry or approximate symmetry like isospin is a very powerful idea, but note that in this example

- Isotopic spin symmetry, just as for any ordinary symmetry, *relates particles that are of the same quantum statistical type* (the neutron and the proton are both spin-$\frac{1}{2}$ fermions).
- There is no known instance of a set of particles that appear to be related by an isotopic spin symmetry in which some of the particles are bosons and some are fermions.

26.1.3 Symmetries Relating Fermions and Bosons

There are theoretical reasons to believe that the Universe will eventually be found to exhibit a fundamental symmetry that *goes beyond normal symmetries and relates fermions to bosons*.

- This conjectured "super" symmetry is termed, appropriately, *supersymmetry.*
- In the supersymmetry picture,
  - *every fermion* in the Universe has a *partner boson of the same mass*, and
  - *every boson* in the Universe has a *partner fermion of the same mass*.

Example: The Universe contains elementary particles called *electrons* that are

- *spin-$\frac{1}{2}$ fermions* with a
- mass of 511 keV.

According to the supersymmetry idea, there is also a *supersymmetric partner of the electron* called a *selectron* that

- is a *spin-0 boson* with
- a mass of 511 keV.

The problem with the supersymmetry idea is that *no one has ever seen the supersymmetric partner of any elementary particle* that we know to exist.

- However, there are theoretical arguments suggesting that supersymmetry may be a *broken symmetry*.
- Then the masses of supersymmetric partners of known elementary particles might be *pushed up to a high value by the symmetry breaking*.
- If the broken supersymmetric masses are large enough, they would not have been produced in any accelerator experiments carried out so far.

Thus, there is strong theoretical prejudice that *broken supersymmetry exists in the Universe*, though we have yet to find any direct evidence for it.

The most favored explanation for dark matter has been that it consists of as-yet undiscovered elementary particles that are *supersymmetric partners of known elementary particles*.

- Then the very weak coupling to ordinary matter, and thus the "darkness" of the dark matter, could be explained as a *quantum selection rule effect*.
- That is, supersymmetric particles would be expected to carry *supersymmetric quantum numbers not carried by ordinary particles*.
- Hence, the probability for interaction of supersymmetric particles with ordinary matter would be *strongly suppressed by conservation laws and selection rules* associated with these quantum numbers.

  For those lacking a background in quantum mechanics, this can be translated roughly as *supersymmetric matter and ordinary matter are so different that they don't even know how to interact with each other*.

- Supersymmetry may also play an important role in *relating dark energy to quantum fluctuations of the vacuum,* though we shall see that our understanding of this issue is far from satisfactory.

Experiments as of late 2017 at the Large Hadron Collider that were predicted by various educated guesses to have high enough energy to produce supersymmetric particles *have failed to find evidence for them*.

- This does not necessarily rule out broken supersymmetry as the origin of dark matter.
- However, it suggests that
  - at best that the simplest ideas about supersymmetry and dark matter may have to be discarded, and
  - at worst supersymmetry may have nothing to do with dark matter.

Thus, there is *abundant gravitational evidence that dark matter exists*, but at present there *isn't a scintilla of evidence* in support of what has been the leading candidate to explain it.

26.2 Vacuum Energy from Quantum Fluctuations

We have seen empirical evidence that the Universe contains a mysterious *"dark energy"* that permeates all of space and is *causing the expansion to accelerate*.

- Supposing this to be true, what is the source of this remarkable behavior (which we have seen to be *equivalent to the presence of antigravity in the Universe*)?
- Given the successes of relativistic quantum field theory in elementary particle physics, it is natural to seek the source of the dark energy in terms of relativistic quantum fields.

No dilute gas of known particles can exhibit an equation of state similar to that inferred for dark energy.

- Thus, if the source of dark energy lies in quantum fields, those fields must be associated with as-yet undiscovered elementary particles.
- But the empirical evidence also suggests that dark energy is a property of space itself, and would still exist even in *empty space* (no particles and no fields).
- Quantum physicists term the ground state (lowest energy state) of a system the *vacuum state.*

Classically, the vacuum state of the Universe would consist of a state with no matter or fields, and no energy. But quantum field theory complicates this issue in two ways:

1. Because of *spontaneous symmetry breaking* (*hidden symmetry,*) the vacuum state of a system (the lowest-energy state) need not be a state with no fields present.

   Various examples are known where a state with no fields present is actually higher in energy than a state with a particular combination of fields, so that the lowest-energy state has fields present.

2. Classically, if there are no fields the *energy is zero*.

   - But even a state that, on average, has no fields has non-zero energy ΔE because of fields that fluctuate into and out of existence over a period Δt such that the uncertainty principle relation $\Delta E \cdot \Delta t \geq \hbar$ is satisfied.
   - These are termed *vacuum fluctuations* and the associated energy is termed the *zero-point energy* or the *vacuum energy.*

   The *reality of the vacuum energy* is demonstrated

   - indirectly by the many successes of the Standard Model of elementary particle physics, and
   - directly through phenomena measurable in the laboratory like the *Casimir effect.*

In light of point 2, the vacuum state of the Universe ("empty space") will have a non-trivial content because of vacuum fluctuations.

- It is tempting to try to associate the effects of dark energy with these vacuum fluctuations.
- Let us attempt a quantitative estimate of the energy density of empty space associated with quantum vacuum fluctuations to see if these effects could account for the accelerated Universe.

26.2.1 Vacuum Energy for Bosonic Fields

Bosonic fields are simpler than fermionic fields, so we first assume the vacuum energy to be associated with bosonic fields.

- In quantum field theory, bosonic fields may be expanded in terms of an *infinite number of harmonic oscillators*.
- Assuming all fields to be bosonic and expanding them as a collection of harmonic oscillators, the total energy is

$$E = \sum_i (n_i + \tfrac{1}{2})\hbar\omega_i,$$

  where n_i can be zero or any positive integer (bosons).
- Thus, even if the Universe is "empty", *there is a vacuum energy* associated with *zero-point energies of the fields*

$$E = E_\Lambda = \tfrac{1}{2}\sum_i \hbar\omega_i.$$

- This sum is over oscillators defined at each point, so vacuum energy density should be *constant over all space.*
- Converting the sum to an integral over momentum using

$$\hbar\omega = \sqrt{p^2c^2 + m^2c^4},$$

  and neglecting masses, yields a *constant energy density* $\varepsilon_\Lambda^{\mathrm{b}}$ *for bosonic vacuum fluctuations*,

$$\varepsilon_\Lambda^{\mathrm{b}} \propto \int_0^\infty p^3 dk = \infty.$$

Since we have found

$$\varepsilon_\Lambda^{\mathrm{b}} \propto \int_0^\infty p^3 dk = \infty,$$

our first simplistic attempt to estimate the energy density of the vacuum *leads to nonsense*.

- But general relativity likely *breaks down on the Planck scale*.
- Thus it is plausible that *new physics on that scale* introduces a *momentum cutoff in the preceding integral*.
- Setting the upper limit of the integral to the *Planck momentum* $P_{\mathrm{pl}} \simeq 10^{19}\,\mathrm{GeV\,c^{-1}}$ gives

$$\varepsilon_\Lambda^{\mathrm{b}} \simeq \frac{(cP_{\mathrm{pl}})^4}{16\pi^2\hbar^3c^3} \simeq 0.824 \times 10^{118}\,\mathrm{MeV\ cm^{-3}},$$

  which is *very large but at least finite*.
- But cosmological observations indicate that *vacuum energy density is the same order of magnitude as the critical (closure) density*, which is only about $10^{-2}\,\mathrm{MeV\ cm^{-3}}$.

Therefore, this simple estimate of the vacuum energy density gives a value that is about *120 orders of magnitude too large* to account for the observed acceleration of the Universe!

This estimate of the vacuum energy density has been termed

> *The most spectacular failure in all of theoretical physics.*

It is not yet clear whether this means that

1. vacuum fluctuations are not responsible for the acceleration of the Universe, or whether it means that
2. we have a (monumental) lack of understanding of how to properly calculate the vacuum energy density.

Some improvement in our estimate is afforded by assuming

- the Universe to be *composed of fermionic fields in addition to bosonic ones*, and
- to assume that there is a *supersymmetry that relates the boson and fermion fields*.

> A serious treatment of this subject is beyond our scope, but we now make some simple estimates of the vacuum energy density expected if a supersymmetry between fermions and bosons exists.

26.2.2 Vacuum Energy for Fermionic Fields

Because fermionic wavefunctions are antisymmetric and bosonic wavefunctions are symmetric under exchange,

- fermionic fields have a spectrum that is like a sum over harmonic oscillators for bosons, except that
  1. The sign of the $\frac{1}{2}\hbar\omega_i$ term is *negative* and
  2. Occupations n_i are *restricted to 0 or 1* (Pauli principle):
     $$E = \sum_i (n_i - \tfrac{1}{2})\hbar\omega_i, \qquad (n_i = 0, 1).$$
- Thus, the zero-point energy for fermionic fields (corresponding to $n_i = 0$) is *negative*.

This looks even *less promising* as an explanation of the accelerating universe than bosonic vacuum fluctuations.

Note though that *negative energy densities* are intriguing for three staples of science fiction:

- *warp drives*,
- *wormholes*, and
- *time machines*.

These might be *hypothetically possible* if we had at our disposal *exotic material* with a negative energy density. *Alas, no such material is known*.

26.2.3 Supersymmetry and Vacuum Energy

The potential relevance of supersymmetry to the vacuum energy is clear from the preceding observation that

- the zero-point energy of fermion fields is *opposite in sign to that of boson fields.*
- If by supersymmetry every fermion field has a partner supersymmetric boson field, their contributions to the vacuum energy density can *cancel each other.*
- Indeed, a detailed theoretical treatment suggests that if supersymmetry were exact, the zero-point energy of all boson and fermion fields would *exactly cancel, leaving a vacuum energy density equal to zero.*
- But we have already argued above that supersymmetry (if it exists) must be *at least partially broken* to be in accord with observations.

Therefore, if the Universe has a *broken supersymmetry*,

- the contributions of fermion and boson fields to zero-point energy could *almost, but not quite, cancel*,
- leaving a *small net vacuum energy density*.

A supersymmetric quantum field estimate gives for the vacuum energy density associated with zero-point motion of the fields

$$\varepsilon_\Lambda = \left(\frac{m_\mathrm{f}^2 - m_\mathrm{b}^2}{P_\mathrm{pl}^2} \right) \varepsilon_\Lambda^\mathrm{b},$$

- where $\varepsilon_\Lambda^\mathrm{b}$ is the earlier bosonic estimate and
- $\Delta m^2 \equiv m_\mathrm{f}^2 - m_\mathrm{b}^2$ is the difference in mass square scales between bosonic (b) and fermionic (f) supersymmetric partners.

Hence, our earlier absurdly high estimate will be reduced by the initial factor involving the mass square difference Δm^2:

$$\varepsilon_\Lambda = \underbrace{\left(\frac{\Delta m^2}{P_\mathrm{pl}^2} \right)}_{\text{reduction}} \varepsilon_\Lambda^\mathrm{b},$$

- But the *failure so far to observe supersymmetric particles* in experiments places a lower limit $\Delta m^2 / P_\mathrm{pl}^2 \geq 10^{-34}$.
- Thus our new estimate of Δm^2 is still more than *80 orders of magnitude too large to accord with observation.*

Even by cosmological standards that is a little beyond the pale, strongly suggesting either that

- dark energy is not vacuum energy, or
- we don't really understand vacuum energy!

Table 26.1: Quantities characteristic of the Planck scale

Quantity	Value
Planck mass	$1.2 \times 10^{19}\,\mathrm{GeV}/c^2$
Planck length	$1.6 \times 10^{-33}\,\mathrm{cm}$
Planck time	$5.4 \times 10^{-44}\,\mathrm{s}$
Planck temperature	$1.4 \times 10^{32}\,\mathrm{K}$

26.3 Quantum Gravity

As we imagine extrapolating the history of the Universe backward in time,

- The big bang theory tells us that the Universe becomes more dense and hotter, and the relevant distance scales become shorter.
- But if the distance scales become short enough (of atomic dimensions or smaller), the theory of quantum mechanics must be used to describe physical events.
- Therefore, as we extrapolate back in time to the beginning of the Universe, eventually we reach a state of sufficient temperature and density that a *fully quantum mechanical theory of gravitation would be required.*
- This is called the *Planck era,* and the corresponding scales of distance, energy, and time are called the *Planck scale.*.

Quantities characteristic of the Planck scale are listed for reference in Table 26.1.

It is instructive to compare the Planck scale with the scale of actual data for properties of elementary particles.

- In the largest particle accelerators, energies of several hundred GeV can be reached.
  - The temperature of a gas having particles of this average energy is approximately 10^{15} K, and
  - the time after the big bang when the temperature would have dropped to this value is about 10^{-10} s.
- Therefore, all speculation about the Universe at times earlier than this is based on theoretical inference.
- Clearly the Planck scale lies far beyond our present or foreseeable ability to probe it directly, but presumably was relevant in the first instants of the big bang.
- At the Planck scale we must apply the principles of quantum mechanics to the gravitational force.
- But our best theory of gravity is general relativity and it does not respect the principles of quantum mechanics.
- Likewise, our best microscopic theory is quantum mechanics, and it does not respect the principles of general relativity.

- What is required then is a theory of gravity that also is consistent with quantum mechanics. This could be termed a theory of *quantum gravity.*
- Unfortunately, no one has yet understood how to accomplish this very difficult task, and we do not have an internally consistent theory of quantum gravity.

26.3.1 Superstrings and Branes

Our ordinary description of the microscopic world, quantum mechanics, is based on the idea that elementary particles are point-like (*exist at a point in space*).

- *A point has zero dimension*, since it has neither breadth, width, nor height.
- This feature of quantum mechanics leads to *serious technical mathematical problems*.
- In our description of the strong, weak, and electromagnetic interactions, a *complex mathematical prescription has been worked out* that avoids these problems.
- The technical term for this prescription is *renormalization.*

Renormalization systematically removes infinite quantities that would otherwise crop up in the theory and leaves us with a *logically consistent quantum description*.

Gravity is Special

- Because of its fundamental properties, the renormalization procedure that works for the other three fundamental interactions *fails for gravity*.

Ultimately this failure is because the

- electromagnetic, weak, and strong interactions are *mediated by spin-1 fields*, but
- gravity is *mediated by a spin-2 field*.

- Thus, if we try to apply ordinary quantum mechanics to gravity we end up with quantities that become infinite in the theory.
- Since we don't understand how to deal with these infinities, quantum gravity based on point-particle quantum mechanics *is not logically consistent*.
- This has two related implications.

1. We cannot describe gravity on the Planck scale.
2. We cannot join gravity with the other three forces into a unified description of all forces.

The most promising present alternative for a theory of quantum gravity is called *superstring theory*, but it is not yet clear whether it can provide a correct picture.

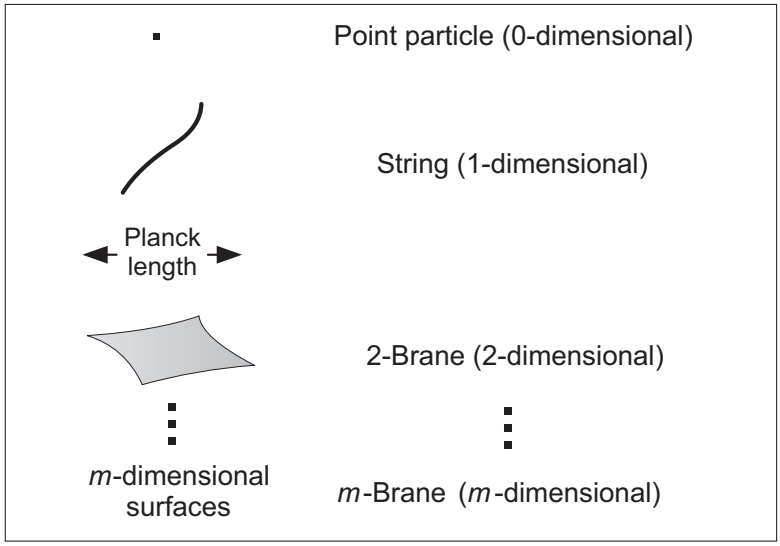
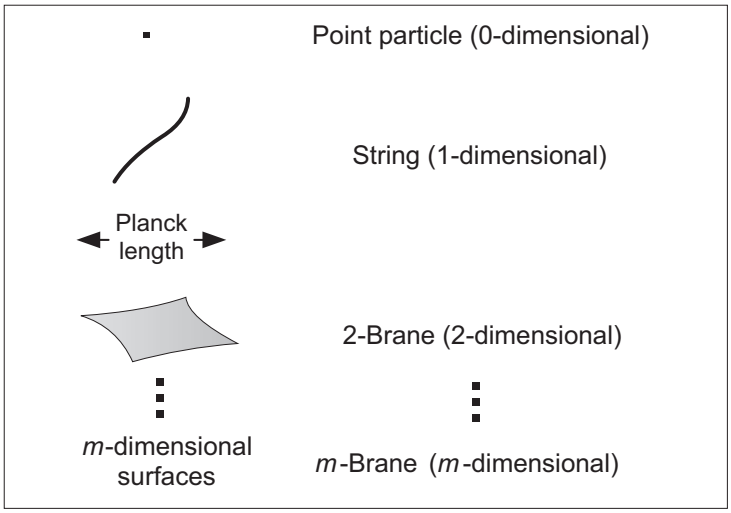


Figure 26.1: Schematic illustration of points, strings, and branes.

One-Dimensional Building Blocks

The basic idea of superstrings *changes the assumption* that elementary building blocks of the Universe are point-like particles.

- Instead, superstring theory assumes that they are tiny (Planck length) objects that *are not points but instead have a length: they are like strings.*
- These strings are assumed to *possess a supersymmetry*, so these elementary building blocks are called *superstrings.*
- Because they have a length, *they are 1-dimensional*, rather than the 0-dimensional particles of ordinary quantum mechanics. The top portion of Fig. 26.1 illustrates.
- Assuming that the building blocks are not pointlike *permits many of the troublesome infinities to be avoided.*

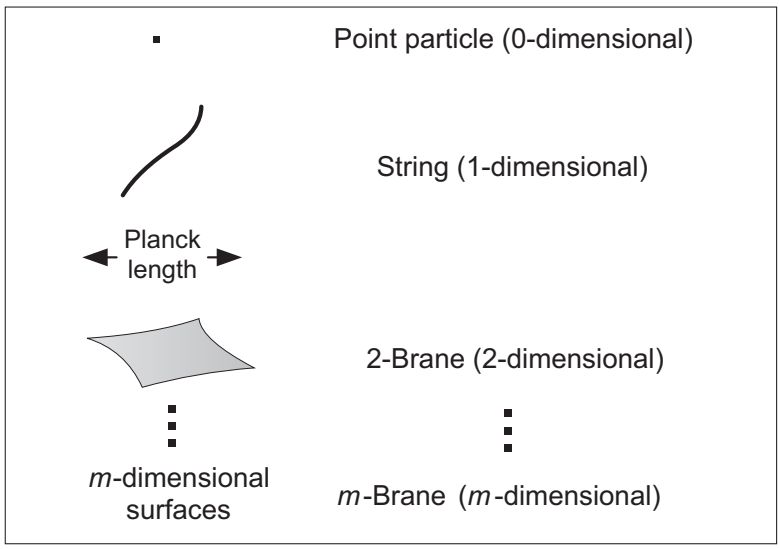


Brane Theory

Once the basic idea of superstrings were introduced in the early 1980s, five different versions of string theory were developed.

- In the 1990s, it was realized that these five versions were actually *strongly related to each other*.
- They seemed different because the discussion to that point had been based on *approximate solutions*.
- More exact solutions indicated that existing versions of superstrings could be *unified in a more general theory*.
- This unified theory is called *m-brane theory*— often shortened to *M*-theory—because it generalizes the idea of fundamental particles having 1 dimension to m dimensions.
- The resulting geometrical surfaces are called m-branes, with the integer m signifying the number of dimensions.

The lower portion of the above figure illustrates a *2-brane:* a 2D object with dimensions comparable with the Planck length.



The Basic Stuff of the Universe

Superstring theory, and its generalization *M*-theory, imply that our current "elementary particles" *are not really elementary*.

- They have an *internal structure on the Planck scale*.
- This internal structure consists of elementary building blocks that are *not point particles* but are instead *1-branes* (strings), *2-branes*, *3-branes*, ... (theory suggests that branes up to *9-branes can* exist).
- We have no hope in the foreseeable future to probe the Planck scale directly.
- Thus the challenge to these new theories is whether they can make any testable scientific predictions at energies below the Planck scale that would allow their validity to be checked.

26.3.2 Testing Branes and Superstrings

The present discussion implies that a logically-consistent quantum theory of gravity and a unification of all four fundamental forces may be possible based on *M-theory*.

- It is difficult to be certain because *its mathematical methods are challenging* and still being developed.
- As a consequence, the theory can't yet be used to make systematic predictions that can be tested by currently-feasible experiments.
- Recall that a hallmark of modern science is experimental verification of hypotheses: *a theory must pass the experimental test to be acceptable as a description of nature*.

It is hoped that *M-theory* will be capable of a quantitatively testable prediction within a decade or so, but it is not yet clear that this is feasible.

One idea leading to a qualitative testable prediction concerns the *strength of the gravitational force*.

- According to brane theory the true number of dimensions for spacetime is more than the four in evidence.
- It has been proposed that the well-known weakness of gravity is an artifact of our observations being confined to four spacetime dimensions.
- In this proposal, *gravity really is very strong* but *most of its strength has "leaked" into other dimensions* that are not visible in our low-energy world.
- Thus, as particle accelerators probe higher energies and therefore shorter length scales,
- eventually they might begin to see *evidence for additional dimensions well before the Planck scale*.
- As a consequence, *the effective strength of gravity might suddenly begin to grow* as the energy of the probe is increased.

With any foreseeable technology there is *no hope of compressing matter into a tiny black hole* if the strength of gravity varies in the same manner measured at larger scales in the laboratory.

But *if the strength of gravity were to grow more rapidly than expected* at very short distances,

- it would be much easier to make a small black hole.
- Hence the *appearance of mini black holes* in high-energy particle collisions,
- which could be detected through their *rapid decay in a burst of Hawking radiation,*
- would be *indirect confirmation of the brane-theory hypothesis*.
- The highest-energy collisions presently available are for
  - the *Large Hadron Collider (LHC)* near Geneva, and for
  - the highest-energy *cosmic rays*.

There is *no evidence thus far* for the production of black holes in such collisions.

26.3.3 The AdS/CFT Correspondence

One development that has received considerable attention is called *AdS/CFT correspondence* or more generally *gauge/gravity duality.*

- This is a conjectured relationship between
  - a theory of quantum gravity in a particular spacetime and
  - a quantum field theory formulated on the *boundary* of that spacetime.
  - Hence the quantum field theory is defined in a space with *one less dimension* than the spacetime in which the gravity is defined.
- The AdS/CFT correspondence is a specific example of the *holographic principle* discussed in earlier chapters.

AdS refers to

- a 5-dimensional *anti-de Sitter space,*
- which corresponds to a metric formulated in *five dimensions* that is *like the de Sitter metric* but with the *opposite sign* for the cosmological constant.
- AdS has constant negative curvature, with gravity that is *attractive* (unlike the *repulsive gravity* of de Sitter space).

The boundary of 5D AdS is 4D.

- It is proposed that a *conformal field theory (CFT),* (quantum field theory in 4D that is *invariant under conformal transformations*), lives there, and that
- the CFT is *dual* to the theory of gravity in AdS space in that solutions in one space imply corresponding solutions in the other space.

In this picture

- the anti-de Sitter space is called the *bulk space* and
- the quantum field theory lives in the *boundary space*.

> 5D AdS is particularly useful as the bulk space because it possesses *symmetries analogous to those of conformal transformations in 4D*.

AdS does not resemble reality: it is 5D and our accelerated expansion suggests de Sitter rather than anti-de Sitter space.

- The importance of AdS lies in the *duality* mapping a string theory in AdS to a quantum field theory on its boundary.
- This implies that 4D QFT solutons can be obtained from 5D gravity solutions through *duality relations*.
- Most interesting is that if the theory on one side of the duality is *strongly coupled* (difficult to solve) then the theory on the other side is *weakly coupled* (easier to solve).

Thus, it has been proposed that the *AdS/CFT duality* might allow quantum field theories for say

- quantum chromodynamics (QCD) or
- highly-correlated electron systems

(strongly coupled and difficult to solve) to be *solved using the weakly-coupled dual* in anti-de Sitter space.

- In current implementations the CFT side does not look like real quantum field theories.
- For example, the CFT side has a high degree of supersymmetry not exhibited by real QCD.

However *some important features* of the CFT resemble those of actual 4D quantum field theories.

How Many Dimensions?

Superstring theories indicate that even our ideas about the *number of dimensions in spacetime* may require revision.

- These theories suggest that *spacetime has more dimensions than the four* (three space and one time) that we are used to dealing with in our everyday lives.
- However, the extra dimensions are conjectured to not be visible until we get down to distances close to the Planck length.
- This is perhaps not a completely crazy idea:
- We know of examples where we can be fooled about the number of dimensions for a space if we cannot resolve it on a sufficiently microscopic scale.

A simple analogy is a *cylindrical pipe*.

- A cylinder is a 2D surface, but
- if we view the pipe from a distance it looks like a line, which is a 1D surface.

Only when we are close can we see the *"hidden" extra dimension*.

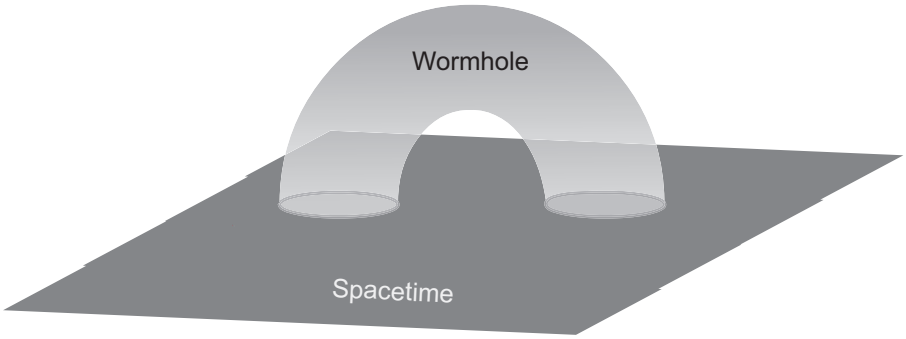


Figure 26.2: A wormhole that connects two regions of the same approximately flat space.

26.3.4 Spacetime Foam, Wormholes, and Such

The domain of quantum gravity is presumably bizarre by our usual standards. Since we do not have an adequate theory, we cannot make very precise statements, but

- Qualitatively we have reason to believe that on this scale even space and time may become something other than our usual conceptions.
- For example, spacetime may develop "wormholes", such as the one illustrated in Fig. 26.2 for the more easily visualized 2-dimensional case.
- A wormhole could connect two regions of the Universe without going through the normal space in between the two points.

Even more disconcerting is the *possibility that spactime may no longer even be continuous* below the Planck scale.

- This has been described poetically as dissolution of space-time into a frothing and bubbling *"spacetime foam"*.
- Relativity implies that space and time are not what they seem, but with relativity we could at least retain the idea of spacetime as a continuous thing.
- With quantum gravity, even that may not be possible.

Quantum Fluctuations of the Metric

Gravity differs fundamentally from the other basic forces, which act *in spacetime*.

- For other basic forces, spacetime is often roughly a *passive "stage" for physical events*.
- But gravity *distorts spacetime* and is in turn *generated by distortion of spacetime*.
- Thus, quantum fluctuations of the gravitational field *don't occur on a passive stage.*
- Rather, it is *the spacetime stage itself (the metric) that fluctuates* at the quantum level of the gravitational field.

This accounts for many of the strange possibilities described in this section.

26.3.5 The Ultimate Free Lunch?

In quantum mechanics even "empty" space is fluctuating with energy and particle–antiparticle pairs can materialize as excitations out the vacuum.

- The strangest of all the strange ideas associated with quantum gravity is that
- perhaps the Universe itself is a fluctuation in the "spacetime vacuum" (which corresponds to the *absence of space and time*).
- That is, perhaps at creation *an expanding spacetime appeared out of "nothing" as a quantum fluctuation*, giving birth to our Universe.

This idea has been dubbed *the "ultimate free lunch",* since it corresponds to creating a Universe from literally nothing, not even space or time.

26.3.6 The Breakdown of Current Physical Laws

Since we do not yet have a consistent wedding of general relativity with quantum mechanics, the presently understood laws of physics may be expected to break down on the Planck scale.

- Therefore, our standard picture of inflation followed by the standard big bang says nothing about the Universe at those very early times preceding inflation.
- In this respect then, we can be relatively certain that *our currently understood laws of physics are not complete.*
- However, the Planck scale is so incredibly small that this may have been significant only in the fleeting instants corresponding to the creation of the Universe, so *is it relevant?*
- One viewpoint is that we have no method to probe the Planck scale, so it is of no significance in the here and now.
- However, if the Universe passed through the Planck scale early in its history, *it is possible that events at the Planck scale influenced the the nature of the Universe today.*

26.3.7 Does the Planck Scale Matter?

If the Universe passed through the Planck scale early in its history, it is possible that the Universe is the way that it is because of events that happened at the Planck scale.

- Perhaps the "ground rules" governing physical law in our Universe were *laid down at the Planck scale*.
- Then those rules *could have been different* if different things had happened at the Planck time.
- Hence the Planck scale could be scientifically relevant to a present understanding of the Universe by *delimiting the possibilities through setting initial conditions*.

As an example, *M-theory* implies that the actual number of spacetime dimensions is greater than the four that are perceived in our low-energy world, but that

- the "extra" dimensions are not visible to low-energy probes.
- One paradigm for rendering the extra dimensions invisible is *compactification.*
- In compactification the extra dimensions are *"rolled up" to such small dimensions* that that they can be seen only by probes having Planck-scale energies.
- In this regard, recall the earlier analogy of whether a pipe appears to be 1D or 2D depending on spatial resolution.

In condensed matter physics it is often possible to find or construct materials that effectively have *fewer than three spatial dimensions* because their interactions in the directions defined by other dimensions are negligible.

- There are many examples from effective 1D and 2D materials that the number of spatial dimensions for a many-body system
- can have a profound effect on the types of collective (*emergent,* in the jargon) states that exist at low energy.
- The properties of the everyday world are *strongly influenced by such emergent states*.
- Thus the nature of reality at the scale of daily existence *may depend fundamentally on the number of spacetime dimensions compactified* at the Planck scale.

Finally, there might be *undiscovered objects in the present Universe* that carry information about the Planck scale.

- For example, suppose that *Hawking mini black holes* were created early in the big bang.
- Then the potentially-observable endpoint for evaporation of those black holes through Hawking radiation could *probe the Planck scale, even in the present Universe*.
- If that were true, then even science in the here and now might be impacted directly by the Planck scale.